EXACT-SOLUTION VOLUME AND LENGTH GENERALIZATION IN TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Research on transformer expressivity shows whether a transformer is capable of solving a given task, but gives little indication of whether the solution, if learned, is generalizable to longer input lengths. We study this question through *normalized exact-solution volume (NESV)*: the fraction of a bounded parameter region that achieves an exact solution on every input of length n. For fixed-width, single-layer transformers with $\log n$-scaled attention, we establish asymptotic bounds on NESV for four tasks: FIRST ($\Theta(1)$), MAJORITY ($\Theta(1/(n \log n))$), INDEX ($\Theta(1/n^3)$), and PARITY (0). These results are consistent with previous empirical results: the faster the exact-solution volume decays with input length, the harder it is to length-generalize on that task. Looking deeper into INDEX, our volume analysis reveals two error sources that grow with n. Consequently, we study a transformer model that would structurally eliminate one of the terms, theoretically improving the NESV bound to $\Theta(n^{-1})$, and empirically achieving 85% accuracy when tested at $10\times$ the training length, compared with the 60% accuracy of the original model. We conclude that volume analysis may be a useful approach to identify concrete sources of length sensitivity and thus provide insights into task-specific model refinements.

1 INTRODUCTION

Theoretical work on the expressivity of transformers (cf. the survey by Strobl et al., 2024) can be seen as a prelude to the study of learnability by transformers, since the existence of a transformer and parameters that solve a given problem is a prerequisite for a transformer to learn to solve the problem. But the existence of a single point in parameter space does not tell us much about whether that point is actually learnable. In this paper, following Kover et al. (2026), we ask a more difficult question: given that a problem is solvable by a transformer with some parameters, how large is the set of successful parameters? While Kover et al. (2026) arrived at somewhat pessimistic conclusions, we look at single-layer transformers with $\log n$*-scaling* and find that, the volume of the set of successful parameters, as a function of the sequence length n, provides a way to quantify the difficulty of length-generalization in various problems.

The drive for looking beyond expressivity is also visible in experiments on simple sequence tasks. Chiang & Cholak (2022) construct exact transformers for FIRST (whether the first symbol is a `1`) and PARITY (whether the number of `1`'s is odd), but report that standard training fails on PARITY, and that FIRST does not necessarily extrapolate to longer inputs without $\log n$-scaling. Hahn & Rofin (2024) report that fitted PARITY solutions become much sharper as input length grows than fitted solutions for FIRST and MAJORITY (whether there are more `1`'s than `0`'s). Thus, tasks that are all expressible can behave differently under training and length extrapolation.

Such findings motivate examining the geometry of the parameter sets that implement a task. In related work on loss landscapes, Chiang et al. (2023) argue that the geometry and volume of low-loss regions can explain part of neural-network generalization without relying entirely on the implicit bias of gradient descent. We study a related exact-solution quantity at each finite input length. For a fixed task and architecture, we restrict the trainable parameters to a bounded region and measure the fraction that correctly solves the task on every input of length n. We call this quantity the *normalized exact-solution volume (NESV)*. Its dependence on n describes how the full set of exact solutions changes as the computation is required to handle longer inputs. With uniform sampling over the bounded parameter

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.

Table 1: NESV bounds and mean accuracy at $10\times$ the maximum training length for the four tasks. The models $\mathcal{A}_{\text{gen}}$, $\mathcal{A}_{\text{sin}}$, and $\mathcal{A}_{\text{pos}}$ are described in Section 2.2, while $\mathcal{A}_{\text{norm}}$ is described in Section 4.2.

Task	Model							
	Positional ($\mathcal{A}_{\text{pos}}$)		General ($\mathcal{A}_{\text{gen}}$)		Sinusoidal ($\mathcal{A}_{\text{sin}}$)		Normalized ($\mathcal{A}_{\text{norm}}$)	
	NESV	Acc. (%)	NESV	Acc. (%)	NESV	Acc. (%)	NESV	Acc. (%)
FIRST	$\Theta(1)$	100.0	0	51.12	0	55.06	$\Theta(1)$	100.0
MAJORITY	—	92.7	$\Theta((n \log n)^{-1})$	82.6	$\Theta((n^3 \log n)^{-1})$	87.5	—	92.7
INDEX	$\Theta(n^{-3})$	59.6	0	51.9	0	50.2	$\Theta(n^{-1})$	85.1
PARITY	0	—	0	—	0	—	0	—

region, this is the probability of sampling an exact solution. Comparing this baseline with empirical training results can help distinguish the effects of initialization and optimization from the nature of the task and the selected transformer model in terms of learnability.

We consider four binary classification tasks on binary strings. FIRST returns the first input symbol, whereas INDEX returns the symbol at a position specified by one of the input tokens. MAJORITY returns whether more than half of input symbols are 1's, whereas PARITY returns whether the number of 1's is odd. We refer to the first pair as *positional* tasks, and the second pair as *counting* tasks.

We conduct our analyses on the simplest models suited to solve the corresponding problems. In all cases, we only consider single-layer transformers. For the positional tasks, we use a model with positional encoding (j, j^2) where the Query and Key maps only depend on token positions and the Value map only depends on the tokens themselves (following the form of the construction of Bhattamishra et al. (2024)). For the counting tasks, the positions of the tokens are irrelevant, so we primarily use a model without positional encoding. We also examined a model with sinusoidal positional encoding ($\sin(j)$, $\cos(j)$).

See Table 1 for a summary of the asymptotic bounds obtained with respect to input length n. Under the positional model $\mathcal{A}_{\text{pos}}$, the NESV for INDEX is $\Theta(n^{-3})$, while FIRST has constant NESV. The sinusoidal model $\mathcal{A}_{\text{sin}}$ has volume 0 for both. Under the position-free model $\mathcal{A}_{\text{gen}}$, the NESV for MAJORITY is $\Theta(1/(n \log n))$. Under the sinusoidal model $\mathcal{A}_{\text{sin}}$, the NESV for MAJORITY is instead $\Theta(1/(n^3 \log n))$. PARITY has a zero NESV whenever $n > h + 1$ for both models, h being the width of the transformer. This is a direct consequence of the standard one-dimensional capacity of shallow ReLU networks (Arora et al., 2018) in the finite-length setting, and is consistent with the broader result that no one-layer, one-head transformer computes PARITY across all lengths (Kozachinskiy et al., 2026b).

We train those tasks under the same protocol. The empirical results are in line with our NESV analysis: the task-model pair whose NESV shrinks faster as n grows shows worse length generalization. The only counterexample was that, under the general sinusoidal model, MAJORITY trains as well as under the position-free model $\mathcal{A}_{\text{gen}}$, although the former's volume shrinks faster with n.

Among the tasks we examine, INDEX demonstrates particularly bad length generalization (60% average accuracy at ten times training length) despite being expressible (Kozachinskiy et al., 2026a). Our NESV analysis reveals the underlying mechanisms. Exact retrieval of a designated position k requires that position to receive enough attention, so its attention score must remain within a constant of the maximum attention score. Expressing the difference in attention scores between neighboring positions as a function of k reveals two key error components, one proportional to k^2 and another proportional to k, which are the sources of failing length generalization.

Guided by this analysis, we modify the model so that the error term proportional to k^2 no longer grows with input length. This is called the normalized positional model. The NESV improves from $\Theta(n^{-3})$ to $\Theta(n^{-1})$. In our experiments, the modified model also generalizes substantially better on INDEX, achieving an average accuracy of 85% percent at ten times the training length.

Taken together, the results show that expressivity alone can conceal qualitatively different parameter-space behavior. Volume analysis, on the other hand, may offer another useful tool to further identify the computational constraints responsible for learning behaviors and offer insight for task-specific model refinements.

2 SETUP

In this section, we describe the tasks to be analyzed and the transformers to be used in the analysis.

2.1 TASKS AND EXACT-SOLUTION VOLUME

For $n \geq 1$, write $[n] = \{1, \ldots, n\}$ and let $w = (w_1, \ldots, w_n) \in \{0,1\}^n$. We study

$$\begin{aligned} \textsc{First}_n(w) &= w_1 \\ \textsc{Index}_n(w,k) &= w_k, \quad k \in [n], \\ \textsc{Majority}_n(w) &= \mathbf{1}\left\{ \sum_{j=1}^{n} w_j \geq \frac{n+1}{2} \right\} \quad \text{for odd } n, \\ \textsc{Parity}_n(w) &= \left(\sum_{j=1}^{n} w_j \right) \bmod 2. \end{aligned}$$

For INDEX, the n tokens $w_1, \ldots, w_n$, which we call *data* tokens, are followed by a single token, which we call the *index* token, encoding k. We refer to this as *right-hand sequence* INDEX*ing*. For the other tasks, there is an index token for consistency's sake, but it is fixed.

Definition 1 (Normalized exact-solution volume, NESV)**.** For a task T and transformer model $\mathcal{A}$, let $S_n^{T,\mathcal{A}}$ be the parameters for which the output is strictly correct on every length-n input, and on every index k. For a fixed parameter bound $B > 0$, define

$$V_B^{T,\mathcal{A}}(n) = \frac{\lambda_{D_\mathcal{A}}\big(S_n^{T,\mathcal{A}} \cap [-B,B]^{D_\mathcal{A}}\big)}{(2B)^{D_\mathcal{A}}}. \tag{1}$$

Equivalently, (1) is the probability of obtaining a correct solution when each scalar parameter is sampled independently and uniformly from $[-B, B]$. Unless stated otherwise, B and the hidden width of the FFNN, h, are fixed as n grows. Throughout, $\log$ denotes the natural logarithm; $\|\cdot\|_2$ and $\|\cdot\|_\infty$ denote the Euclidean and maximum-coordinate norms. Throughout, $c, C > 0$ and the constants implicit in the subscripted asymptotic notation such as $\Theta_{B,h}$ may depend on B and h, but never on n; c, C may change from line to line.

2.2 TRANSFORMER MODELS

We consider only single-layer transformers. Let x_j denote the j-th data token and let z denote the index token. The query, key, and value maps have the form $Q(x) = W_Q x + b_Q, K(x) = W_K x + b_K, V(x) = W_V x + b_V$. Writing $q = Q(z), k_j = K(x_j), v_j = V(x_j)$, the attention output is

$$\alpha_j = \frac{\exp\big((\log n) q^\top k_j\big)}{\sum_{i=1}^{n} \exp((\log n) q^\top k_i)}, \qquad A = \sum_{j=1}^{n} \alpha_j u_j. \tag{2}$$

The $\log n$ factor was proposed by Chiang & Cholak (2022) and Nakanishi (2025) and has been used in practice in Qwen (Bai et al., 2023). Chen et al. (2026) provide additional theoretical justification for this choice. In this paper, this choice is natural since we study positional tasks such as INDEX, where attention must continue to select a particular position as the number of competing positions grows with the input length. We omit the $1/\sqrt{d}$ factor in conventional $\log n$ scaling because d is constant in n and does not affect the asymptotic bounds in n.

The self-attention is followed by a feed-forward network (FFNN) classifier with hidden width h:

$$y = W_2 \,\mathrm{ReLU}(W_1 A + b_1) + b_2 \in \mathbb{R}^2, \tag{3}$$

where $W_1 \in \mathbb{R}^{h \times 2}, b_1 \in \mathbb{R}^h, W_2 \in \mathbb{R}^{2 \times h}, b_2 \in \mathbb{R}^2$.

The two coordinates of y are indexed by the labels $\{0,1\}$, and the predicted label is $\widehat{y} = \mathrm{argmax}_{c \in \{0,1\}}(y)_c$. When the input needs to be explicit, we write $A(w,k)$ and $y(w,k)$.

We do not use residual connections or layer normalization. Adding them could alter the size and geometry of the exact-solution set; we leave this extension for future work.

In this paper, we consider four models. We describe three of them here; the fourth, $\mathcal{A}_{\mathrm{norm}}$, we leave for Section 4.2.

In the **general** model $\mathcal{A}_{\mathrm{gen}}(h)$, the data tokens are $x_j = e_{w_j} \in \mathbb{R}^2$, and the index token is $z = 0_2$. The maps $Q, K, V : \mathbb{R}^2 \to \mathbb{R}^2$ are unrestricted affine maps, so $q = Q(0_2)$, $k_j = K(e_{w_j})$, and $v_j = V(e_{w_j})$. So $\mathcal{A}_{\mathrm{gen}}(h)$ has $D_{\mathrm{count}} = 5h + 20$ trainable scalar parameters.

We also examined a model, $\mathcal{A}_{\sin}(h)$, with a **sinusoidal** positional encoding $(\sin(j), \cos(j))$.

For the positional tasks, we use the **positional model** $\mathcal{A}_{\mathrm{pos}}(h)$. The data token at position j and the index token encoding k are

$$x_j^{\mathrm{pos}} = \begin{bmatrix} e_{w_j} \\ j \\ j^2 \end{bmatrix} \in \mathbb{R}^4, \qquad z_k^{\mathrm{pos}} = \begin{bmatrix} 0_2 \\ k \\ 0 \end{bmatrix} \in \mathbb{R}^4$$

(where 0_2 is the two-dimensional zero vector). For FIRST, the index is fixed at $k = 1$. For INDEX, the index $k \in [n]$ is supplied as part of the input. Because j and j^2 grow without bound in the input length, allowing them to propagate through the network would like lead to instability; positional-encoding choices are known to strongly affect length extrapolation (Kazemnejad et al., 2023). Accordingly, we constrain the model to have the form of the index-lookup construction of Bhattamishra et al. (2024):

$$\begin{aligned} W_Q^{\mathrm{pos}} &= [0_{2\times 2} \quad a \quad 0_2], & b_Q^{\mathrm{pos}} &= b, \\ W_K^{\mathrm{pos}} &= [0_{2\times 2} \quad r \quad s], & b_K^{\mathrm{pos}} &= c_K, \\ W_V^{\mathrm{pos}} &= [C_V \quad 0_{2\times 2}], & b_V^{\mathrm{pos}} &= b_V, \end{aligned}$$

where $a, b, r, s, c_K, b_V \in \mathbb{R}^2, C_V \in \mathbb{R}^{2\times 2}$ (and $0_{2\times 2}$ is the 2×2 zero matrix). This model has $D_{\mathrm{pos}} = 5h + 18$ trainable scalar parameters.

3 NESV BOUNDS AND LENGTH GENERALIZATION

We first establish the volume bounds for the four tasks and then compare such results with those of length-generalization experiments.

3.1 MAIN RESULT

Theorem 1 (NESV bounds)**.** *Fix $B > 0$ and $h \geq 2$. As $n \to \infty$,*

$$V_B^{\textsc{First},\mathcal{A}_{\mathrm{pos}}(h)}(n) = \Theta_{B,h}(1) \tag{4}$$

$$V_B^{\textsc{Index},\mathcal{A}_{\mathrm{pos}}(h)}(n) = \Theta_{B,h}(n^{-3}) \tag{5}$$

$$V_B^{\textsc{Majority},\mathcal{A}_{\mathrm{gen}}(h)}(n) = \Theta_{B,h}\big((n \log n)^{-1}\big) \qquad n \text{ odd} \tag{6}$$

$$V_B^{\textsc{Parity},\mathcal{A}_{\mathrm{gen}}(h)}(n) = 0 \qquad n > h + 1. \tag{7}$$

Proof sketch. For FIRST, one can choose parameters such that the attention score decreases quadratically away from position one. A fixed neighborhood of these parameters preserves that decrease, and the resulting concentration of attention lets the classifier return the first symbol for all sufficiently large lengths. The neighborhood does not shrink with n, thus ensuring a constant NESV bound.

However, the same argument cannot be used for INDEX, because the attention score must peak at $j = k$ for every possible k, and small errors that are harmless at one position can be amplified at another. We examine the two relevant error sources in Section 4.

In the position-free model $\mathcal{A}_{\mathrm{gen}}$, a word is represented only by its number of 1's. For MAJORITY, the output changes only once, between counts $(n-1)/2$ and $(n+1)/2$. These counts differ by just one, so the interval that separates them narrows as n increases. Consequently, placing the decision boundary between them requires increasingly precise parameter choice.

For PARITY, by contrast, every increment in the count must flip the output. A fixed number of hidden ReLU units cannot provide arbitrarily many flips in output. The upper bound of zero, meaning that PARITY is not solvable by the class of transformers considered here, is consistent with previous theoretical results (Kozachinskiy et al., 2026b) and experiments (Chiang & Cholak, 2022; Hahn & Rofin, 2024).

The complete proofs of the four bounds, including the technical details omitted from the INDEX proof sketch in Section 4.1, are given in Appendix A. □

For detailed results related to the sinusoidal model $\mathcal{A}_{\text{sin}}$, see Appendix C.

3.2 EXPERIMENT

To compare these bounds with empirical results, we train the FIRST, INDEX, and MAJORITY models with hidden width $h = 8$. We use 20 random seeds per task and draw a fresh batch at each update, sampling its length uniformly from $1, \ldots, N$; for MAJORITY, we use only odd lengths. The main results use $N = 51$. Checkpoints are selected using validation data at length N, without inspecting longer inputs, and tested on random examples at lengths up to 511. Additional experiments on $N = 11, 21, 31, 41$, which give similar results, are in Appendix D.

Table 1 shows that FIRST remains accurate at the longest test length (under $\mathcal{A}_{\text{pos}}$), MAJORITY retains substantial accuracy (under $\mathcal{A}_{\text{gen}}$), and INDEX falls much closer to chance (under $\mathcal{A}_{\text{pos}}$). These observations agree with the ranking of the three volume bounds.

We found the results of INDEX particularly interesting. Despite the task being expressible and learnable on lengths up to 51, it showed poor length generalization, much worse than that of MAJORITY. We take a closer look at this task in the next section.

4 THE INDEX TASK

We first revisit the volume argument from the previous section, to identify the parameter combinations responsible for the n-dependent decay. We then modify the model to control one of these combinations structurally and compare the resulting theoretical and empirical behavior.

4.1 LENGTH-SENSITIVE COMPONENTS OF THE ATTENTION SCORE

Recall that, in the positional model $\mathcal{A}_{\text{pos}}(h)$, $q = ak + b, k_j = rj + sj^2 + c_K, u_j = C_V e_{w_j} + b_V$. The attention score is

$$g_j = q^\top k_j = (ak + b)^\top (rj + sj^2 + c_K). \tag{8}$$

We compare the scores at position k and position $k + \Delta$. The expansion will produce the following quantities used repeatedly in later proofs, so we introduce some notations for future convenience.

$$\chi = a^\top s \qquad \ell = 2b^\top s + a^\top r \qquad \delta = b^\top s \qquad \phi = b^\top r. \tag{9}$$

Now consider position $k + \Delta \in [n]$. The *score gap* of this position is

$$g_{k+\Delta} - g_k = \Delta \underbrace{\left(2\chi k^2 + \ell k + \phi\right)}_{C_{\text{lin}}(k)} + \Delta^2 \underbrace{\left(\chi k + \delta\right)}_{C_{\text{quad}}(k)}. \tag{10}$$

We want this gap to be negative. The coefficient $C_{\text{quad}}(k)$ controls the curvature of the score around the indexed position, while $C_{\text{lin}}(k)$ measures its first-order drift and therefore whether the score remains centered at $j = k$; we want it to be zero. Because χ and ℓ are multiplied by k and k^2, small errors in these parameter combinations can be amplified at large k (i.e., at large n).

4.1.1 SKETCH OF LOWER BOUND

See Section A.2.1 for the full proof of the lower bound.

Let $\eta = \min\{B/16, 1/16\}$ and consider the solution (which we call the *center*):

$$a^0 = (2\eta, 0), \qquad b^0 = (0, -\eta), \qquad r^0 = (\eta, 0), \qquad s^0 = (0, \eta). \tag{11}$$

We want to find a region around the center that still solves the problem. The volume of this region will serve as our lower bound. At this center, $\chi = \ell = \phi = 0$, $\delta = -\eta^2$, and the score gap is

$$g_{k+\Delta} - g_k = -\eta^2 \Delta^2. \tag{12}$$

In a sufficiently small neighborhood of the center, we may restrict

$$|\chi| \le cn^{-2}, \qquad |\ell| \le cn^{-1}, \tag{13}$$

while allowing a, b, r_2, and s_2 (the second coordinates of r and s, resp.) to vary over fixed-width intervals independent of n. Every parameter choice in this region then satisfies $g_{k+\Delta} - g_k \le -\frac{\eta^2}{4}\Delta^2$ for every $k \in [n]$ and every nonzero Δ such that $k + \Delta \in [n]$. Under the $\log n$ attention scale, the total attention outside the indexed position then tends to zero as $n \to \infty$ uniformly in k. Hence, for all sufficiently large n, the indexed position receives at least a fixed fraction of the total attention, uniformly over all inputs and indexes. We may then restrict (C_V, b_V) and (W_1, b_1, W_2, b_2) to fixed-size neighborhoods, independent of n, such that, for some constant $c_0 > 0$ independent of n, $y_{w_k}(w, k) - y_{1-w_k}(w, k) \ge c_0$ for every input (w, k). To produce a exact solution, only the width of the interval χ and ℓ need to depend on n. Therefore the constructed set of exact solutions has normalized volume

$$\Omega(n^{-2}) \cdot \Omega(n^{-1}) = \Omega(n^{-3}). \tag{14}$$

4.1.2 Sketch of upper bound

See Section A.2.2 for the full proof of the upper bound.

For each indexed position k, compare a constant word with a word that differs from it only at position k. These two inputs must be assigned different labels. We consider two cases. Let $m = y_{w_k}(w, k) - y_{1-w_k}(w, k)$. In any correct solution, we will have $m > 0$, but if $m \le n^{-5}$, we call the solution *small-margin*; otherwise, we call it *large-margin*.

Take any one of the inequalities $y_{w_k}(w, k) > y_{1-w_k}(w, k)$; there are $4n$ such inequalities in total. Fix arbitrary values of $a, b, r, s, c_K, C_V, b_V, W_1, b_1, W_2$, and $b_{2,1-w_k}$. As b_{2,w_k} varies, the margin of the inequality varies with slope one. To satisfy $0 < y_{w_k}(w, k) - y_{1-w_k}(w, k) \le n^{-5}$, b_{2,w_k} needs to lie in an interval of length n^{-5}. Thus the set of small-margin solutions has normalized volume at most $4n \cdot n^{-5}/(2B) = O(n^{-4})$.

For every large-margin solution, changing only the symbol at position k from τ to $\sigma \ne \tau$ must change the predicted label from τ to σ. By the definition of a large-margin solution, for each of the two words w in this comparison, $y_{w_k}(w, k) - y_{1-w_k}(w, k) > n^{-5}$. Since the $y_\sigma(w, k) - y_\tau(w, k)$, viewed as a function of the attention output $A(w, k)$ is Lipschitz, this is impossible if the attention assigned to position k is too small. Hence the indexed position receives at least an inverse polynomial amount of attention. The softmax formula in (2) then implies that g_k lies within a constant of the maximum score. Comparing it with the two neighboring scores in (10) gives $C_{\text{quad}}(k) + |C_{\text{lin}}(k)| \le M_0$, where M_0 is independent of n. Since every coordinate of a, b, s lies in $[-B, B]$, both $\chi = a^\top s$ and $\delta = b^\top s$ are bounded independently of n. Hence, for $k \le n$, $|C_{\text{quad}}(k)| = |\chi k + \delta| = O(n)$, and thus $|C_{\text{lin}}(k)| \le M_0 + |C_{\text{quad}}(k)| = O(n)$. Let $M = \lfloor n/4 \rfloor$. Since $C_{\text{lin}}(k) = 2\chi k^2 + \ell k + \phi$, we have $C_{\text{lin}}(3M) - 2C_{\text{lin}}(2M) + C_{\text{lin}}(M) = 4\chi M^2$. Because each term on the left is $O(n)$ and $M = \Theta(n)$, this gives $|\chi| = O(n^{-1})$. Moreover, $C_{\text{lin}}(2M) - C_{\text{lin}}(M) = 6\chi M^2 + \ell M$. Using $|\chi| = O(n^{-1})$ in this identity gives $|\ell| = O(1)$. Repeating the same argument gives

$$|\chi| = O(n^{-2}), \qquad |\ell| = O(n^{-1}). \tag{15}$$

We used the fact

$$\Pr\left(|a^\top s| \le \varepsilon_1,\ |2b^\top s + a^\top r| \le \varepsilon_2\right) = O_B(\varepsilon_1 \varepsilon_2) \tag{16}$$

whenever $0 < \varepsilon_1 \le \varepsilon_2$ are sufficiently small (for detailed proof see Section A.2.2, Lemma 3). Apply this fact to (15) and adding the $O(n^{-4})$ small-margin volume would gives the upper bound $O(n^{-3})$.

4.1.3 Experimental analysis

We next check whether $\chi = a^\top s$ and $\ell = 2b^\top s + a^\top r$ identified by the volume proof track the behavior of trained models. In (10), χ controls the error term whose effect grows quadratically with k, while ℓ controls the error term whose effect grows linearly with k. For a model trained up to

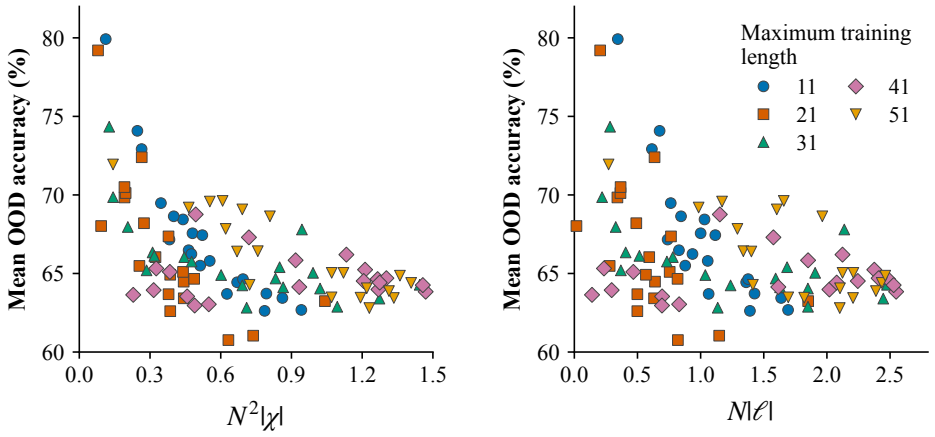

Figure 1: Scale-normalized drift coefficients and mean OOD accuracy for trained positional INDEX models. Markers indicate the maximum training length N.

maximum length N, we therefore consider the quantities $N^2|\chi|$ and $N|\ell|$, allowing models trained at different maximum lengths to be comparable. Across the 100 positional models trained at five maximum lengths, smaller values of both quantities are associated with higher mean accuracy on unseen lengths; see Fig. 1. Ranking the models separately within each maximum training length gives correlations of -0.655 for $N^2|\chi|$ and -0.516 for $N|\ell|$. Both have $p < 10^{-3}$ under permutation tests that preserve the five training-length groups. The training and evaluation protocol is given in Sections D.2 and D.3, and the coefficient analysis is detailed in Section D.4.

The above analysis suggests a direct architectural change: instead of relying on training to place two free coefficients inside shrinking intervals, control one of the length-growing error terms through the model itself.

4.2 THE NORMALIZED MODEL

We introduce a modified positional model, denoted $\mathcal{A}_{\mathrm{norm}}(h)$. It uses the same input vectors, value map, ReLU classifier, and trainable positional parameters a, b, r, s, c_K as $\mathcal{A}_{\mathrm{pos}}(h)$. We split the computation of the attention score into two *channels*: a linear channel that does not depend on the j^2 coordinate of the keys' positional encoding, and a quadratic channel that depends only on the j^2 coordinate. In the quadratic channel, we modify the queries as follows. Define the fixed smooth map

$$F(v) = \frac{v}{\sqrt{1 + \|v\|_2^2}} \qquad (v \in \mathbb{R}^2). \tag{17}$$

The linear and quadratic positional channels are

$$q^{\mathrm{lin}} = ak + b, \qquad q^{\mathrm{quad}} = b + F(ak), \qquad k_j^{\mathrm{lin}} = rj + c_K, \qquad k_j^{\mathrm{quad}} = sj^2. \tag{18}$$

Thus the linear channel uses the full query $ak + b$, whereas the quadratic channel applies F to the index-dependent contribution ak. The attention score is

$$g_j^{\mathrm{norm}} = \left(q^{\mathrm{lin}}\right)^\top k_j^{\mathrm{lin}} + \left(q^{\mathrm{quad}}\right)^\top k_j^{\mathrm{quad}} = (ak + b)^\top (rj + c_K) + \left(b + F(ak)\right)^\top sj^2. \tag{19}$$

The attention weights are obtained by applying the same $\log n$-scaled softmax to (19); the values and the one-hidden-layer ReLU classifier are unchanged. The normalization introduces no new trainable parameters, so $\mathcal{A}_{\mathrm{norm}}(h)$ has the same parameter dimension and uses the same B as $\mathcal{A}_{\mathrm{pos}}(h)$.

This modification could be applied to a general transformer by excluding j^2 from the position encoding at position j but adding an additional term $f(q_i, k_j)j^2$ directly into the attention score, with f being a scalar-valued function whose output remains bounded. We leave exploration of this more general setting for future work.

The map F is approximately the identity near the origin, but its output has norm strictly smaller than one. Consequently, the component of the query that depends on k and is multiplied with j^2 no

longer grows linearly with k. To make this precise, suppose first that $a \neq 0$ and write $\widehat{a} = a/\|a\|_2$. Retaining δ and ϕ from (9), define

$$\widehat{\chi} := \widehat{a}^\top s, \qquad \widehat{\ell} := \ell + 2\widehat{\chi}, \qquad \varepsilon_\chi(k) := \big(F(ak) - \widehat{a}\big)^\top s. \tag{20}$$

Then the modified quadratic and linear coefficients are

$$\begin{aligned} C^{\mathrm{norm}}_{\mathrm{quad}}(k) &= \widehat{\chi} + \delta + \varepsilon_\chi(k), \\ C^{\mathrm{norm}}_{\mathrm{lin}}(k) &= \widehat{\ell} k + \phi + 2k\varepsilon_\chi(k). \end{aligned} \tag{21}$$

Thus,

$$g^{\mathrm{norm}}_{k+\Delta} - g^{\mathrm{norm}}_k = \Delta C^{\mathrm{norm}}_{\mathrm{lin}}(k) + \Delta^2 C^{\mathrm{norm}}_{\mathrm{quad}}(k). \tag{22}$$

Moreover,

$$|\varepsilon_\chi(k)| \le \frac{\|s\|_2}{2k^2\|a\|_2^2}, \qquad |2k\varepsilon_\chi(k)| \le \frac{\|s\|_2}{k\|a\|_2^2}. \tag{23}$$

Thus the k^2 drift in (10) is replaced by a linear term and a bounded error $2k\varepsilon_\chi(k) = O(k^{-1})$. Only the coefficient $\widehat{\ell}$ remains to be confined to an interval of width $O(n^{-1})$.

Theorem 2 (NESV of INDEX under $\mathcal{A}_{\mathrm{norm}}(h)$)**.** *For $B > 0$ and $h \ge 2$.*

$$V_B^{\textsc{Index},\mathcal{A}_{\mathrm{norm}}(h)}(n) = \Theta_{B,h}(n^{-1}). \tag{24}$$

Proof. The proof follows the same strategy as that of the INDEX bounds for the positional model in Section 4.1. The full proof is given in Appendix B. □

Compared with the original positional model, selective normalization removes one shrinking direction from the volume calculation. In $\mathcal{A}_{\mathrm{pos}}(h)$, the two effective constraints have widths n^{-2} and n^{-1}; in $\mathcal{A}_{\mathrm{norm}}(h)$, the saturation of the quadratic query channel leaves only the n^{-1} constraint. The improvement from $\Theta(n^{-3})$ to $\Theta(n^{-1})$ therefore follows from the same score-level mechanism identified by the original proof.

4.3 LENGTH-GENERALIZATION EXPERIMENTS

We train $\mathcal{A}_{\mathrm{pos}}(h)$ and $\mathcal{A}_{\mathrm{norm}}(h)$ under the same protocol, using paired seeds, the same training batches, and identical initial values for all parameters shared by the two models. The main comparison uses classifier width $h = 8$, maximum training length $N = 51$, and 20 seeds.

As shown in Fig. 2, the original model degrades rapidly beyond the training range, whereas the selectively normalized model retains substantially higher accuracy. At length 511, approximately ten times the maximum training length, mean accuracy increases from 60% for $\mathcal{A}_{\mathrm{pos}}(h)$ to approximately 85% for $\mathcal{A}_{\mathrm{norm}}(h)$. Experiments with maximum training lengths $N \in \{11, 21, 31, 41\}$ show similar results and are reported in Section D.5.

Taken together, the coefficient analysis and the architectural intervention give two complementary forms of evidence. The original volume proof identifies the parameter combinations whose effects grow with the index k, and those combinations are associated with the behavior of trained models. Selectively normalizing the quadratic query channel removes the faster-growing contribution, enlarges the exact-solution volume from $\Theta(n^{-3})$ to $\Theta(n^{-1})$, and improves extrapolation in the model for which the volume calculation is carried out.

5 RELATED WORK

The INDEX task has been studied many times before. Bhattamishra et al. (2024) give an efficient one-layer construction in which the data appear before the indexed INDEX, as in our right-hand INDEX task. This differs from left-hand INDEXing, where the INDEX appears before the data and the answer is produced later (Kozachinskiy et al., 2025; 2026a). One-layer constructions for related right-hand function-evaluation problems are also studied by Strobl et al. (2025).

A line of work related to our normalized model $\mathcal{A}_{\mathrm{norm}}$ controls attention by normalizing query–key interactions. QKNorm ℓ_2-normalizes complete query and key vectors and learns the attention

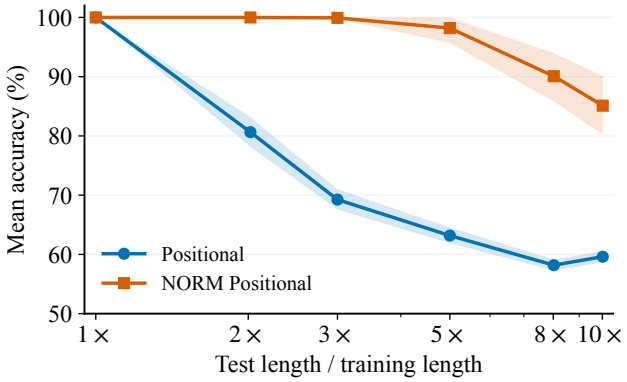

Figure 2: Length generalization on INDEX for the positional and $\mathcal{A}_{\mathrm{norm}}(h)$ models trained up to $N = 51$. Shading shows 95% bootstrap confidence intervals.

scale (Henry et al., 2020), while scaled cosine attention likewise removes dependence on their magnitudes (Liu et al., 2022). Separately, approaches to length extrapolation often modify the positional contribution to attention, for example by adding distance-dependent score biases (Press et al., 2022) or by designing relative position encoding to improve attention resolution beyond the training lengths (Sun et al., 2023). Our $\mathcal{A}_{\mathrm{norm}}(h)$, on the other hand, does not normalize the full query or key vectors, nor the quadratic positional coordinate j^2. Instead, it applies a fixed smooth bounded map to the index-dependent query component in the quadratic positional channel.

Another related line of work studies which functions occupy large or stable regions of transformer parameter space. Hahn & Rofin (2024) connect Boolean sensitivity to transformer norms and perturbation behavior, while Kover et al. (2026) study the parameter-space geometry of transformers computing Boolean functions. Empirical work on simplicity bias likewise suggests that the parameter-to-function map favors some Boolean functions over others (Bhattamishra et al., 2023). Our volume measure tracks how volume changes with sequence length, rather than assigning a single complexity or sensitivity measure to the target task. For INDEX, such analysis not only shows that the volume decreases with longer input lengths, but also isolates the sources of such shrinkage.

6 DISCUSSION AND LIMITATIONS

Initialization and optimization Parameter-space volume depends on the distribution placed on the parameters. Equation (1) assumes a uniform distribution over a box of size $2B$ in the raw coordinates. But initialization and optimization can concentrate strongly on particular parts of parameter space, so differences in exact-solution volume might not translate exactly into differences in accuracy. Tightening the correspondence would therefore require combining the present analysis with the distributions over parameters induced by initialization and optimization.

Depth Our theoretical results concern fixed-width, one-layer, one-head transformers. This restriction can matter qualitatively. Bhattamishra et al. (2024) show examples where two-layer transformers of logarithmic size solve problems for which one-layer transformers require linear size. The parameter constraints governing length generalization may therefore change substantially with depth, and extending the volume analysis to deeper transformers is an important next step.

7 CONCLUSION

Parameter-space volume is useful not only for measuring how exact solutions change with input length, but also for revealing *why* they change. On the INDEX task, our volume analysis isolates the length-sensitive components, suggesting a modification to the model itself that improves length extrapolation. This suggests a broader use of volume analysis as a diagnostic for transformer model refinement, especially task-specific ones.

Reproducibility Statement

The baseline architectures, parameter prior, and strict-correctness criterion are defined in Section 2. The $\mathcal{A}_{\mathrm{norm}}(h)$ architecture is defined in Section 4.2, and the fixed-sinusoidal MAJORITY model is defined in Section C.1. The complete volume arguments, experimental protocol, checkpoint-selection rule, evaluation-bank construction, and supplementary results are presented in the appendix.

AI Use Statement

In this work, we used generative AI tools to aid literature exploration, assist in the writing of proofs (prose editing, LaTeX preparation, proofreading), create or modify scientific figures, and edit code (for experiments). We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, or assist with translation. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/pdf?id=B1J_rgWRW.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Satwik Bhattamishra, Arkil Patel, Varun Kanade, and Phil Blunsom. Simplicity bias in transformers and their ability to learn sparse Boolean functions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 5767–5791, 2023. doi:10.18653/v1/2023.acl-long.317.

Satwik Bhattamishra, Michael Hahn, Phil Blunsom, and Varun Kanade. Separations in the representational capabilities of transformers and recurrent architectures. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi:10.52202/079017-1135.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=7SLtElfqCW.

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*, pp. 7654–7664, 2022. doi:10.18653/v1/2022.acl-long.527.

Ping-yeh Chiang, Renkun Ni, David Yu Miller, Arpit Bansal, Jonas Geiping, Micah Goldblum, and Tom Goldstein. Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=QC10RmRbZy9.

Michael Hahn and Mark Rofin. Why are sensitive functions hard for transformers? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 14973–15008, 2024. doi:10.18653/v1/2024.acl-long.800.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020. doi:10.18653/v1/2020.findings-emnlp.379.

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, pp. 24892–24928, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4e85362c02172c0c6567ce593122d31c-Abstract-Conference.html.

Blanka Kover, Alexandra Butoi, Anej Svete, Michael Hahn, and Ryan Cotterell. Understanding the parameter space geometry of transformers encoding boolean functions. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/pdf/2606.08768.

Alexander Kozachinskiy, Felipe Urrutia, Hector Orellana, Tomasz Steifer, German Pizarro, Matias Fuentes, Francisco Meza Vasquez, Cristian Buc Calderon, and Cristobal Rojas. Strassen attention, split VC dimension and compositionality in transformers. In *Advances in Neural Information Processing Systems*, volume 38, pp. 10718–10749, 2025. doi:10.52202/085713-0327.

Alexander Kozachinskiy, Vicente Opazo, and Felipe Urrutia. Indexing: The beginning and the end, 2026a. URL https://arxiv.org/abs/2607.22361.

Alexander Kozachinskiy, Tomasz Steifer, and Przemyslaw Walega. Parity, sensitivity, and transformers, 2026b. URL https://arxiv.org/abs/2602.05896.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12009–12019, 2022. doi:10.1109/CVPR52688.2022.01170.

Ken M. Nakanishi. Scalable-softmax is superior for attention, 2025. URL https://arxiv.org/abs/2501.19399.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=R8sQPpGCv0.

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.

Lena Strobl, Dana Angluin, and Robert Frank. Concise one-layer transformers can do function evaluation (sometimes), 2025. URL https://arxiv.org/abs/2503.22076.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14590–14604, 2023. doi:10.18653/v1/2023.acl-long.816.

A Proofs of the Baseline NESV Bounds

This appendix proves the four baseline bounds in Theorem 1.

A.1 First

The lower bound constructs fixed neighborhoods of parameters independent of input length that gives exact solutions to the task. The upper bound follows from symmetry between the two labels.

Proof of the FIRST *bound in Theorem 1.* Choose $\eta = \min\{B/16, 1/16\}$ and set

$$a^0 = (2\eta, 0), \qquad b^0 = (0, -\eta), \qquad r^0 = (\eta, 0), \qquad s^0 = (0, \eta), \tag{25}$$

with $c_K^0 = b_V^0 = 0$ and $C_V^0 = \eta I_2$. We call this set of (a^0, b^0, r^0, s^0) the center. Put $q_F := a + b$, $\widehat{\ell}_1 := q_F^\top r$, $\widehat{\ell}_2 := q_F^\top s$. At the center, $\widehat{\ell}_1^0 = 2\eta^2$ and $\widehat{\ell}_2^0 = -\eta^2$. By continuity, there is a fixed positive-volume neighborhood of (a, b, r, s) on which

$$|\widehat{\ell}_1 - 2\eta^2| \le \frac{\eta^2}{8}, \qquad |\widehat{\ell}_2 + \eta^2| \le \frac{\eta^2}{8}. \tag{26}$$

For $j = 1 + \Delta$ with $\Delta \ge 1$, the key bias cancels and

$$\begin{aligned} g(1) - g(1+\Delta) &= \Delta\big[(-\widehat{\ell}_1 - 2\widehat{\ell}_2) + (-\widehat{\ell}_2)\Delta\big] \\ &\ge \frac{\eta^2}{2}\Delta^2. \end{aligned} \tag{27}$$

Consequently, throughout this fixed positional-parameter neighborhood,

$$1 - \alpha_1 \le \sum_{\Delta=1}^{\infty} n^{-(\eta^2/2)\Delta^2} \longrightarrow 0. \tag{28}$$

Restrict (C_V, b_V) to a fixed neighborhood on which

$$\|C_V e_\sigma + b_V - \eta e_\sigma\|_\infty \le \frac{\eta}{16} \qquad (\sigma \in \{0, 1\}). \tag{29}$$

For all sufficiently large n, (28) gives $\alpha_1 \ge 7/8$. If $y^\star = w_1$ and $\bar{y}^\star = 1 - y^\star$, the total attention masses on the two symbols satisfy

$$p_{y^\star} - p_{\bar{y}^\star} \ge \frac{3}{4}. \tag{30}$$

Convexity of the value perturbation in (29) therefore gives

$$A_{y^\star} - A_{\bar{y}^\star} \ge \eta\big(p_{y^\star} - p_{\bar{y}^\star}\big) - \frac{\eta}{8} \ge \frac{5\eta}{8}. \tag{31}$$

Choose $W_1^0, b_1^0, W_2^0, b_2^0$ as follows. Take the first two rows of W_1^0 to be $\eta e_0^\top$ and $\eta e_1^\top$, with both corresponding biases equal to η^2. Set every remaining row of W_1^0 to zero and its bias to $-\eta$. In W_2^0, give output row zero coefficient η on hidden unit zero and output row one coefficient η on hidden unit one, with all other entries zero; set $b_2^0 = 0$. The first two hidden units are active throughout a sufficiently small neighborhood of (29), whereas the remaining units are inactive, and

$$y_0 - y_1 = \eta^2(A_0 - A_1). \tag{32}$$

Thus (31) gives a correct-logit margin of at least that $y_0 - y_1 \ge 5\eta^3/8$ at the center. Allow c_K to vary over a sufficiently small fixed neighborhood of 0. This does not affect the attention weights, since $q_F^\top c_K$ is added to every attention score. Together with the fixed neighborhoods chosen above for (a, b, r, s), (C_V, b_V), and (W_1, b_1, W_2, b_2), this gives a set of normalized volume $c_F > 0$, independent of n.

For the upper bound, swap the two rows of W_2 and the two entries of b_2, leaving every other coordinate unchanged. This is a measure-preserving involution of the raw parameter cube. It maps the exact solutions of FIRST to the exact solutions of the complementary task $w \mapsto 1 - w_1$. The two exact-solution sets are disjoint, so each occupies at most one half of the cube. In particular, the NESV is $O(1)$. □

A.2 INDEX

We use the notation from Section 4.1. Thus the attention score at position j is

$$g_j = (ak + b)^\top (rj + sj^2 + c_K), \tag{33}$$

and

$$\chi = a^\top s, \qquad \ell = 2b^\top s + a^\top r, \qquad \delta = b^\top s, \qquad \phi = b^\top r. \tag{34}$$

For $k + \Delta \in [n]$, the *score gap* at position $k + \Delta$ is

$$g_{k+\Delta} - g_k = \Delta C_{\mathrm{lin}}(k) + \Delta^2 C_{\mathrm{quad}}(k), \tag{35}$$

where

$$C_{\mathrm{lin}}(k) = 2\chi k^2 + \ell k + \phi, \qquad C_{\mathrm{quad}}(k) = \chi k + \delta. \tag{36}$$

A.2.1 LOWER BOUND

Choose $\eta = \min\{B/16, 1/16\}$ and consider the solution (which we call the *center*):

$$a^0 = (2\eta, 0), \qquad b^0 = (0, -\eta), \qquad r^0 = (\eta, 0), \qquad s^0 = (0, \eta). \tag{37}$$

At this point,

$$\chi = \ell = \phi = 0, \qquad \delta = -\eta^2, \tag{38}$$

and hence

$$g_k - g_j = \eta^2 (j - k)^2. \tag{39}$$

To obtain a positive-volume family around this point, consider the local coordinate map

$$\Psi(a, b, r, s) = (\chi, \ell, a_1, a_2, b_1, b_2, r_2, s_2). \tag{40}$$

We call (a, b, r, s) the untransformed coordinates and $(\chi, \ell, a_1, a_2, b_1, b_2, r_2, s_2)$ the transformed coordinates. For fixed (a, b, r_2, s_2), the Jacobian of $(s_1, r_1) \mapsto (\chi, \ell)$ is

$$\begin{bmatrix} a_1 & 0 \\ 2b_1 & a_1 \end{bmatrix}, \qquad \det = a_1^2. \tag{41}$$

The determinant is nonzero at (37). The inverse function theorem therefore gives a fixed neighborhood on which Ψ is one-to-one and the Jacobians of Ψ and Ψ^{-1} are bounded. Hence there are constants $c, C > 0$, independent of n, such that every measurable set E in the transformed-coordinate neighborhood satisfies $c\,\mathrm{vol}(E) \le \mathrm{vol}\big(\Psi^{-1}(E)\big) \le C\,\mathrm{vol}(E)$.

Restrict

$$|\chi| \le \frac{\eta^2}{64n^2}, \qquad |\ell| \le \frac{\eta^2}{64n}, \tag{42}$$

while restricting the remaining transformed coordinates to fixed-width intervals independent of n sufficiently close to the center that

$$\delta \le -\frac{3\eta^2}{4}, \qquad |\phi| \le \frac{\eta^2}{16}. \tag{43}$$

For every $k \le n$, these conditions imply

$$C_{\mathrm{quad}}(k) \le -\frac{\eta^2}{2}, \qquad |C_{\mathrm{lin}}(k)| \le \frac{\eta^2}{4}. \tag{44}$$

Indeed, the contribution of χk to $C_{\mathrm{quad}}(k)$ is at most $\eta^2/64$, and the three contributions to $C_{\mathrm{lin}}(k)$ are bounded by $2\eta^2/64$, $\eta^2/64$, and $\eta^2/16$, respectively. Therefore, for every nonzero integer Δ with $k + \Delta \in [n]$,

$$g_k - g_{k+\Delta} = -\Delta C_{\mathrm{lin}}(k) - \Delta^2 C_{\mathrm{quad}}(k) \qquad \ge \frac{\eta^2}{4}\Delta^2. \tag{45}$$

Under the $\log n$ attention scale, the total attention outside the indexed position is consequently at most

$$2\sum_{d=1}^{\infty} n^{-(\eta^2/4)d^2}, \tag{46}$$

which tends to zero as $n \to \infty$ uniformly in the index k and over all positional parameters (a, b, r, s) in the local neighborhood above satisfying (42) and (43).

Set $C_V^0 = \eta I_2$, $b_V^0 = 0$, and take the parameter values

$$W_1^0 = \begin{bmatrix} \eta I_2 \\ 0_{(h-2)\times 2} \end{bmatrix}, \qquad b_1^0 = (\eta^2, \eta^2, -\eta, \ldots, -\eta)^\top,$$

$$W_2^0 = \begin{bmatrix} \eta I_2 & 0_{2\times(h-2)} \end{bmatrix}, \qquad b_2^0 = 0.$$

For all sufficiently large n, (46) makes the indexed position receive at least $3/4$ of the total attention. If $w_k = \sigma$, the σ coordinate of the attention output exceeds the other coordinate by at least $\eta/2$, even if every other token has symbol $1 - \sigma$. With the chosen parameters, the first two hidden units are active and, for $c \in \{0, 1\}$, the c-th unit outputs $\eta A_c + \eta^2$. The corresponding row of W_2 multiplies

this quantity by η, so $y_c = \eta^2 A_c + \eta^3$. The common term η^3 cancels between the two logits, and therefore $y_\sigma - y_{1-\sigma} = \eta^2(A_\sigma - A_{1-\sigma}) \geq \frac{\eta^3}{2}$. This margin is uniform over all inputs and indexes, so there exists fixed neighborhoods independent of n of $C_V, b_V, W_1, b_1, W_2, b_2$ that produce exact solutions. The key bias c_K may also vary over a fixed neighborhood because it cancels from the attention weights.

The two intervals in (42) contribute normalized widths of orders n^{-2} and n^{-1}, whereas all other coordinates vary over intervals of width independent of n. Therefore

$$V_B^{\text{INDEX},\mathcal{A}_{\text{pos}}(h)}(n) = \Omega_{B,h}(n^{-3}). \tag{47}$$

A.2.2 Upper bound

For each indexed position k, compare a constant word with a word that differs from it only at position k. These two inputs must be assigned different labels. We consider two cases. Let $m = y_{w_k}(w,k) - y_{1-w_k}(w,k)$. In any correct solution, we will have $m > 0$, but if $m \leq n^{-5}$, we call the solution *small-margin*; otherwise, we call it *large-margin*.

Take any one of the inequalities $y_{w_k}(w,k) > y_{1-w_k}(w,k)$; there are $4n$ such inequalities in total. Fix arbitrary values of $a, b, r, s, c_K, C_V, b_V, W_1, b_1, W_2$, and $b_{2,1-w_k}$. As b_{2,w_k} varies, the margin of the inequality varies with slope one. To satisfy $0 < y_{w_k}(w,k) - y_{1-w_k}(w,k) \leq n^{-5}$, b_{2,w_k} needs to lie in an interval of length n^{-5}. Thus the set of small-margin solutions has normalized volume at most $4n \cdot n^{-5}/(2B) = O(n^{-4})$.

For every large-margin solution, changing only the symbol at position k from τ to $\sigma \neq \tau$ must change the predicted label from τ to σ. By the definition of a large-margin solution, for each of the two words w in this comparison, $y_{w_k}(w,k) - y_{1-w_k}(w,k) > n^{-5}$. Since the $y_\sigma(w,k) - y_\tau(w,k)$, viewed as a function of the attention output $A(w,k)$ is Lipschitz, this is impossible if the attention assigned to position k is too small. To see this, let w be the constant-τ word and let w' be the word obtained by changing only position k to σ. Because the query and key maps read only positional coordinates, the two words have the same attention weights. Their attention outputs therefore satisfy

$$A(w',k) - A(w,k) = \alpha_k C_V(e_\sigma - e_\tau), \tag{48}$$

so

$$\|A(w',k) - A(w,k)\|_\infty \leq 2B\alpha_k. \tag{49}$$

Moreover, ReLU is 1-Lipschitz, each row of W_1 has ℓ_1-norm at most $2B$, and the difference between rows σ and τ of W_2 has ℓ_1-norm at most $2hB$. Hence

$$\begin{aligned}&\big|\big(y_\sigma(w',k) - y_\tau(w',k)\big) - \big(y_\sigma(w,k) - y_\tau(w,k)\big)\big| \\ &\qquad \leq 4hB^2\|A(w',k) - A(w,k)\|_\infty \leq 8hB^3\alpha_k.\end{aligned} \tag{50}$$

The first difference is greater than n^{-5} and the second is less than $-n^{-5}$, so

$$\alpha_k \geq \frac{n^{-5}}{4hB^3}. \tag{51}$$

Hence the indexed position receives at least an inverse polynomial amount of attention. The softmax formula in (2) then implies that g_k lies within a constant of the maximum score. Indeed,

$$\frac{n^{-5}}{4hB^3} \leq \alpha_k = \frac{n^{g_k}}{\sum_{j=1}^n n^{g_j}} \leq n^{g_k - \max_j g_j}, \tag{52}$$

and therefore

$$\max_j g_j - g_k \leq 5 + \frac{\log(4hB^3)}{\log n} \leq M_0 \tag{53}$$

for all sufficiently large n, where M_0 is independent of n. For every interior index $2 \leq k \leq n-1$, comparing g_k with the two neighboring scores in (10) gives

$$g_{k+1} - g_k = C_{\text{lin}}(k) + C_{\text{quad}}(k) \leq M_0 \tag{54}$$

and

$$g_{k-1} - g_k = -C_{\text{lin}}(k) + C_{\text{quad}}(k) \leq M_0. \tag{55}$$

Thus

$$C_{\mathrm{quad}}(k) + |C_{\mathrm{lin}}(k)| \le M_0, \tag{56}$$

where M_0 is independent of n. Since every coordinate of a, b, s lies in $[-B, B]$, both $\chi = a^\top s$ and $\delta = b^\top s$ are bounded independently of n. Hence, for $k \le n$, $|C_{\mathrm{quad}}(k)| = |\chi k + \delta| = O(n)$, and thus $|C_{\mathrm{lin}}(k)| \le M_0 + |C_{\mathrm{quad}}(k)| = O(n)$. Let $M = \lfloor n/4 \rfloor$. Since $C_{\mathrm{lin}}(k) = 2\chi k^2 + \ell k + \phi$, we have $C_{\mathrm{lin}}(3M) - 2C_{\mathrm{lin}}(2M) + C_{\mathrm{lin}}(M) = 4\chi M^2$. Because each term on the left is $O(n)$ and $M = \Theta(n)$, this gives $|\chi| = O(n^{-1})$. Moreover, $C_{\mathrm{lin}}(2M) - C_{\mathrm{lin}}(M) = 6\chi M^2 + \ell M$. Using $|\chi| = O(n^{-1})$ in this identity gives $|\ell| = O(1)$. These preliminary bounds make

$$|C_{\mathrm{quad}}(k)| = |\chi k + \delta| = O(1) \tag{57}$$

uniformly for $k \le n$. Equation (56) then gives $|C_{\mathrm{lin}}(k)| = O(1)$. Repeating the same argument, now using these improved bounds, gives

$$|\chi| = O(n^{-2}), \qquad |\ell| = O(n^{-1}). \tag{58}$$

We used the fact $\Pr\left(|a^\top s| \le \varepsilon_1,\ |2b^\top s + a^\top r| \le \varepsilon_2\right) = O_B(\varepsilon_1\varepsilon_2)$ whenever $0 < \varepsilon_1 \le \varepsilon_2$ are sufficiently small.

Lemma 3 (Joint two-slab estimate)**.** *Let a, b, r, s be independent and uniform on $[-B, B]^2$. There are constants $C_{\mathrm{slab}}, \varepsilon_B > 0$, depending only on B, such that, whenever $0 < \varepsilon_1 \le \varepsilon_2 \le \varepsilon_B$,*

$$\Pr\left(|a^\top s| \le \varepsilon_1,\ |2b^\top s + a^\top r| \le \varepsilon_2\right) \le C_{\mathrm{slab}}\varepsilon_1\varepsilon_2. \tag{59}$$

Proof. Scaling changes only the constant, so take $B = 1$. Fix (a, s), and set

$$\varrho = (\|a\|_2^2 + \|s\|_2^2)^{1/2}. \tag{60}$$

The second constraint defines a slab in (b, r) with normal vector $(2s, a)$. One coordinate of this vector has magnitude at least a constant multiple of ϱ, so, after fixing the other three coordinates, the allowed fraction of the remaining coordinate is at most

$$C \min\left\{1, \frac{\varepsilon_2}{\varrho}\right\}. \tag{61}$$

To estimate the first constraint at fixed radius ϱ. Write

$$a = \varrho\cos\vartheta\,\widehat{a}, \qquad s = \varrho\sin\vartheta\,\widehat{s}, \qquad 0 \le \vartheta \le \frac{\pi}{2}, \tag{62}$$

where $\widehat{a}, \widehat{s} \in \mathbb{S}^1$. If φ is the angle between $\widehat{a}$ and $\widehat{s}$, then

$$|a^\top s| = \varrho^2 \sin\vartheta\cos\vartheta\,|\cos\varphi|. \tag{63}$$

For fixed ϑ, the fraction of angle φ satisfying the first constraint is at most

$$C \min\left\{1, \frac{\varepsilon_1}{\varrho^2\sin\vartheta\cos\vartheta}\right\}. \tag{64}$$

The angular density contains the factor $\sin\vartheta\cos\vartheta$, therefore integrating over ϑ bounds the angular fraction by

$$C \min\left\{1, \frac{\varepsilon_1}{\varrho^2}\right\}. \tag{65}$$

because when $t \le 1$, $\int_0^{\pi/2} \min\{\sin\vartheta\cos\vartheta, t\}\,d\vartheta = O(t)$. The radial volume element for $(a, s) \in \mathbb{R}^4$is proportional to $\varrho^3 d\varrho$. Enlarging the cube to a fixed ball and combining (61) and (65), the required probability is at most a constant times

$$\int_0^{R_0} \varrho^3 \min\left\{1, \frac{\varepsilon_1}{\varrho^2}\right\} \min\left\{1, \frac{\varepsilon_2}{\varrho}\right\} d\varrho. \tag{66}$$

Assume $0 < \varepsilon_1 \le \varepsilon_2 \le 1$. If $\varepsilon_2 \le \sqrt{\varepsilon_1}$, split the integral at ε_2 and $\sqrt{\varepsilon_1}$ to obtain

$$\begin{aligned} I &\le \int_0^{\varepsilon_2} \varrho^3\,d\varrho + \varepsilon_2 \int_{\varepsilon_2}^{\sqrt{\varepsilon_1}} \varrho^2\,d\varrho + \varepsilon_1\varepsilon_2 \int_{\sqrt{\varepsilon_1}}^{R_0} d\varrho \\ &\le C\varepsilon_1\varepsilon_2. \end{aligned} \tag{67}$$

If $\sqrt{\varepsilon_1} \le \varepsilon_2$, split instead at $\sqrt{\varepsilon_1}$ and ε_2. Rescaling restores B and proves the lemma. □

Applying the above lemma to (58), the large-margin solutions have normalized volume $O(n^{-3})$: the entries of a, b, r, s are independent parameter coordinates, and all remaining coordinates integrate out. Adding the $O(n^{-4})$ small-margin volume gives

$$V_B^{\text{INDEX},\mathcal{A}_{\text{pos}}(h)}(n) = O_{B,h}(n^{-3}). \tag{68}$$

Together with (47), this proves the INDEX claim in Theorem 1.

A.3 MAJORITY

In the position-free model $\mathcal{A}_{\text{gen}}$, every input with the same number of ones produces the same attention output. The proof therefore reduces MAJORITY to a one-dimensional sequence indexed by the count and measures how much of the parameter space places the decision boundary between the two central counts. Let

$$t = \sum_{j=1}^{n} w_j, \qquad v_b = V(e_b), \qquad q = Q(0_2), \tag{69}$$

and define the query–key score contrast

$$X = q^\top W_K(e_1 - e_0), \qquad \Gamma = \exp((\log n)X). \tag{70}$$

All zero tokens share one score and all one tokens share another, so the total attention mass on the one tokens and the attention output are exactly

$$\rho_t = \frac{t\Gamma}{n - t + t\Gamma}, \qquad A_t = v_0 + \rho_t(v_1 - v_0). \tag{71}$$

Thus the 2^n inputs reduce to $n+1$ possible count levels. Throughout this subsection, write $n = 2m+1$. For an attention scale $\beta > 0$ and a score contrast x, let

$$\Delta\rho_{n,\beta}(x) = \rho_{m+1} - \rho_m, \tag{72}$$

where $\Gamma = e^{\beta x}$. The counts m and $m+1$ are the two adjacent levels at which the majority label changes.

Lemma 4 (Central attention gap)**.** *For $\Gamma = e^{\beta x}$,*

$$\Delta\rho_{n,\beta}(x) = \frac{ne^{\beta x}}{(m+1+me^{\beta x})(m+(m+1)e^{\beta x})} \tag{73}$$

$$= \frac{n}{n^2 + (n^2-1)\sinh^2(\beta x/2)}. \tag{74}$$

Moreover,

$$0 < \Delta\rho_{n,\beta}(x) \le \frac{1}{n}, \qquad \int_{\mathbb{R}} \Delta\rho_{n,\beta}(x)\,dx = \frac{4\,\operatorname{artanh}(1/n)}{\beta}. \tag{75}$$

Proof. Substitute $t = m, m+1$ into (71) and subtract to obtain (73). Dividing its denominator by $e^{\beta x}$ and using $n = 2m+1$ gives (74). The maximum is attained at $x = 0$. For the integral, set $z = \tanh(\beta x/2)$. Then

$$dx = \frac{2\,dz}{\beta(1-z^2)}, \qquad \Delta\rho_{n,\beta}(x) = \frac{n(1-z^2)}{n^2 - z^2}. \tag{76}$$

Therefore

$$\int_{\mathbb{R}} \Delta\rho_{n,\beta}(x)\,dx = \frac{2n}{\beta}\int_{-1}^{1} \frac{dz}{n^2 - z^2} = \frac{4\,\operatorname{artanh}(1/n)}{\beta}. \tag{77}$$

□

For fixed values of all parameters except $b_{2,1}$, exact correctness requires $b_{2,1}$ to place the decision threshold between the two central count levels. Define the bias-free logit difference

$$H(z) = (e_1 - e_0)^\top W_2\,\mathrm{ReLU}(W_1 z + b_1). \tag{78}$$

Lemma 5 (Output-bias fiber)**.** *Condition on every raw parameter except $b_{2,1}$. The set of values of $b_{2,1}$ that can extend the conditioned coordinates to an exact* MAJORITY *solution has length at most*

$$8hB^3\Delta\rho_{n,\beta}(X). \tag{79}$$

Proof. Correctness on the count-m and count-$(m+1)$ levels requires

$$H(A_m) + b_{2,1} - b_{2,0} < 0, \tag{80}$$
$$H(A_{m+1}) + b_{2,1} - b_{2,0} > 0. \tag{81}$$

Thus $b_{2,1}$ must lie in

$$\big(b_{2,0} - H(A_{m+1}),\ b_{2,0} - H(A_m)\big), \tag{82}$$

whose length is $H(A_{m+1}) - H(A_m)$ when this difference is positive. ReLU is 1-Lipschitz, each row of W_1 has ℓ_1-norm at most $2B$, and the difference between the two rows of W_2 has ℓ_1-norm at most $2hB$. Hence H is $4hB^2$-Lipschitz in ℓ_∞. Moreover, $A_{m+1} - A_m = \Delta\rho_{n,\log n}(X)(v_1 - v_0)$, so

$$\|A_{m+1} - A_m\|_\infty \le 2B\Delta\rho_{n,\beta}(X). \tag{83}$$

Multiplying the two bounds proves the claim. □

To average the fiber length over the query and key parameters, we use a uniform density bound for X.

Lemma 6 (Density of the score contrast)**.** *Under the uniform raw-cube measure, the random variable $X = q^\top W_K(e_1 - e_0)$ has a density f_X satisfying*

$$\|f_X\|_\infty \le B^{-2}. \tag{84}$$

Proof. Write $\Delta_K = W_K(e_1 - e_0)$. Each coordinate of Δ_K is the difference of two independent uniform variables and has triangular density bounded by $(2B)^{-1}$. Conditional on $q_{\mathrm{agg}} \neq 0$, choose a coordinate with magnitude $\|q_{\mathrm{agg}}\|_\infty$ and condition on the other coordinate of Δ_K. Scaling gives

$$\|f_{X|q_{\mathrm{agg}}}\|_\infty \le \frac{1}{2B\|q_{\mathrm{agg}}\|_\infty}. \tag{85}$$

Since $q_{\mathrm{agg}} = Q(0) = b_Q$ is uniform on $[-B, B]^2$,

$$\mathbb{E}\big[\|q_{\mathrm{agg}}\|_\infty^{-1}\big] = \frac{2}{B}. \tag{86}$$

Averaging proves the result. □

By Fubini's theorem and Lemmas 4 to 6,

$$\begin{aligned} V_B^{\mathrm{MAJORITY},\mathcal{A}_{\mathrm{gen}}(h)}(n) &\le 4hB^2\,\mathbb{E}\big[\Delta\rho_{n,\log n}(X)\big] \\ &\le 4hB^2\|f_X\|_\infty \int_{\mathbb{R}} \Delta\rho_{n,\log n}(x)\,dx \\ &\le \frac{16h\,\mathrm{artanh}(1/n)}{\log n}. \end{aligned} \tag{87}$$

Since $\mathrm{artanh}(1/n) \le n/(n^2-1)$, this is $O_h((n\log n)^{-1})$.

For the matching lower bound, we need a positive mass of score contrasts on which the central gap is of order $1/n$ and a fixed-volume set of the remaining parameters that turns this gap into a correct bias interval.

Lemma 7 (Lower contrast mass)**.** *There is a constant $c_X = c_X(B) > 0$ such that, for all sufficiently large n,*

$$\Pr\left(|X| \le \frac{1}{\log n}\right) \ge \frac{c_X}{\log n}. \tag{88}$$

Proof. Write $\Delta_K = W_K(e_1 - e_0)$ and $q_{\mathrm{agg}} = (q_1, q_2)$. Restrict to

$$q_1 \in [B/2, B], \qquad |q_2| \le B/4, \qquad |\Delta_{K,2}| \le B/2. \tag{89}$$

These independent events have probabilities at least $1/4$, $1/4$, and $3/8$. Conditional on them, the inequality

$$|q_1\Delta_{K,1} + q_2\Delta_{K,2}| \le \varepsilon \tag{90}$$

asks that $\Delta_{K,1}$ lie in an interval of length at least $2\varepsilon/B$, centered in $[-B/4, B/4]$. If $\varepsilon \le B^2/8$, the interval lies in $[-B/2, B/2]$, where the triangular density of $\Delta_{K,1}$ is at least $3/(8B)$. Hence

$$\Pr(|X| \le \varepsilon) \ge \frac{9\varepsilon}{512B^2}. \tag{91}$$

Taking $\varepsilon = 1/\log n$ proves the lemma. □

Lemma 8 (Positive-volume monotone decoder family)**.** *There are constants $c_\Omega, \kappa > 0$ and a set of values of $(W_V, b_V, W_1, b_1, W_2, b_{2,0})$ of normalized volume at least c_Ω such that the bias-free logit difference along $v_0 + p(v_1 - v_0)$ is affine with derivative at least κ for all $p \in [0, 1]$. For every parameter choice in this set, $b_{2,1}$ may vary over an interval of length at least*

$$\kappa(\rho_{m+1} - \rho_m) \tag{92}$$

while preserving exact MAJORITY *correctness.*

Proof. Let $e_V = (1, 0)^\top$ and choose W_V, b_V so that

$$v_0 = 0, \qquad v_1 = \eta e_V. \tag{93}$$

Give one hidden unit first-layer row $\eta e_V^\top$ and bias η, make every other hidden unit strictly inactive, and set the difference between output rows on the active hidden unit to η. Set the other output-row differences and $b_{2,0}$ to zero. Along the value segment, define

$$\psi(p) = H(v_0 + p(v_1 - v_0)). \tag{94}$$

For these parameter values,

$$\psi(p) = \eta^2 + \eta^3 p. \tag{95}$$

The active pre-activation lies in $[\eta, \eta + \eta^2]$, while every other pre-activation equals $-\eta$. A fixed open raw-coordinate box around these parameter values therefore preserves the activation pattern, keeps ψ affine with derivative at least $\eta^3/2$, and keeps $b_{2,0} - \psi$ inside $(-B, B)$. Take $\kappa = \eta^3/2$ and let c_Ω be the normalized volume of this box. For parameters in the box, the interval

$$\big(b_{2,0} - \psi(\rho_{m+1}),\ b_{2,0} - \psi(\rho_m)\big) \tag{96}$$

has length at least $\kappa(\rho_{m+1} - \rho_m)$ and lies in the raw bias range. Since both ρ_t and ψ are increasing, every $t \le m$ gives a negative logit difference and every $t \ge m + 1$ gives a positive one. Thus all count levels are classified correctly. □

On the event in (88), $|\log n\, X| \le 1$, so (74) gives $\rho_{m+1} - \rho_m \ge \frac{1}{n\left(1+\sinh^2(1/2)\right)} = \Omega(n^{-1})$. Restrict the $(W_V, b_V, W_1, b_1, W_2, b_{2,0})$ to the set in Lemma 8. The remaining bias interval then has normalized length $\Omega(1/n)$. Multiplying this length by the $\Omega(1/\log n)$ contrast event and the fixed parameter box volume gives

$$V_B^{\mathrm{MAJORITY}, \mathcal{A}_{\mathrm{gen}}(h)}(n) \ge \frac{c_-}{n \log n}. \tag{97}$$

Together with (87), this proves the MAJORITY claim in Theorem 1.

A.4 PARITY

For PARITY, the count reduction in (71) gives a sequence of points along the value segment on which the required label flips. The proof compares the resulting number of sign changes with the number of activation changes available to a one-hidden-layer ReLU classifier. If $v_0 \neq v_1$, then

$$0 = \rho_0 < \rho_1 < \cdots < \rho_n = 1. \tag{98}$$

Since every count-t attention output has the form $A_t = v_0 + \rho_t(v_1 - v_0)$, all such outputs lie on the line segment joining v_0 and v_1. Parameterize this segment by $A(p) = v_0 + p(v_1 - v_0)$, $p \in [0, 1]$ and define the difference between the two output logits along it by

$$f(p) = (e_1 - e_0)^\top \big[W_2 \,\mathrm{ReLU}(W_1 A(p) + b_1) + b_2\big]. \tag{99}$$

This is a continuous piecewise-affine function of p, and each hidden preactivation can change sign at most once.

Proof of the PARITY *bound in Theorem 1.* If $v_0 = v_1$, every count level has the same attention output, so exact parity is impossible. Otherwise, exact parity makes $f(\rho_t)$ alternate in sign. By continuity, for each $t = 0, \ldots, n-1$, choose a zero $z_t \in (\rho_t, \rho_{t+1})$. For each $t = 1, \ldots, n-1$, the interval (z_{t-1}, z_t) contains ρ_t, where f is nonzero. If no hidden pre-activation changed sign in this interval, the activation pattern would be fixed and f would be affine there. An affine function that vanishes at both endpoints is identically zero, contradicting $f(\rho_t) \neq 0$. Hence each of the $n-1$ disjoint intervals (z_{t-1}, z_t) contains a point at which a hidden ReLU pre-activation changes sign. Since each hidden pre-activation changes sign at most once, this requires $h \geq n - 1$. Therefore the exact-solution set is empty whenever $n > h + 1$. □

B PROOF OF THE $\mathcal{A}_{\mathrm{norm}}(h)$ INDEX BOUND

This appendix proves Theorem 2. Recall the fixed map

$$F(v) = \frac{v}{\sqrt{1 + \|v\|_2^2}} \tag{100}$$

and the $\mathcal{A}_{\mathrm{norm}}(h)$ positional score

$$g_j^{\mathrm{norm}} = (ak + b)^\top (rj + c_K) + \big(b + F(ak)\big)^\top sj^2. \tag{101}$$

$$\widehat{\chi} = \widehat{a}^\top s, \qquad \widehat{\ell} = \ell + 2\widehat{\chi}, \qquad \varepsilon_\chi(k) = \big(F(ak) - \widehat{a}\big)^\top s, \tag{102}$$

with $\delta = b^\top s$ and $\phi = b^\top r$ as in (9). Then,

$$C_{\mathrm{quad}}^{\mathrm{norm}}(k) = \widehat{\chi} + \delta + \varepsilon_\chi(k), \qquad C_{\mathrm{lin}}^{\mathrm{norm}}(k) = \widehat{\ell}k + \phi + 2k\varepsilon_\chi(k), \tag{103}$$

Whenever $k + \Delta \in [n]$,

$$g_{k+\Delta}^{\mathrm{norm}} - g_k^{\mathrm{norm}} = \Delta C_{\mathrm{lin}}^{\mathrm{norm}}(k) + \Delta^2 C_{\mathrm{quad}}^{\mathrm{norm}}(k). \tag{104}$$

Since $F(ak)$ is collinear with a,

$$F(ak) = \lambda_a(k)\widehat{a}, \qquad \lambda_a(k) = \frac{k\|a\|_2}{\sqrt{1 + k^2\|a\|_2^2}}. \tag{105}$$

Moreover,

$$0 \leq 1 - \lambda_a(k) \leq \frac{1}{2k^2\|a\|_2^2}, \qquad |2k\varepsilon_\chi(k)| \leq \frac{\|s\|_2}{k\|a\|_2^2}. \tag{106}$$

B.1 LOWER BOUND

Use the same positional center as in (37). At this point, $\widehat{a}^0 = (1, 0)$ is orthogonal to $s^0 = (0, \eta)$, and hence

$$\widehat{\ell} = \phi = 0, \qquad Q_{a^0}(k) = -\eta^2 \qquad (k \geq 1). \tag{107}$$

Thus

$$C^{\mathrm{norm}}_{\mathrm{quad}}(k) = -\eta^2, \qquad C^{\mathrm{norm}}_{\mathrm{lin}}(k) = 0, \tag{108}$$

and therefore,

$$g^{\mathrm{norm}}_k - g^{\mathrm{norm}}_j = \eta^2(j-k)^2. \tag{109}$$

Work in a fixed neighborhood of this center on which $a_1 > \eta$. The map $a \mapsto \widehat{a}$ is smooth there. Replace the raw coordinate r_1 by $\widehat{\ell}$, keeping (a, b, r_2, s) as the other coordinates. Since

$$\frac{\partial \widehat{\ell}}{\partial r_1} = a_1, \tag{110}$$

the corresponding coordinate change is locally invertible, with Jacobian and inverse Jacobian bounded on a sufficiently small fixed neighborhood. Thus boxes in these transformed coordinates and their preimages in the original coordinates have volumes within fixed multiplicative constants, independent of n. Shrink the fixed intervals for the other transformed coordinates so that throughout the neighborhood

$$\|a\|_2 \geq \eta, \qquad b^\top s \leq -\frac{7\eta^2}{8}, \qquad |\widehat{a}^\top s| \leq \min\left\{\frac{\eta^2}{8}, \frac{\eta^4}{16}\right\}, \qquad |\phi| \leq \frac{\eta^2}{16}. \tag{111}$$

By (105), for every $k \geq 1$,

$$C^{\mathrm{norm}}_{\mathrm{quad}}(k) = \delta + \lambda_a(k)\widehat{\chi} \leq -\frac{3\eta^2}{4}. \tag{112}$$

Furthermore, (106) and (111) give

$$|2k\varepsilon_\chi(k)| \leq \frac{|\widehat{a}^\top s|}{\|a\|_2^2} \leq \frac{\eta^2}{16} \qquad (k \geq 1). \tag{113}$$

Here we used that $2k\varepsilon_\chi(k) = 2k(\lambda_a(k) - 1)\widehat{a}^\top s$.

Now restrict the only shrinking transformed coordinate by

$$|\widehat{\ell}| \leq \frac{\eta^2}{16n}. \tag{114}$$

For every $k \leq n$, equations (111), (113), and (114) imply

$$|\widehat{\ell}k + \phi + 2k\varepsilon_\chi(k)| \leq \frac{3\eta^2}{16}. \tag{115}$$

Therefore, for every nonzero integer Δ with $k + \Delta \in [n]$,

$$\begin{aligned} g^{\mathrm{norm}}_k - g^{\mathrm{norm}}_{k+\Delta} &\geq \frac{3\eta^2}{4}\Delta^2 - \frac{3\eta^2}{16}|\Delta| \\ &\geq \frac{\eta^2}{2}\Delta^2. \end{aligned} \tag{116}$$

Under the $\log n$ attention scale, the total attention outside position k is at most

$$2\sum_{d=1}^{\infty} n^{-(\eta^2/2)d^2}, \tag{117}$$

which tends to zero uniformly in k and in the parameters in the neighborhood above. Thus the same fixed neighborhoods of $C_V, b_V, W_1, b_1, W_2, b_2$ used in the baseline INDEX lower bound preserve a positive logit difference for the correct label, and c_K may vary over a fixed neighborhood because it cancels from the attention weights. The transformed coordinate $\widehat{\ell}$ varies over an interval of width $\Theta(n^{-1})$, whereas $a, b, r_2, s, c_K, C_V, b_V, W_1, b_1, W_2, b_2$ vary over fixed-width intervals. Boundedness of the local Jacobian then gives

$$V_B^{\mathrm{INDEX}, \mathcal{A}_{\mathrm{norm}}(h)}(n) = \Omega_{B,h}(n^{-1}). \tag{118}$$

B.2 Upper bound

Use the same $4n$ inequalities $y_{w_k}(w,k) > y_{1-w_k}(w,k)$ as in the original INDEX upper bound. The same output-bias argument shows that exact solutions for which one of these margins lies in $(0, n^{-5}]$ have normalized volume $O_B(n^{-4})$. For every remaining exact solution, the indexed position receives at least $c_0 n^{-5}$ attention in each constant-word/single-change comparison. Hence its score lies within a constant M_0 of the maximum score, uniformly in n. Comparing the indexed position with its two neighbors in (104) gives

$$C^{\mathrm{norm}}_{\mathrm{quad}}(k) + |C^{\mathrm{norm}}_{\mathrm{lin}}(k)| \le M_0 \qquad (2 \le k \le n-1). \tag{119}$$

Since $\|F(ak)\|_2 < 1$ and the parameters are bounded, there exists a constant $C_0 = C_0(B)$ such that

$$|C^{\mathrm{norm}}_{\mathrm{quad}}(k)| = |\big(b + F(ak)\big)^\top s| \le C_0 \qquad (a,b,s \in [-B,B]^2,\ k \ge 1). \tag{120}$$

It follows from (119) that

$$|\widehat{\ell} k + \phi + 2k\varepsilon_\chi(k)| \le M_1 \qquad (2 \le k \le n-1) \tag{121}$$

for another constant $M_1 = M_1(B,h)$.

Split the raw parameter cube according to whether

$$\|a\|_2 \le n^{-1/2}. \tag{122}$$

Because a consists of two independent raw coordinates, the normalized volume of this part is at most the area of a radius-$n^{-1/2}$ disk divided by $(2B)^2$, and hence is $O_B(n^{-1})$. On its complement, take $k_0 = \lfloor n/2 \rfloor$. For all sufficiently large n, k_0 is interior and $k_0 \ge n/3$. By (106),

$$|2k\varepsilon_\chi(k_0)| \le \frac{\|s\|_2}{k_0 \|a\|_2^2} = O_B(1) \tag{123}$$

because $\|a\|_2 > n^{-1/2}$. Boundedness also gives $|\phi| = O_B(1)$. Equation (121) at k_0 therefore forces

$$|\widehat{\ell}| \le \frac{C}{n} \tag{124}$$

for $C = C(B,h)$.

Conditional on (a,b,s),

$$\widehat{\ell} = a^\top r + \big(2b^\top s + 2\widehat{a}^\top s\big), \tag{125}$$

so (124) defines a slab in the independent variable r with normal a. Choose a coordinate i with $|a_i| \ge \|a\|_2/\sqrt{2}$ and fix the other coordinate of r. The allowed values of r_i lie in an interval of length at most $C/(n\|a\|_2)$, up to a constant depending only on B and h. Thus the normalized volume is at most

$$C_B \min\left\{1, \frac{1}{n\|a\|_2}\right\}. \tag{126}$$

Enlarging the square for a to a fixed-radius disk and integrating in polar coordinates on the complement of (122) gives

$$\begin{aligned} \int_{\substack{a \in [-B,B]^2 \\ \|a\|_2 > n^{-1/2}}} \frac{C_B}{n\|a\|_2}\, da &\le \frac{C'_B}{n} \int_{n^{-1/2}}^{\sqrt{2}B} d\varrho \\ &= O_B(n^{-1}). \end{aligned} \tag{127}$$

All remaining raw coordinates integrate out. Adding the $O(n^{-1})$ small-a region and the $O(n^{-4})$ small-margin region gives

$$V_B^{\mathrm{INDEX}, \mathcal{A}_{\mathrm{norm}}(h)}(n) = O_{B,h}(n^{-1}). \tag{128}$$

Together with (118), this proves Theorem 2.

C Fixed Sinusoidal Encoding on Majority

This appendix analyzes the fixed-frequency sinusoidal MAJORITY model used in the discussion.

C.1 Model and matched bound

Let $n = 2m + 1$ be odd, let $p_j = (\sin j, \cos j)^\top$, and define

$$x_j^{\sin} = \begin{bmatrix} e_{w_j} \\ p_j \end{bmatrix} \in \mathbb{R}^4, \qquad z^{\sin} = 0_4. \tag{129}$$

The frequency is fixed at one and is not a trainable coordinate. The query, key, and value maps are unrestricted affine maps from $\mathbb{R}^4$ to $\mathbb{R}^2$, the attention multiplier is $\beta_n = \log n$, and the output is the same width-h one-hidden-layer ReLU classifier as in Section 2. We denote this architecture by $\mathcal{A}_{\sin}(h)$. All scalar parameters are bounded in $[-B, B]$.

Theorem 9 (Fixed-sinusoidal Majority volume)**.** *For every fixed $B > 0$ and $h \geq 2$,*

$$V_B^{\textsc{Majority},\mathcal{A}_{\sin,\mathrm{gen}}(h)}(n) = \Theta(\frac{1}{n^3 \log n}). \tag{130}$$

Write the columns of W_K and W_V as

$$W_K = [k_0\ k_1\ k_{\sin}\ k_{\cos}], \qquad W_V = [v_0\ v_1\ v_{\sin}\ v_{\cos}], \tag{131}$$

and define

$$\begin{aligned} &q = b_Q, && \lambda = \beta_n q^\top (k_1 - k_0), && \Gamma = e^\lambda, \\ &a = \beta_n \begin{bmatrix} q^\top k_{\sin} \\ q^\top k_{\cos} \end{bmatrix}, && P_j = e^{a^\top p_j}, && T = \sum_{j=1}^n P_j, \\ &v_b = v_b + b_V \quad (b \in \{0,1\}), && C_V = [v_{\sin}\ v_{\cos}], && \mu = \frac{1}{T}\sum_{j=1}^n P_j p_j. \end{aligned} \tag{132}$$

The common key-bias and symbol-zero score contributions cancel, so the attention weights are

$$\alpha_j(w) = \frac{P_j \Gamma^{w_j}}{\sum_{i=1}^n P_i \Gamma^{w_i}}. \tag{133}$$

For $t \in [0,1]^n$, introduce the algebraic extensions

$$\mathsf{D}(t) = T + (\Gamma - 1)\sum_j P_j t_j, \tag{134}$$

$$\mathsf{N}(t) = \sum_j P_j (v_0 + C_V p_j) + \sum_j t_j P_j \big[\Gamma v_1 - v_0 + (\Gamma - 1) C_V p_j\big]. \tag{135}$$

Then $\mathsf{D}(t) > 0$ on the cube and $A(w) = \mathsf{N}(w)/\mathsf{D}(w)$ on binary inputs.

C.2 The necessary positional compensation

Let $d_i = (W_2)_{1i} - (W_2)_{0i}$ and $b = b_{2,1} - b_{2,0}$. After multiplying the logit difference by the positive denominator $\mathsf{D}(t)$, the resulting function is continuous and piecewise affine in t. For an activation pattern $\sigma \subseteq [h]$, define

$$\begin{aligned} &\ell_\sigma = \sum_{i \in \sigma} d_i (W_1)_{i,:}^\top, && c_\sigma = b + \sum_{i \in \sigma} d_i b_{1,i}, \\ &\delta_\sigma = \ell_\sigma^\top (v_1 - v_0), && \xi_\sigma = C_V^\top \ell_\sigma, \end{aligned} \tag{136}$$

and

$$c_\sigma^* = -\ell_\sigma^\top \frac{v_0 + \Gamma v_1}{1 + \Gamma} - \xi_\sigma^\top \mu. \tag{137}$$

Lemma 10 (Positional compensation)**.** *Outside a set of raw-cube measure zero, every exact solution at a sufficiently large odd length has an activation pattern $\sigma \subseteq [h]$ such that $\delta_\sigma \neq 0$ and*

$$\left\| a + \frac{\sinh \lambda}{\delta_\sigma} \xi_\sigma \right\|_2 \leq \frac{C}{n}, \tag{138}$$

while

$$|c_\sigma - c_\sigma^*| \le \frac{C|\delta_\sigma|}{n}\frac{\Gamma}{(1+\Gamma)^2}. \tag{139}$$

The constants are uniform over the bounded raw parameter cube, including when q and δ_σ are arbitrarily small but nonzero.

Proof. Exclude first the finite union of sets on which one of the affine pieces defined below vanishes at $t = 0$; each is null by slicing along $b_{2,1}$. On activation pattern σ, the denominator-scaled logit difference is an affine function $L_\sigma(t) = L_\sigma(0) + \sum_j \widehat{\ell}_{\sigma,j} t_j$, where

$$\widehat{\ell}_{\sigma,j} = P_j(\kappa_\sigma + \nu_\sigma^\top p_j), \qquad \kappa_\sigma = \ell_\sigma^\top(\Gamma v_1 - v_0) + (\Gamma - 1)c_\sigma, \qquad \nu_\sigma = (\Gamma - 1)\xi_\sigma. \tag{140}$$

Fix $t \in [0,1]^n$ with $\sum_j t_j = n/2$. Preserving $\sum_j t_j$, $\sum_j P_j t_j$, $\sum_j P_j \sin j\, t_j$, and $\sum_j P_j \cos j\, t_j$, one can move to a point with at most four fractional coordinates. Rounding those coordinates to words with respectively m and $m+1$ ones and using exact correctness forces a zero of one affine piece within four coordinate changes. Applying this to the centrally symmetric convex body generated by $P_j(1, p_j) \in \mathbb{R}^3$ and using the elementary convex-body slab bound yields, for some σ,

$$\sum_{j=1}^{n} \left|\widehat{\ell}_{\sigma,j} - \operatorname{med}(\widehat{\ell}_{\sigma,1}, \dots, \widehat{\ell}_{\sigma,n})\right| \le C_h M_\sigma, \qquad \left|L_\sigma\left(\tfrac{1}{2}\mathbf{1}\right)\right| \le C_h M_\sigma, \tag{141}$$

where $M_\sigma = \max_j |\widehat{\ell}_{\sigma,j}|$. The fixed phases $j \bmod 2\pi$ are equidistributed because $|\sum_{j=1}^n e^{ikj}| = O_k(1)$ for each fixed nonzero integer k. Applied to the sequence $e^{a^\top p_j}(\kappa_\sigma + \nu_\sigma^\top p_j)$, the first inequality in (141) first bounds $\|a\|_2$, then a compactness argument forces a and ν_σ/κ_σ to approach zero, and finally the first two discrete Fourier modes of the second-order Taylor expansion give

$$\kappa_\sigma \ne 0, \qquad M_\sigma \asymp |\kappa_\sigma|, \qquad \|a\|_2^2 \le \frac{C}{n}, \qquad \left\|a + \frac{\nu_\sigma}{\kappa_\sigma}\right\|_2 \le \frac{C}{n}. \tag{142}$$

To return to the raw parameters, put $e_c = c_\sigma - c_\sigma^*$ and $\tau = 2\Gamma\delta_\sigma/(1+\Gamma)$. Direct substitution gives

$$L_\sigma\left(\tfrac{1}{2}\mathbf{1}\right) = \frac{(1+\Gamma)T}{2} e_c, \qquad \kappa_\sigma = \tau - \nu_\sigma^\top \mu + (\Gamma - 1)e_c. \tag{143}$$

Since $T \asymp n$, (141) and (142) imply $|e_c| \le C|\kappa_\sigma|/((1+\Gamma)n)$ and $|\kappa_\sigma - \tau| \le Cn^{-1/2}|\kappa_\sigma|$. Hence $\delta_\sigma \ne 0$, $|\tau| \asymp |\kappa_\sigma|$, and $\|\nu_\sigma/\tau - \nu_\sigma/\kappa_\sigma\|_2 \le C/n$. Using $\nu_\sigma/\tau = (\sinh\lambda/\delta_\sigma)\xi_\sigma$ proves (138), and the bound on e_c proves (139). □

C.3 Volume integration

Proof of Theorem 9. For the upper bound, fix one of the finitely many activation patterns σ and condition on q, the two symbol key columns, and all value and classifier coordinates except $b_{2,1}$, leaving the two positional key columns unconditioned. Each coordinate of (138) restricts one independent positional key column to a slab whose conditional normalized width is at most

$$C_B \min\left\{1, \frac{1}{n\beta_n\|q\|_\infty}\right\}. \tag{144}$$

The two columns use disjoint raw coordinates. After they are fixed, (139) restricts $b_{2,1}$ to an interval of normalized length $O_{B,h}(1/n)$. Summing over σ and integrating $t = \|q\|_\infty$, whose density under the uniform square measure is proportional to t, gives

$$\begin{aligned} V_B^{\text{MAJORITY},\mathcal{A}_{\sin,\text{gen}(h)}}(n) &\le \frac{C_{B,h}}{n}\int_0^B t \min\left\{1, \frac{1}{n\beta_n t}\right\}^2 dt \\ &\le \frac{C_{B,h}(1+\log(n\beta_n))}{n^3\beta_n^2} = O_{B,h}\left(\frac{1}{n^3\log n}\right). \end{aligned} \tag{145}$$

For the lower bound, restrict the value and classifier coordinates to a fixed positive-volume open box on which the bias-free logit difference has the form $\ell^\top A + c_H$ for every attainable attention output,

with constants $0 < \delta_- \le \delta := \ell^\top(v_1 - v_0) \le \delta_+$ and $\xi = C_V^\top \ell$ satisfying $\|\xi\|_2/\delta \le 1/100$. Such a box is obtained from an interior center with $v_0 = 0$, $v_1 = \eta(1,0)^\top$, $C_V = 0$, and one uniformly active hidden unit. Set

$$A^* = \frac{v_0 + \Gamma v_1}{1+\Gamma} + C_V \mu, \qquad b^*_{2,1} = b_{2,0} - c_H - \ell^\top A^*. \tag{146}$$

At this centered bias, the exact logit difference is

$$(y(w))_1 - (y(w))_0 = \frac{\sum_j (w_j - 1/2)\widehat{\ell}_j}{\mathsf{D}(w)}, \qquad \widehat{\ell}_j = P_j \left[\frac{2\Gamma\delta}{1+\Gamma} + (\Gamma - 1)\xi^\top (p_j - \mu) \right]. \tag{147}$$

If $|\lambda| \le c_1 n^{-1/2}$ and $\|a + (\sinh\lambda/\delta)\xi\|_2 \le c_2/n$, a uniform first-order expansion of (147) gives $\widehat{\ell}_j/\tau = F_0 + O(1/n)$ with $\tau = 2\Gamma\delta/(1+\Gamma)$ and $F_0 \ge 3/4$. Therefore the numerator has the MAJORITY sign and magnitude bounded below by a positive constant, while $\mathsf{D}(w) = O(n)$; an interval $|b_{2,1} - b^*_{2,1}| < c_3/n$ preserves strict correctness on every input.

It remains to measure a full-dimensional set satisfying these conditions. Restrict the query bias to

$$q_1 > 0, \qquad |q_2| \le q_1/4, \qquad r := \beta_n q_1 \in \left[\frac{L}{n}, \frac{c}{\sqrt{n}} \right]. \tag{148}$$

For sufficiently small c, the symbol key columns may vary over their full raw squares while $|\lambda| \le c_1 n^{-1/2}$. Conditional on these choices and on the decoder box, the two coordinates of the compensation condition restrict one entry of each positional key column to an interval of length $\Theta(1/(nr))$; fixed-width restrictions on the other entries keep these intervals inside the raw cube. Their joint normalized mass is therefore $\Omega_{B,h}(1/(n^2 r^2))$, and the admissible output-bias interval contributes $\Omega_{B,h}(1/n)$. Since the query-cone measure after integrating q_2 is a constant multiple of $r\,dr/\beta_n^2$,

$$\begin{aligned} V_B^{\text{MAJORITY}, \mathcal{A}_{\sin,\mathrm{gen}}(h)}(n) &\ge \frac{c_{B,h}}{n^3 \beta_n^2} \int_{L/n}^{c/\sqrt{n}} \frac{dr}{r} \\ &\ge \frac{c'_{B,h}}{n^3 \log n}. \end{aligned} \tag{149}$$

Every restricted raw coordinate varies over an interval of positive width and every unrestricted coordinate integrates out, so this is a full-dimensional lower bound. Together with (145), it proves the theorem. □

D Experimental Details and Selected Additional Results

This appendix specifies the experimental protocol used for the main-text comparisons and reports the additional results referenced in the discussion of sinusoidal positional encoding.

D.1 Reported models and benchmark provenance

The baseline positional implementation is the model $\mathcal{A}_{\mathrm{pos}}(h)$ from Section 2, with quadratic positional coordinates (j, j^2), and the selectively normalized implementation is $\mathcal{A}_{\mathrm{norm}}(h)$ from Section 4.2. For the baseline MAJORITY experiment, we use the unrestricted affine NoPE implementation of $\mathcal{A}_{\mathrm{gen}}(h)$. The additional sinusoidal MAJORITY control is the fixed-frequency general model $\mathcal{A}_{\sin}(h)$ defined in Section C.1.

We use the same 20 seeds, 2000–2019, at each maximum training length

$$N \in \{11, 21, 31, 41, 51\}. \tag{150}$$

Training and qualification did not have access to the test banks, and every newly trained checkpoint was selected and frozen before its test evaluation.

D.2 Training, selection, and initialization

Training uses Adam with learning rate 3×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, numerical stabilizer 10^{-8}, no weight decay, batch size 256, and gradient-norm clipping at one. Each update uses a fresh batch of independent uniformly random bits. For INDEX, the index k is sampled uniformly from the positions in the sampled sequence. For FIRST and INDEX, one data length is sampled uniformly from $\{1, \ldots, N\}$ per batch. For MAJORITY, the sampled length is uniform over the odd integers in that range. Computation uses `float64`. At data length L, attention scores are multiplied by $\log L$, without an additional $1/\sqrt{d}$ factor.

Every run is trained for at most 30,000 updates and is validated every 500 updates on a fixed bank of 1,000 examples at length N. Training stops after three consecutive perfect validations. The selected checkpoint maximizes validation accuracy; ties are resolved first by lower cross-entropy and then by the earlier update. No input longer than N is used in checkpoint selection.

For the $\mathcal{A}_{\mathrm{pos}}(h)$ and $\mathcal{A}_{\mathrm{norm}}(h)$ models, the vectors a, r, s, b, c_K are initialized from centered normal distributions with standard deviations $0.1/N$, $0.1/N$, $0.1/N^2$, 0.1, 0.1, respectively. Other parameters use Xavier-uniform initialization and their biases are zero. $\mathcal{A}_{\mathrm{pos}}(h)$/$\mathcal{A}_{\mathrm{norm}}(h)$ pairs use identical initial tensors and the same training streams. For $\mathcal{A}_{\mathrm{gen}}$ and $\mathcal{A}_{\mathrm{sin}}$, content query/key blocks and the bounded sinusoidal positional blocks use Xavier-uniform initialization, while shared value, bias, and classifier blocks are copied from the paired $\mathcal{A}_{\mathrm{pos}}$ seed.

D.3 Evaluation banks, metrics, and uncertainty

For a maximum training length N, the common test grid is

$$N, \quad 2N+1, \quad 3N, \quad 5N, \quad 8N+1, \quad 10N+1. \tag{151}$$

At each test length, all seeds and model variants for the same task use the same saved bank of 1,000 examples. For FIRST, the first bit is exactly balanced and the remaining bits are random. For MAJORITY, 500 random strings are paired with their complements. For INDEX, each of 500 random string–query pairs is paired with the complemented string and the same query.

Test accuracy is the fraction correct on one saved bank. The longest-test statistic is the mean across the 20 seeds at $L = 10N + 1$. For coefficient analyses and other OOD summaries, the seed-level mean OOD accuracy is the equal-weight mean across the five test lengths strictly greater than N. Confidence intervals are pointwise percentile 95% intervals.

D.4 Coefficient analysis

The coefficient analysis uses the 100 selected INDEX checkpoints on $\mathcal{A}_{\mathrm{pos}}$: 20 seeds at each of the five values of N in (150). For each checkpoint, we compute

$$\chi = a^\top s, \qquad \ell = 2b^\top s + a^\top r \tag{152}$$

directly from the learned parameters.

To avoid treating differences among the five training maxima as evidence for a coefficient association, we rank each coefficient magnitude and the seed-level mean OOD accuracy separately within each value of N, and then pool the within-N ranks. The resulting correlations are -0.655 for $N^2|\chi|$ and -0.516 for $N|\ell|$. The two-sided permutation test uses 10,000 permutations and shuffles ranks only within the same value of N; both tests give $p < 10^{-3}$.

D.5 Selected results across maximum training lengths

Table 2 reports the longest-test results. Each entry is the mean percentage accuracy across 20 seeds, followed by its pointwise 95% seed-bootstrap interval.

D.6 Sinusoidal positional encoding on Majority

Relevant results on the performance of $\mathcal{A}_{\mathrm{sin}}$ compared to $\mathcal{A}_{\mathrm{gen}}$ on MAJORITY is displayed in Table 3.

Table 2: Accuracy at the longest tested length $L = 10N + 1$. Entries are percentages, with pointwise 95% seed-bootstrap intervals in parentheses.

N	L	FIRST ($\mathcal{A}_{\mathrm{pos}}$)	MAJORITY ($\mathcal{A}_{\mathrm{gen}}$)	INDEX ($\mathcal{A}_{\mathrm{pos}}$)	INDEX ($\mathcal{A}_{\mathrm{norm}}$)
11	111	100.0 (100.0–100.0)	83.8 (78.0–89.4)	57.2 (55.9–58.6)	63.4 (62.7–64.1)
21	211	100.0 (100.0–100.0)	91.3 (86.2–95.3)	55.2 (54.1–56.5)	74.1 (68.8–79.9)
31	311	100.0 (100.0–100.0)	87.1 (82.6–91.5)	59.2 (58.3–60.0)	90.6 (85.6–95.4)
41	411	100.0 (100.0–100.0)	88.5 (84.2–92.2)	55.1 (54.5–55.8)	84.4 (78.1–90.6)
51	511	100.0 (100.0–100.0)	82.6 (77.6–87.5)	59.6 (58.6–60.6)	85.1 (80.2–90.1)

Table 3: Mean OOD accuracy on MAJORITY for $\mathcal{A}_{\mathrm{gen}}$ and $\mathcal{A}_{\mathrm{sin}}$ general models. Entries are percentages, averaged equally over the five test lengths above N, with 95% confidence intervals in parentheses.

N	NoPE General	Sinusoidal General
11	91.6 (88.1–95.0)	95.3 (93.6–96.8)
21	95.1 (91.8–97.6)	93.8 (91.5–95.9)
31	92.9 (90.0–95.6)	91.2 (88.2–94.0)
41	93.4 (90.6–95.8)	91.5 (88.7–94.1)
51	89.8 (86.2–92.9)	92.9 (90.0–95.6)

D.7 SANITY CHECKS FOR $\mathcal{A}_{\mathrm{norm}}(h)$ ON OTHER TASKS

We also compared $\mathcal{A}_{\mathrm{pos}}(h)$ to $\mathcal{A}_{\mathrm{norm}}(h)$ on FIRST and MAJORITY. For FIRST, both positional models achieve 100% mean accuracy at $L = 10N + 1$ for every N in (150).

For our setup of MAJORITY, $\mathcal{A}_{\mathrm{pos}}$ and $\mathcal{A}_{\mathrm{norm}}$ are equivalent, hence producing same results for every seed.