
Pause Tokens Strictly Increase the Expressivity of Constant-Depth Transformers

Charles London

Department of Computer Science
University of Oxford
Oxford, OX1 3QG
charles.london@cs.ox.ac.uk

Varun Kanade

Department of Computer Science
University of Oxford
Oxford, OX1 3QG
varun.kanade@cs.ox.ac.uk

Abstract

Pause tokens, simple filler symbols such as "...", consistently improve Transformer performance on both language and mathematical tasks, yet their theoretical effect remains unexplained. We provide the first formal separation result, proving that adding pause tokens to constant-depth, logarithmic-width Transformers strictly increases their computational expressivity. With bounded-precision activations, Transformers without pause tokens compute only a strict subset of AC^0 functions, while adding a polynomial number of pause tokens allows them to express the entire class. For logarithmic-precision Transformers, we show that adding pause tokens achieves expressivity equivalent to TC^0 , matching known upper bounds. Empirically, we demonstrate that two-layer causally masked Transformers can learn parity when supplied with pause tokens, a function that they appear unable to learn without them. Our results provide a rigorous theoretical explanation for prior empirical findings, clarify how pause tokens interact with width, depth, and numeric precision, and position them as a distinct mechanism, complementary to chain-of-thought prompting, for enhancing Transformer reasoning.

1 Introduction

Transformers [38] have become ubiquitous in modern machine learning, achieving state-of-the-art performance across a wide range of tasks, particularly in natural language processing. Despite their empirical success, many aspects of their computational behaviour remain poorly understood. An intriguing recently observed phenomenon is that adding pause tokens (e.g. a "..." token) to Transformer inputs can improve performance on some question-answering and mathematical tasks [30, 14]. This phenomenon has raised questions about the computational role of such seemingly meaningless tokens, particularly in the context of chain-of-thought prompting and the faithfulness of intermediate reasoning steps.

In this paper, we provide a theoretical explanation for how pause tokens might improve Transformer performance. Specifically, we analyse how they affect the expressivity of Transformers through the lens of circuit complexity. Our contributions are as follows:

1. **Pause tokens allow Transformers to simulate Boolean circuit classes exactly.** We prove that constant-precision Transformers with a polynomial number of pause tokens are equivalent in expressivity to AC^0 (constant depth circuits with and/or/not gates), while logarithmic precision Transformers with pause tokens match TC^0 (circuits like AC^0 with threshold gates).
2. **Pause tokens strictly increase expressivity in constant-precision Transformers.** We prove a separation: Transformers without pause tokens are fundamentally less expressive than those with.

3. **Empirically, pause tokens help causally masked Transformers learn functions requiring global computation.** We show that two-layer causally masked Transformers struggle to learn parity (which belongs to TC^0) unless pause tokens are introduced. This provides some insight into how pause tokens interact with architectural constraints in practice.

We consider both uniform and non-uniform Transformer and circuit classes (see Definition 3.6), covering the cases where Transformer parameters and positional encodings can be entirely arbitrary, and where they must be generated by an efficiently computable procedure.

Our results provide a principled explanation for recent empirical findings on pause tokens while also highlighting broader implications for Transformer expressivity. In particular, they suggest that a Transformer’s computational power is not solely determined by depth and width, but also by how computation is distributed across the token sequence. The addition of pause tokens effectively expands the available computational workspace of the model, allowing it to implement more expressive computations within the same architectural constraints. These results have implications for the faithfulness of chain-of-thought reasoning, the tradeoffs between width and depth in Transformer architectures, and the impact of quantization on Transformer computation.

The remainder of this paper is organised as follows. Section 2 reviews prior work on Transformer expressivity and the use of pause tokens (more related work is in the appendix). Section 3 introduces key concepts from circuit complexity and formally defines the Transformer models we study. Section 4 presents our main theoretical results, establishing the expressivity of Transformers with and without pause tokens. Section 5 provides empirical evidence that pause tokens facilitate learning of parity under causal masking. Finally, Section 6 discusses implications and Section 7 poses open questions.

2 Related Work

We first note the difference between the use of pause tokens and chain-of-thought (CoT, Wei et al. [39]) in Transformer computation. While both methods purport to increase the computational space or memory of the Transformer to enable it to solve more problems, they do so in very different ways. In CoT, a Transformer generates tokens autoregressively and attends to its own output, enabling deeper sequential computation. With pause tokens, we simply pad the input to the Transformer with filler tokens, enabling “wider” parallel computation. While pause tokens are more efficient (as they can be computed in parallel), they are strictly less expressive than CoT [28, 22].

Transformers with pause tokens: The notion of adding padding to a Transformers input to improve performance traces back to Burtsev et al. [5], in which they prepended memory tokens to the input to a Transformer encoder, but achieved only marginal gains. Lanham et al. [21] experimented with replacing Transformer CoT tokens with blank “...” tokens at test time, and found they decreased performance versus CoT. The first work to empirically demonstrate an advantage in using pause tokens was Goyal et al. [14], in which they pre-trained a Transformer to use between 0 and 50 pause tokens before responding. This led to increased performance on natural language and mathematical benchmarks including CommonsenseQA [35] and GSM8K [10]. Since then, Pfau et al. [30] have replicated this improvement on simple mathematical tasks and proposed quantifier depth as the theoretical explanation.

Formal models of Transformers: Since the invention of the Transformer, there have been many attempts to formally characterise their computational power. Initial works determined that Transformers were Turing complete [31], but required infinite precision for the result. More recent works have used tools from theoretical computer science, including circuit complexity, communication complexity, and formal language and automata theory, to analyse more realistic models of the Transformer and derive limitations on the languages they can recognise and problems they can solve. For machine learning, language recognition can be thought of as binary classification of input strings. This section covers work using circuit complexity, but further related work can be found in Appendix A.

Circuit complexity. Hao et al. [16] showed that log-precision Transformers using unique hard attention (attending to only a single position) are contained within AC^0 , while averaging hard attention Transformers can express non- AC^0 languages. Merrill and Sabharwal [27] proved that log precision Transformers (with standard softmax attention) are contained within logspace-uniform TC^0 , while Chiang [9] extended these results to show that Transformers with $O(\text{poly}(n))$ precision are in uniform TC^0 . Further works have also used the circuit complexity perspective to determine

the expressive power of both constant- and log-precision Transformers using intermediate chain-of-thought steps, showing that they are equivalent to P (uniformly, or P/poly non-uniformly) [22, 28].

3 Preliminaries

In this section, we introduce the relevant computational complexity concepts, including circuit classes and logspace uniformity, as well as definitions of the Transformer model we are considering and fixed precision arithmetic.

3.1 Boolean Circuits

Boolean circuits are a model of computation that uses logical operations to process binary inputs and produce binary outputs. They are a natural way to formalise parallel computations (like Transformers), as their depth corresponds roughly to the parallel time required to compute a function, and their width corresponds to resource requirements at each level of parallel computation.

Definition 3.1. (Boolean circuits). A Boolean circuit C_n is a directed acyclic graph with n sources and one sink, computing a function $\{0, 1\}^n \rightarrow \{0, 1\}$. Non-source vertices are gates labeled with logical operations. The output $C_n(x)$ is determined recursively. $|C_n|$ is the number of gates; depth is the longest path from any source to the sink.

We restrict ourselves to two classes: AC^0 and TC^0 , with constant $O(1)$ depth and at most polynomial $\text{poly}(n)$ size. AC^0 circuits use NOT, AND, and OR gates with unbounded fan-in. TC^0 circuits use threshold gates $f(x) = I[\sum_{i=1}^m x_i > \theta]$ and $f(x) = I[\sum_{i=1}^m x_i < \theta]$. It is known $AC^0 \subsetneq TC^0$ [13]. AC_ℓ^0 and TC_ℓ^0 are these classes restricted to depth $\leq \ell$.

3.2 Circuit Families and Logspace Uniformity

Definition 3.2. (Circuit families). Let $T : \mathbb{N} \rightarrow \mathbb{N}$ be a function. A $T(n)$ -size circuit family is a sequence $\{C_n\}_{n \in \mathbb{N}}$ of Boolean circuits, where C_n has n inputs and a single output, and $|C_n| \leq T(n)$ for all n .

Circuit families implicitly define a language that they accept, as follows:

Definition 3.3. (Language recognition). A circuit family $\{C_n\}$ recognises $L \subset \{0, 1\}^*$ if, for all $x \in \{0, 1\}^*$, $C_{|x|}(x) = 1$ if and only if $x \in L$.

Without further constraints, circuit families can recognise undecidable languages, such as variants of the halting problem. To avoid this, we use uniform circuits, which can be constructed by some computable function. A family $\{C_n\}$ is logspace uniform if there is a Turing machine using $O(\log n)$ space that outputs C_n given 1^n (for a formal definition, see Definition B.2). Logspace uniformity ensures that circuits can be constructed systematically, avoiding non-uniform “hard-coded” solutions for each input size.

3.3 Limited Precision Arithmetic

We follow Li et al. [22] in using a fixed-point numerical representation, with 1 bit for the sign, $p(n)$ bits before the point, and $p(n)$ bits afterwards. When working with constant precision $p(n) = O(1)$, and for logarithmic precision $p(n) = O(\log n)$. We will use the shorthand $p = p(n)$.

Definition 3.4 (Fixed precision arithmetic). For a fixed precision level p :

- Values are represented in $\mathbb{F}_p = \{c \cdot k \cdot 2^{-p} : c \in \{-1, 1\}, k \in \mathbb{N}, 0 \leq k \leq 2^{2p} - 1\}$.
- Arithmetic operations (e.g., addition, multiplication) round results to the nearest representable value in $\mathbb{F}_p$. We define by $[x]_p : \mathbb{R} \rightarrow \mathbb{F}_p$ the operation that rounds to the nearest representable number. When addition is iterated, we operate left to right, rounding after each operation (so it is no longer associative).
- Saturation arithmetic is used: values exceeding the representable range $[-B_p, B_p]$, where $B_p = 2^p - 2^{-p}$, are clamped to $\pm B_p$.

We justify these choices as a reasonable approximation of how arithmetic is done in quantized neural networks. These quantized networks often use a form of outlier clipping (saturation arithmetic) to avoid issues with overflow, and round numbers to the closest number in the representable range [8].

3.4 Transformer Definition

All theoretical results in this work apply to both decoder-only and encoder-only Transformers models (i.e. with and without causal masking). In our proofs on the lower bounds of Transformer expressivity, we assume causal masking as this setting is more restrictive, but results still hold if it is removed.

Definition 3.5 (Transformer). A Transformer consists of l layers, each composed of a multi-head masked self-attention mechanism and a feedforward network. Given an input sequence $X = (x_1, \dots, x_n)$, where $x_i \in \mathbb{R}^d$:

- **Self-Attention:** Each head computes a weighted sum of a linear transform of the input tokens

$$X \leftarrow XW^V \cdot \text{softmax}(XW^{QK}X^\top + M) \quad (1)$$

where $W^{QK}, W^V \in \mathbb{R}^{d \times d}$ and $M \in \mathbb{R}^{n \times n}$. In the case that the Transformer is causally masked M is lower triangular, with $-B_p$ above the diagonal, and all other elements zero. Without masking, M is the zero matrix. The models we consider have a fixed number of attention heads.

- **Feedforward Network:** A position-wise feedforward network, with constant depth and $O(d)$ width, is applied to each token after each attention layer.
- **Positional Encodings:** Fixed or learned positional encodings $\phi(i, n) \in \mathbb{R}^d$ are added or appended to each input token x_i to encode token positions.
- **Residual Connections:** Each sublayer (attention/feedforward) has residual connections.

We will define by $\text{TF}[p(n), d(n), b(n)]$ the family of problems $\mathcal{L} : \{0, 1\}^n \rightarrow \{0, 1\}$ for which there is an integer l such that there exists an l -layer Transformer with precision $2p(n)$ ¹, embedding dimension $d(n)$, and $b(n)$ additional blank tokens that computes $\mathcal{L}(x)$ on any $x \in \{0, 1\}^n$.

In this work, we will largely operate with classes of Transformers with some range of functions $p(n), d(n), b(n)$. In the constant precision case, the classes we consider are:

1. $\text{TF}[1, L, 0] := \bigcup_{c \in \mathbb{N}} \text{TF}[c, c \log n, 0]$,
2. $\text{TF}[1, L, P] := \bigcup_{c \in \mathbb{N}} \text{TF}[c, c \log n, n^c]$,
3. $\text{TF}[1, L, Q] := \bigcup_{c \in \mathbb{N}} \text{TF}[c, c \log n, 2^{(\log n)^c}]$.

These are Transformers with constant precision, logarithmic embedding size, and no, polynomial, and quasi-polynomial pause tokens, respectively. We choose logarithmic embedding size as we feel it is the most practically relevant to current architectures, while still enabling attention over the entire input sequence. To define the uniform class $\text{TF}[1, L, Q]$, we also have to define a notion of polylogspace-uniformity, which can be defined analogously to logspace-uniformity (see Definition B.3).

In the logarithmic precision case, the classes $\text{TF}[L, L, \cdot]$ are defined analogously, with $p(n) = c$ replaced by $p(n) = c \log n$. We denote by TF_l the family of Transformers with depth $\leq l$.

3.5 Logspace Uniform Transformers

Prior work on Transformer expressivity has largely focused on two extremes of uniformity. On the one hand, Li et al. [22] analyses non-uniform Transformers, where separate parameters can be freely defined for each input length n , effectively allowing different models for different input sizes. On the other hand, Merrill and Sabharwal [27] consider Transformers where the positional encoding function $\phi(i)$ is the same for all input lengths. If we want to allow the embedding size to increase for longer input sequences (necessary in the constant precision case), our embedding function must be

¹ $2p(n)$ as we have $p(n)$ bits before the fixed point and $p(n)$ bits after.

able to depend on n i.e. $\phi(i, n)$. In practice, when a model with a larger context window is needed, often the embedding size is increased (and a whole new model is trained).²

We introduce a logspace-uniform Transformer class as a middle ground between these approaches. Our construction captures a more realistic computational model while still enforcing constraints that prevent edge cases where Transformer families could encode undecidable problems through arbitrary variations in their parameters.

Definition 3.6 (Logspace uniform Transformer families). A family of Transformers $\{T_n\}$ is logspace uniform if there exists Turing machines M_1 and M_2 operating with logarithmic space such that: M_1 takes as input 1^n and outputs the weights and biases of T_n ; M_2 takes as input 1^n and the binary encoding of an index i and outputs $\phi(i, n)$.

This formulation allows for naturally size-dependent embeddings and weights but prevents freely designing entirely unrelated models for different n . This definition also allows us to draw natural connections with uniform circuit classes, ensuring that expressivity results reflect constraints on efficiently computable models rather than arbitrary, input-length-specific designs.

4 Expressiveness of Transformers with pause tokens

Given the empirical improvements observed with pause tokens, a natural question is whether these tokens simply aid optimisation, or fundamentally alter the Transformer’s expressivity. In this section, we present our main theoretical contributions, which establish a rigorous computational foundation for increasing Transformer expressivity with pause tokens.

4.1 Constant precision Transformers and AC^0

We begin with the case where the Transformer operates with constant-precision arithmetic. Understanding expressivity in this regime has become more important due to the widespread use of low-bit quantized LLMs. First, we show that the class of constant-precision Transformers with a polynomial number of pause tokens is equivalent to AC^0 . This demonstrates that the pause tokens can act as intermediate computational units, allowing a Transformer to effectively encode the computation of a Boolean circuit within its residual stream.

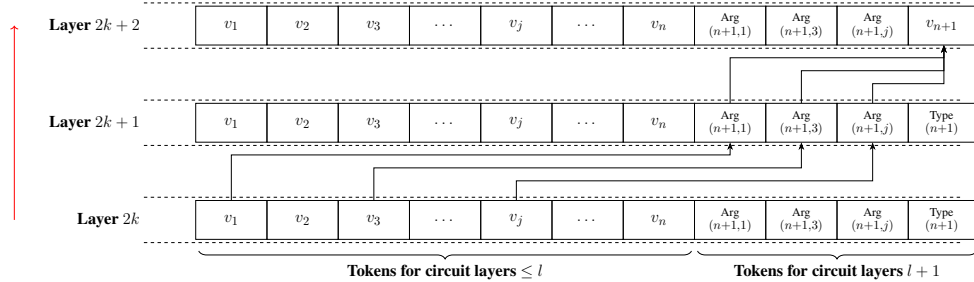


Figure 1: Two layers of a Transformer with pause tokens can simulate a layer of a Boolean circuit. In the first layer, inputs to the gates in the layer are copied to argument positions. In the second layer, these arguments are combined at the gate position to compute the output of the gate. v_i represents the value of vertex i in the circuit, whether that be an input or a gate. $\text{Arg}(i, j)$ tokens denote an edge from gate j to gate i , and $\text{Type}(i)$ tokens denote a gate i . The red arrow represents the direction of computation.

Our first result demonstrates that $\text{TF}[1, L, P]$ is computationally equivalent to AC^0 (both uniformly and non-uniformly), even when the Transformer has causal masking.

Theorem 4.1. $\text{TF}[1, L, P] = AC^0$.

²For example, when moving from Llama2 to Llama3, Meta doubled the embedding dimension of the largest model and increased the context window from 4096 to 128,000 tokens [36, 12].

To prove this equivalence, we show that (1) any function in AC^0 can be computed by a Transformer in $TF[1, L, P]$, and (2) every function computed by such a Transformer is in AC^0 . We sketch a proof below.

Simulating AC^0 with Transformers. Let C_n be a logspace-uniform AC^0 circuit family computing a Boolean function $f : \{0, 1\}^n \rightarrow \{0, 1\}$. We construct a causally masked Transformer that simulates C_n as follows:

- **Token representation:** The Transformer receives as input a sequence of length n encoding a binary string x . A polynomial number of pause tokens are appended, each representing an edge or vertex in C_n .
- **Positional encodings:** The Transformer's positional encodings represent the structure of C_n , specifying the connectivity between input bits, intermediate gates, and the final output gate. Attention weights ensure that each gate token attends only to its input tokens.
- **Computation by Transformer layers:** Each pair of Transformer layers $2k + 1$ and $2k + 2$ mimic a single layer of the circuit C_n . The first attention layer copies the values of previously computed gates to the edge tokens, and the second attention layer computes the (rounded) sum of the inputs to each gate (see Figure 1). The feedforward layers are used as thresholds to compute the exact bit value of the gate output.

A full construction, including explicit positional encodings, weight matrices, and feedforward computations can be found in Theorem C.1.

Transformers are contained in AC^0 . Every function necessary to implement a Transformer in $TF[1, L, P]$ can be implemented by an AC^0 circuit. The positional encoding function is logspace-uniform, with $O(\log n)$ size outputs, and is thus computable in (uniform) AC^0 [27]. Attention mechanisms and feedforward layers involve multiplication, division and exponentiation of a constant amount of constant precision numbers, and addition of at most a polynomial amount of constant precision numbers, all of which are in (uniform) AC^0 . Since a Transformer has a fixed number of layers, its overall computation is in AC^0 . See Theorem C.5 for the detailed proof.

Given this equivalence, we can move on to proving separations between Transformers with and without pause tokens. Our next result demonstrates that in the non-uniform case, a Transformer with a polynomial number of pause tokens is strictly more expressive than one without.

Corollary 4.2. *For the non-uniform class of Transformers, $TF[1, L, 0] \subsetneq TF[1, L, P]$.*

It follows from the result of Li et al. [22] that $TF[1, L, 0] \subsetneq AC^0$. Without pause tokens, the number of available computational units in the Transformer is limited by the sequence length n . We can obtain an upper bound on the size of the circuits it can simulate by considering the size of the circuits that can simulate it. This will be of some fixed polynomial size $O(n^k)$ (dominated by the size of the circuit needed to add n numbers). By the non-uniform size hierarchy for Boolean circuits [1], there exist functions in non-uniform AC^0 that require circuit size greater than $O(n^k)$. The class of Transformers without pause tokens is strictly less expressive than AC^0 , and therefore less expressive than $TF[1, L, P]$.

In the uniform case, the separation is weaker, as there is no equivalent of the non-uniform size hierarchy. Instead, we make use of the *fixed-depth* uniform size hierarchy for AC^0 (see Lemma D.3) to prove a separation for Transformers of some fixed depth l .

Theorem 4.3. *For the uniform class of Transformers, $TF_l[1, L, 0] \subsetneq TF_l[1, L, Q]$.*

In this case, our separation is between fixed-depth Transformers with no blank tokens, and those with a *quasi-polynomial* number of blank tokens (for the proof see Theorem D.4).

It should be clear that the addition of pause tokens does not enable the Transformer class to solve problems that require more serial computation steps, but instead increases the parallel width of computation. While it seems intuitively obvious that a deeper Transformer (still of fixed depth) that can perform more sequential steps is more expressive, it is not immediately obvious what is not computable at a fixed depth with a polynomial number of pause tokens. In the interest of rigour, we prove the following lemma.

Lemma 4.4. *For any depth l there exists a constant $m > l$ such that $TF_l[1, L, P] \subsetneq TF_m[1, L, P]$.*

The proof of this lemma is based on the result of Hastad [17], who showed that there exist functions within AC_{l+1}^0 that are not computable by any poly-size circuit in AC_l^0 .

4.2 Logarithmic Precision Transformers and TC^0

In the previous section, we established that constant precision Transformers with pause tokens are equivalent to AC^0 . We now extend this analysis to (uniform and non-uniform) logarithmic-precision Transformers, showing that they are equivalent to the strictly larger class TC^0 : constant-depth Boolean circuits with threshold gates.

Theorem 4.5. $TF[L, L, P] = TC^0$.

This result is equivalent to that obtained in Merrill and Sabharwal [27], that Transformers with polynomial advice (extra input to a TM that is allowed to depend on the input length n but not the input itself) contain TC^0 , but our definition of Transformer uniformity leads to equality in the uniform case.

The proof structure is very similar to the constant precision case, with the key difference being that logarithmic precision allows attention layers to perform threshold-like computations, as required to simulate TC^0 . In the reverse direction, it is necessary to show that the iterated sum of a polynomial amount of logarithmic precision numbers is in TC^0 . For the full proof, see Appendix C.2.

While this result establishes a theoretical equivalence, we are unable to prove a strict separation between $TF[L, L, 0]$ and $TF[L, L, P]$ as we did in the constant precision case. Unlike AC^0 , where several superpolynomial bounds are well established, TC^0 lower bounds are still an open area of research. The best known *wire* lower bounds for TC^0 are of the form n^{1+c-d} for some $c > 1$, and a TC^0 circuit simulating a Transformer trivially requires $\Omega(n^2)$ wires due to the attention between all tokens [6, 7].

5 Experiments: learning parity with pause tokens

Parity as a TC^0 benchmark. Parity is a canonical problem within TC^0 , separating AC^0 from TC^0 by requiring threshold-like computations to count the number of set bits modulo 2 [13]. While we are unable to theoretically separate $TF[L, L, 0]$ and $TF[L, L, P]$, prior work has established that learning parity is difficult for Transformers without pause tokens [3, 15]. Our experiments investigate whether pause tokens enable practical learning of parity in a 2-layer Transformer.

We hypothesize that some of the observed performance increases afforded by pause tokens are due to the restrictive nature of causal masking in decoder-only Transformers. This masking limits the model’s ability to aggregate information globally over the input sequence, as only the final token can attend to the entire sequence. For parity, a standard construction for a 2-layer threshold circuit involves n threshold gates in the first layer, with each gate taking all n input bits as input (i.e. global information, see Appendix E). A causally masked Transformer may struggle to simulate this, while pause tokens, by providing additional “gates” that can attend to the entire input, offer a mechanism to bypass this bottleneck.

Our main reason for focusing on the TC^0 regime is practical. To effectively simulate the AC^0 regime would require that our precision $< \log n$, requiring either scaling the input to unrealistic lengths (for our compute budget) with full precision or training heavily quantized models, which we found to be unstable and non-performant.

Experimental setup. We train a 2-layer, 4-head GPT-2-style Transformer [32] on bit sequences of length $\in \{20, \dots, 300\}$. The model is trained to compute the parity of the input sequence, where the final token’s output is used for classification. To study the impact of pause tokens and causal masking, we consider the following scenarios: instant answer with causal masking; instant answer without causal masking; and n pause tokens (with causal masking). Training details are in Appendix F.

Hints. We found that in all regimes, gradient-based training for a shallow Transformer struggles to learn parity using only the loss on the final prediction (never performing above chance at any length). Therefore, we follow Pfau et al. [30] in providing additional “hint” labels during training.

For the non-causal and pause token Transformers, this is in the form of the n threshold values $t_j = \mathbb{I}[\sum_{i=1}^n x_i \geq j]$ for all $j \in [1, n]$. This hint is analogous to operations performed by the first layer of the reference 2-layer threshold circuit for parity. When training with pause tokens, we compute the loss of the final prediction, and with respect to the hint over the pause tokens, as depicted below. This additional supervision directly encourages parallel threshold-like computations

Input	x_1	$\dots$	x_n	-1	$\dots$	-1	END
Labels	$\square$	$\dots$	$\square$	t_1	$\dots$	t_n	PARITY $_n$
Loss mask	0	$\dots$	0	1	$\dots$	1	1

When training the non-causal Transformer without pause tokens, we use the same hints, but located as below, with loss computed over all tokens.

Input	x_1	$\dots$	x_n	END
Labels	t_1	$\dots$	t_n	PARITY $_n$.

For the causal Transformer, the hint has to be serial, as it cannot compute threshold values for the entire sequence at any token but the last. We provide serial hints in the form of subparities: for token x_j , the label is the parity until index j , $p_j = \text{PARITY}_j$. In the instant answer regime, the loss is computed over the entire sequence as below.

Input	x_1	$\dots$	x_n	END
Labels	p_1	$\dots$	p_n	PARITY $_n$.

Results. We can see in Figure 2 that with causal masking and no pause tokens, the accuracy of the Transformer on parity falls to random guessing by the time the sequence is 100 bits long. The removal of the causal mask allows the Transformer to achieve near-perfect accuracy until the sequence reaches 200 bits, providing support for our hypothesis that causal masking leads to a bottleneck when the Transformer needs global information. However, pause tokens provide a non-trivial gain above and beyond removing the causal mask, suggesting that there are additional benefits over providing extra tokens for global information aggregation, even for the case of parity where the required circuit size should simply be linear in n . Specifically, we believe these pause tokens can act as dedicated “workspace” positions that simplify how the Transformer routes and aggregates information across the sequence. Having separate tokens for intermediate or partial computations can reduce the complexity of attention patterns in a way that is not possible without them.

Our empirical results suggest that, in causally masked Transformers, if a task requires global information, pause tokens can allow it to learn the task more easily, provided an appropriate signal is provided during training. This could help explain improvements on tasks like question-answering and mathematical reasoning, where the entire input is relevant to the final answer.

6 Discussion

The role of pause tokens. Our findings suggest that pause tokens can enhance the computational expressivity of Transformers, particularly in the constant precision setting. In constant precision settings, pause tokens allow Transformers to express all of AC^0 , and in logarithmic precision settings,

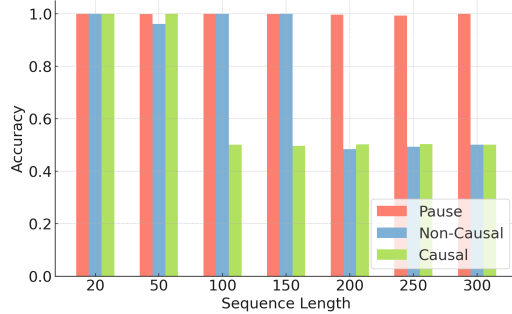


Figure 2: Test accuracy on predicting the parity of a sequence for Transformers with learned positional encodings, with and without pause tokens and causal masking. Averaged over 3 random seeds.

this increases to TC^0 . This provides a theoretical underpinning to the empirical results of Goyal et al. [14] and Pfau et al. [30].

It is known that CoT with a polynomial number of steps allows Transformers to compute everything in P (uniformly and $P/poly$ non-uniformly). It is not proven, but it is widely believed by complexity theorists that $TC^0 \subsetneq P/poly$, which would imply that log-precision Transformers with CoT are strictly more expressive than those with pause tokens. In fact, a Transformer with CoT only requires depth 2 to compute anything in $P/poly$ [22], which suggests that for a fixed depth, it is strictly more powerful. However, CoT requires running multiple forward passes through the Transformer (up to $\text{poly}(n)$ for full $P/poly$ expressivity), while pause tokens provide their additional expressivity within a single forward pass. This suggests that the two mechanisms may be complementary in improving Transformer performance, with pause tokens offering a more parallel form of computation.

That said, the practical utility of pause tokens appears limited. The theoretical gains require a polynomial number of pause tokens, and in practice models struggle to exploit this capacity effectively. These factors likely explain the weak empirical results reported by Burtsev et al. [5] and Lanham et al. [21]. Pause tokens can only help when the task demands additional computational “width” rather than depth, may require orders of magnitude more pause tokens than input tokens, and empirically need specific training strategies to become useful.

Chain-of-thought faithfulness. A natural concern is whether pause tokens introduce an additional layer of obfuscation in Transformer reasoning. Some prior work has suggested that chain-of-thought (CoT) reasoning can be misleading, in the sense that models may generate intermediate reasoning steps that are not causally linked to the final prediction [37]. Given that our results establish a theoretical advantage of pause tokens in expressivity, it raises the question of whether a model could use such tokens to obfuscate computations.

Our findings suggest that the risk of pause tokens being used for deceptive reasoning is limited. Unlike CoT, pause tokens are only useful for a more restricted class of problems (under the conjecture that $TC^0 \subsetneq P$), and training data for reasoning tends to be in the form of sequential computation steps, rather than parallelizable as would be necessary to utilize pause tokens. It seems likely that the default learned behaviour of a Transformer, especially one trained via next-token prediction, will be to generate tokens that are meaningful to the problem it is trying to solve. Even in the case that a model does learn to make use of intermediate filler tokens, Bharadwaj [2] demonstrates that non-obfuscated tokens can be recovered from the intermediate layers.

Quantization. Our results offer insights into how quantization affects Transformer expressivity. Specifically, we show that if both weights and activations are quantized to constant precision, the model is limited to AC^0 expressivity. However, if only weights are quantized while activations retain logarithmic precision, the Transformer remains in TC^0 , with the caveat that we need exact thresholding in the feedforward layers.

This follows from our proof that $TC^0 \subseteq TF[L, L, P]$ (Theorem C.7), which does not rely on weights having logarithmic precision except to perform thresholding in the feedforward network. The self-attention mechanism does not impose an expressivity constraint with constant precision weights (and saturation arithmetic), as long as the embedding dimension and activation precision remain log-size. This suggests that in mixed precision Transformer models, the expressivity bottleneck may be in the non-linearity of the feedforward network. If threshold-like behaviour can be implemented at sufficient precision, the model retains TC^0 expressivity even with weight quantization.

7 Future directions and limitations

Our theoretical results rely on specific choices in modeling Transformer computation, particularly regarding numerical precision and arithmetic operations. Here, we outline future directions for refining these results and addressing some of the limitations.

Saturation arithmetic and circuit complexity assumptions. Our proof that $TF[1, L, P] = AC^0$ relies on the use of saturation arithmetic to allow the Transformer to attend to a polynomial number of locations when simulating AC^0 gates with polynomial fan-in. While this assumption is a reasonable approximation to practical implementations of quantized neural networks, changing the arithmetic definition can change the expressivity of the models. It is possible to replace saturation arithmetic

with modular arithmetic and still demonstrate that $AC^0 \subseteq TF[1, L, P]$, but this breaks the equivalence as modular addition is not in AC^0 (in this case the Transformer belongs to ACC^0). Further work could analyse the expressivity under different numerical assumptions.

Tighter bounds on the expressivity of TF_L . For the non-uniform setting, we are able to prove a strict separation between $TF[1, L, 0]$ and $TF[1, L, P]$. However, in the uniform fixed-depth setting, our separation requires a quasi-polynomial number of pause tokens. Stricter upper and lower bounds on the expressivity of fixed-depth Transformers could lead to stronger separations between uniform Transformers with and without pause tokens. In the logarithmic precision case, a tighter bound might enable the use of the depth 2 TC^0 super-linear gate and super-quadratic wire lower bounds of Kane and Williams [20] to separate very shallow Transformers with and without pause tokens.

Separating $TF[L, L, 0]$ and $TF[L, L, P]$. While we prove that $TF[L, L, P] = TC^0$, we are unable to show that $TF[L, L, 0] \subsetneq TF[L, L, P]$. This is primarily due to the lack of known superpolynomial lower bounds for TC^0 circuits, a long-standing open problem in circuit complexity. If such bounds were established, they could lead to a provable separation between Transformers with and without pause tokens in the logarithmic-precision regime. Alternatively, perhaps approaching the problem from the Transformer side could provide some intuition for improving TC^0 lower bounds. Empirically, our results suggest that pause tokens can significantly improve the ability of causal Transformers to learn global functions, but we leave formal separation results as an open problem.

Expressivity vs. learnability. Our work focuses entirely on what Transformers with pause tokens can compute, but provides no theoretical results on whether they can learn to use this additional expressivity effectively. Empirically, we (and Pfau et al. [30]) observe that pause tokens can improve performance, but only when parallelizable explicit supervision is provided. An avenue for further research is to theoretically analyse the learnability of Transformers with pause tokens, to determine whether gradient-based training can access the full expressivity benefits.

Acknowledgments

The authors would like to thank Hanlin Ren for illuminating discussions on several circuit complexity results. CL was supported by the Engineering and Physical Sciences Research Council, grant number EP/W524311/1.

References

- [1] Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.
- [2] Aryasomayajula Ram Bharadwaj. Understanding hidden computations in chain-of-thought reasoning, 2024. URL <https://arxiv.org/abs/2412.04537>.
- [3] S. Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of transformers to recognize formal languages. In *Conference on Empirical Methods in Natural Language Processing*, 2020. URL <https://api.semanticscholar.org/CorpusID:222225236>.
- [4] Satwik Bhattamishra, Michael Hahn, Phil Blunsom, and Varun Kanade. Separations in the representational capabilities of transformers and recurrent architectures. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=6HUIJoD3wTj>.
- [5] Mikhail S Burtsev, Yuri Kuratov, Anton Peganov, and Grigory V Sapunov. Memory transformer. *arXiv preprint arXiv:2006.11527*, 2020.
- [6] Lijie Chen and Roei Tell. Bootstrapping results for threshold circuits “just beyond” known lower bounds. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019*, page 34–41, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367059. doi: 10.1145/3313276.3316333. URL <https://doi.org/10.1145/3313276.3316333>.

- [7] Lijie Chen, Binghui Peng, and Hongxun Wu. Theoretical limitations of multi-layer transformer. *ArXiv*, abs/2412.02975, 2024. URL <https://api.semanticscholar.org/CorpusID:274464787>.
- [8] Mengzhao Chen, Yi Liu, Jiahao Wang, Yi Bin, Wenqi Shao, and Ping Luo. Prefixquant: Static quantization beats dynamic through prefixed outliers in llms. *arXiv preprint arXiv:2410.05265*, 2024.
- [9] David Chiang. Transformers in uniform TC^0 . *Transactions on Machine Learning Research*, 2025.
- [10] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [11] Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Marcus Hutter, Shane Legg, and Pedro A. Ortega. Neural networks and the chomsky hierarchy. *ArXiv*, abs/2207.02098, 2022. URL <https://api.semanticscholar.org/CorpusID:250280065>.
- [12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [13] Merrick Furst, James B Saxe, and Michael Sipser. Parity, circuits, and the polynomial-time hierarchy. *Mathematical systems theory*, 17(1):13–27, 1984.
- [14] Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *ICLR*, 2024. URL <https://openreview.net/forum?id=ph04CRkPdC>.
- [15] 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. URL <https://aclanthology.org/2020.tacl-1.11/>.
- [16] Yiding Hao, Dana Angluin, and Robert Frank. Formal language recognition by hard attention transformers: Perspectives from circuit complexity. *Transactions of the Association for Computational Linguistics*, 10:800–810, 2022.
- [17] J Hastad. Almost optimal lower bounds for small depth circuits. In *Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*, STOC ’86, page 6–20, New York, NY, USA, 1986. Association for Computing Machinery. ISBN 0897911938. doi: 10.1145/12130.12132. URL <https://doi.org/10.1145/12130.12132>.
- [18] William Hesse. Division is in uniform TC^0 . In *International Colloquium on Automata, Languages, and Programming*, pages 104–114. Springer, 2001.
- [19] Neil Immerman. *Descriptive Complexity*. Springer Verlag, 1998.
- [20] Daniel M Kane and Ryan Williams. Super-linear gate and super-quadratic wire lower bounds for depth-two and depth-three threshold circuits. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 633–643, 2016.
- [21] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.
- [22] Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=3EWTEy9MTM>.
- [23] Nutan Limaye, Karteeek Sreenivasaiah, Srikanth Srinivasan, Utkarsh Tripathi, and S. Venkitesh. A fixed-depth size-hierarchy theorem for $AC^0[\oplus]$ via the coin problem, 2019. URL <https://arxiv.org/abs/1809.04092>.

- [24] Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=De4FYqjFueZ>.
- [25] Robert McNaughton and Seymour A. Papert. *Counter-Free Automata (M.I.T. research monograph no. 65)*. The MIT Press, 1971. ISBN 0262130769.
- [26] William Merrill. Formal languages and neural models for learning on sequences. In *International Conference on Graphics and Interaction*, 2023. URL <https://api.semanticscholar.org/CorpusID:261101973>.
- [27] William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023.
- [28] William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=NjNGlPh8Wh>.
- [29] Binghui Peng, Sridhar Narayanan, and Christos Papadimitriou. On limitations of the transformer architecture. *ArXiv*, abs/2402.08164, 2024. URL <https://api.semanticscholar.org/CorpusID:267636545>.
- [30] Jacob Pfau, William Merrill, and Samuel R. Bowman. Let’s think dot by dot: Hidden computation in transformer language models. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=NikbrdtYvG>.
- [31] Jorge Pérez, Javier Marinković, and Pablo Barceló. On the turing completeness of modern neural network architectures. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=HyGBdo0qFm>.
- [32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [33] Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. *ArXiv*, abs/2306.02896, 2023. URL <https://api.semanticscholar.org/CorpusID:259075636>.
- [34] 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, 2023. URL <https://api.semanticscholar.org/CorpusID:264833196>.
- [35] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, 2019.
- [36] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [37] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don’t always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 2024.
- [38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.
- [39] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: All theoretical claims are supported by complete proofs, and all assumptions are stated. The empirical claim regarding learning parity is clearly presented as supportive but not definitive.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss several limitations in detail in Sections 6 and 7. These include assumptions of fixed-point and saturation arithmetic, and that separations in the logarithmic precision regime remain conjectural. We also note that it is not possible to construct experiments that exactly match theory in the limit.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results in the paper are fully formalised, with explicit assumptions on numerical precision, saturation arithmetic, masking behaviour, and uniformity. All proofs are provided in the appendix and cross-referenced from the main text. Where theorems rely on known results (e.g., Hastad’s switching lemma, depth-hierarchy theorems), these are appropriately cited and the proofs clearly state their dependence.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Yes, code to reproduce the experimental results is provided in a .zip file, and experiment hyperparameters can be found in the appendix (and in the code).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code to reproduce the experimental results is provided in a .zip file, along with a shell script to run all experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Brief experimental details are provided in Section 5 of the main paper, and full details can be found in the appendices and code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report statistical significance or error bars. Our experiments serve primarily as illustrative case studies to support the theoretical claims, and we run a small number of seeds sufficient to observe the key qualitative differences discussed.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All experiments can be run on a single V100 GPU, as detailed in the appendices.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our work consists entirely of theoretical analysis and synthetic experiments on algorithmically generated data (bitstrings). There are no foreseeable ethical concerns, nor is any sensitive data or human subject data involved.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: We assess our work as having no immediate societal impact, as it is a purely theoretical study of Transformer expressivity under idealised conditions. Any downstream impacts would depend on further engineering or empirical deployment, which we do not pursue in this paper.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work does not release any model or data with dual-use or misuse risk.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The original GPT-2 paper [32] is cited, and the Huggingface implementation is used under Apache license. All data used was generated as part of the project.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not release any new datasets, models, or benchmarks.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our study does not involve crowdsourcing or human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subjects research is involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were used only for non-substantive code generation and copy-editing.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Further related work

Communication complexity. A related line of work has considered Transformers from a communication complexity perspective, though these results are mostly applicable to single-layer models. Peng et al. [29] use communication complexity techniques to show limitations in the representational power of single-layer Transformers, while Sanford et al. [33] and Bhattamishra et al. [4] use them to contrast the expressivity of Transformers and recurrent models. More recently, Chen et al. [7] have extended some of these results to the multi-layer case.

Formal languages and automata. Many papers have analysed Transformers through the lens of formal languages [3, 34, 26, 11], attempting to determine the class of languages that they can recognise. Liu et al. [24] proved theoretically that Transformers with $O(\log n)$ depth can simulate all finite-state automaton on inputs of length n , and so require depth that increases (mildly) with input length to capture all regular languages.

B Logspace uniformity and circuit descriptions

A logspace Turing machine may not have the space to write down its output, if the output size is $\Omega(\log n)$. For this reason, we instead require that the TM can compute any desired bit of the output in logarithmic space, i.e. it is *implicitly logspace computable* [1].

Definition B.1. (Implicitly logspace computable) A function $f : \{0, 1\}^* \rightarrow \{0, 1\}^*$ is implicitly logspace computable if the mapping $x, i \mapsto f(x)_i$ can be computed by a Turing machine using logarithmic space.

Definition B.2. (Logspace uniformity) A circuit family $\{C_n\}$ is logspace-uniform if there is an implicitly logspace computable function mapping 1^n to the description of the circuit C_n .

Definition B.3. (Polylogspace uniformity) A circuit family $\{C_n\}$ is logspace-uniform if there is an implicitly polylogspace computable function mapping 1^n to the description of the circuit C_n .

We will define circuit descriptions $\text{Desc}(C_n)$ for C_n as a topologically ordered string of vertices, $\text{Vertex}(1), \dots, \text{Vertex}(|C_n|)$, where the first n vertices correspond to input variables $\text{Inp}(1), \dots, \text{Inp}(n)$, and the remaining vertices represent logic gates. For AC^0 circuits, each gate vertex $\text{Vertex}(i)$ is defined by the string:

1. **Arguments:** $\text{Args}(i, j)$, which specifies that the output of vertex j serves as an input to vertex i .
2. **Type:** $\text{Type}(i)$, indicating the gate's operation (AND, OR, NOT).

For TC^0 circuits, $\text{Vertex}(i)$ is defined by the string:

1. **Arguments:** $\text{Args}(i, j)$, which specifies that the output of vertex j serves as an input to vertex i .
2. **Direction:** $\text{Dir}(i)$, which species if the threshold gate is $>$ or $<$.
3. **Threshold:** $\text{Thresh}(i)$, a numerical value θ representing the threshold for activation of the gate.

The topological order ensures that if we computed the circuit in order from its description, all inputs to a vertex would already have been computed by the time the vertex itself is evaluated. We will define the length of the description $|\text{Desc}(C_n)|$ as the total number of input, argument, and type or threshold strings.

C Transformer-circuit equivalence

In this section, we will demonstrate the equivalence between Transformers with pause tokens, and the circuit classes AC^0 and TC^0 . We will demonstrate the equivalence in the uniform case, but all the proofs still hold if these uniformity requirements are relaxed.

C.1 Constant precision

We will first consider Transformers with constant precision, $p(n) = O(1)$. While a full precision FP32 Transformer will almost always have precision $\geq \log n$, for any realistic input length n , more and more LLMs are being served with aggressive quantization, with 8-bit, 4-bit, and even 1.58-bit models becoming more popular. It therefore becomes necessary to understand the computational power of models under these circumstances. The constant precision regime also allows us to show strong separations between Transformers with and without pause tokens, due to the existence of universal AC^0 lower bounds.

Theorem C.1. $AC^0 \subseteq TF[1, L, P]$

Proof. Consider any logspace uniform AC^0 circuit family $\{C_n\}$. Given a Turing machine that generates the description of $\{C_n\}$, we can modify it to instead generate the position encodings and parameters of a constant precision Transformer family $\{T_n\}$ that simulates the circuits, using $b(n) \leq |\text{Desc}(C_n)|$ pause tokens.

Positional encodings. We will construct a positional encoding for every input symbol $\text{Inp}(i)$, argument symbol $\text{Arg}(i, j)$, and gate symbol $\text{Type}(i)$. All encodings have dimension $d(n) = O(\log n)$, and we will use the notation $\mathbf{bin}(i)$ to refer to the binary vector representation of $0 \leq i \leq b(n)$, and $\mathbf{sbin}(i) = 2\mathbf{bin}(i) - 1$. We will use $\mathbf{x} \frown \mathbf{y}$ to represent the interleaving of two vectors $\mathbf{x}$ and $\mathbf{y}$ of the same dimensionality, so that $(\mathbf{x} \frown \mathbf{y})_{2i-1} = x_i$ and $(\mathbf{x} \frown \mathbf{y})_{2i} = y_i$. Finally, we will define $\mathbf{k}(i) = \mathbf{sbin}(i) \frown \mathbf{1}$, and $\mathbf{q}(i) = B_p(\mathbf{sbin}(i) \frown (-\mathbf{1}))$. This means that

$$\mathbf{k}(i)^T \mathbf{q}(j) = \begin{cases} 0 & \text{if } i = j \\ -B_p & \text{otherwise.} \end{cases} \quad (2)$$

Post-exponentiation, the pre-normalized attention values will be 1 for $i = j$ and 0 everywhere else. We reserve 0 as a privileged index, such that for all indices i in the circuit description $\mathbf{k}(0)^T \mathbf{q}(i) = -B_p$. Our position encodings will then take the following forms:

- Inputs

$$\phi(\text{Inp}(i)) = (0, 0, 0, \mathbf{k}(i), \mathbf{q}(i), \mathbf{k}(i), \mathbf{q}(i)).$$

- Arguments

$$\phi(\text{Arg}(i, j)) = \begin{cases} (0, 0, 1, \mathbf{k}(0), \mathbf{q}(j), \mathbf{k}(i), \mathbf{q}(i)) & \text{if Type}(i) \text{ is OR} \\ (1, 0, 1, \mathbf{k}(0), \mathbf{q}(j), \mathbf{k}(i), \mathbf{q}(i)) & \text{otherwise.} \end{cases}$$

- Gates

$$\phi(\text{Type}(i)) = \begin{cases} (0, 1, 0, \mathbf{k}(i), \mathbf{q}(i), \mathbf{k}(0), \mathbf{q}(i)) & \text{if Type}(i) \text{ is AND} \\ (0, 0, 0, \mathbf{k}(i), \mathbf{q}(i), \mathbf{k}(0), \mathbf{q}(i)) & \text{otherwise.} \end{cases}$$

We will denote the positional encoding corresponding to circuit C_n as $\phi(C_n)$. Given a description of C_n in prefix form, $\phi(C_n)$ has the following structure

$$\phi(C_n) = \left[\begin{array}{c} | \\ \phi(\text{Inp}(1)) \dots \phi(\text{Inp}(n)) \phi(\text{Arg}(n+1, j)) \dots \phi(\text{Arg}(n+1, k)) \phi(\text{Type}(n+1)) \dots \phi(\text{Type}(|C_n|)) \\ | \end{array} \right]$$

The input to the Transformer is then a binary string x of length n , with $b(n) = \text{poly}(n)$ pause tokens appended to it. We choose to use 0 for the pause tokens to simplify the construction, but any fixed $x \in \mathbb{F}_p$ will work, as it can be corrected for in the feedforward layers. We append the position encodings at each position, to give vectors in $O(\log n)$. In matrix form, the encoded input X appears as

$$X = \left[\begin{array}{c|c} x & 0 \\ \hline & \phi(\text{Desc}(C_n)) \end{array} \right]$$

Transformer layers. For each individual layer in the AC^0 circuit, we will construct 2 (causally masked) Transformer layers that perform the same computation. For an AC^0 circuit of depth l , this will result in a Transformer with $2l$ (and hence constant) layers.

The proof proceeds by induction. If we assume that the first l layers of the circuit have been computed correctly, then at each $\text{Type}(i)$ token location in the input preceding layer $l + 1$, we have the computed value of $\text{Vertex}(i) = v_i$ (this is x_i for $\text{Inp}(i)$, and the recursively computed value for $\text{Type}(i)$). Each argument token $\text{Arg}(i, j)$ will have value 0. For $\text{Type}(i)$ tokens in layers $> l$ the value at this point can be arbitrary (denoted by $\square$), as it will be overwritten. For convenience, we will denote the set of vertices in the first l layers of the circuit C_n by $C_n[\leq l]$. We will demonstrate that the next two Transformer layers correctly compute the output values for the gates in layer $l + 1$.

The first attention layer copies previously computed vertex values into the correct argument positions for layer $l + 1$. We construct W^{QK} such that query

- $\text{Inp}(i)$ attends only to itself.
- $\text{Arg}(i, j)$ attends only to $\text{Vertex}(j)$.
- $\text{Type}(i)$ attends only to itself.

After softmax and normalization, we will have an attention value of 1 on the token u it attends to, and 0 everywhere else. The value matrix W^V will pass through the first index of the token attended to (u_1), and zero out everything else (explicit matrices for W^{QK} and W^V can be found in Appendix C.3.1). Concretely, after the attention layer and residual connection, for each token type we will have

$$\text{Inp}(i) \leftarrow 2v_i \tag{3}$$

$$\text{Arg}(i, j) \leftarrow \begin{cases} v_j & \text{if } \text{Vertex}(j) \in C_n[\leq l] \\ \square & \text{otherwise.} \end{cases} \tag{4}$$

$$\text{Type}(i) \leftarrow \begin{cases} 2v_i & \text{if } \text{Vertex}(i) \in C_n[\leq l] \\ \square & \text{otherwise.} \end{cases} \tag{5}$$

We get the factors of 2 due to self-attention and residual connections, but these are corrected for in the feedforward network. We apply a constant-depth feedforward network to each token u , computing $u_1 \leftarrow \mathbb{I}[(u_1 - u_2) \neq 0] - u_1$, $u_{i \neq 1} \leftarrow 0$ (for the ReLU network construction, see Appendix C.3.2). u_2 is 1 if u corresponds to $\text{Arg}(i, j)$ and $\text{Type}(i)$ is AND or NOT, and 0 otherwise. Therefore, post feedforward and residual connection, we have

$$\text{Inp}(i) \leftarrow v_i \tag{6}$$

$$\text{Arg}(i, j) \leftarrow \begin{cases} v_j & \text{if } \text{Type}(i) \text{ is OR and } \text{Vertex}(j) \in C_n[\leq l], \\ 1 - v_j & \text{if } \text{Type}(i) \text{ is AND or NOT and } \text{Vertex}(j) \in C_n[\leq l], \\ \square & \text{otherwise.} \end{cases} \tag{7}$$

$$\text{Type}(i) \leftarrow \begin{cases} v_i & \text{if } \text{Vertex}(i) \in C_n[\leq l] \\ \square & \text{otherwise.} \end{cases} \tag{8}$$

The positional encodings of all tokens are unaffected by these layers, as they zero out the positional information, and the residual connections copy it over.

The second Transformer layer then computes the gate outputs for layer $l + 1$. Due to our use of saturation arithmetic, we can attend an arbitrary subset of input locations with constant precision. The maximum value of the normalizing constant in the attention computation is B_p , and post-exponentiation attention values are either 0 or 1, so our minimal non-zero attention value is $1/B_p \geq 1/2^p$, and so is still representable as a non-zero value. W^{QK} is constructed such that query

- $\text{Inp}(i)$ attends only to itself.
- $\text{Arg}(i, j)$ attends only to itself.
- $\text{Type}(i)$ attends to all $\text{Arg}(i, \cdot)$, and does not attend to itself.

W^{PV} again only allows through u_1 . If we define m_i as the input arity of $\text{Vertex}(i)$, and define $\hat{m}_i = \lceil m_i \rceil_p$, then the representation of each token after the attention layer and residual connection becomes

$$\text{Inp}(i) \leftarrow 2v_i \quad (9)$$

$$\text{Arg}(i, j) \leftarrow \begin{cases} 2v_j & \text{if Type}(i) \text{ is OR and Vertex}(j) \in C_n[\leq l], \\ 2 - 2v_j & \text{if Type}(i) \text{ is AND or NOT and Vertex}(j) \in C_n[\leq l], \\ \square & \text{otherwise.} \end{cases} \quad (10)$$

$$\text{Type}(i) \leftarrow \begin{cases} \frac{1}{\hat{m}_i} \sum_{\{j \mid \text{Arg}(i, j)\}} v_j & \text{if Type}(i) \text{ is OR and Vertex}(i) \in C_n[\leq l + 1], \\ \frac{1}{\hat{m}_i} \sum_{\{j \mid \text{Arg}(i, j)\}} (1 - v_j) & \text{if Type}(i) \text{ is AND or NOT and Vertex}(i) \in C_n[\leq l + 1], \\ \square & \text{otherwise.} \end{cases} \quad (11)$$

The addition is iterated fixed point addition, but all we require is that the result will be non-zero if any of the addition terms is non-zero. The value of $\text{Type}(i)$ will be > 0 if it is: OR and any argument $v_j = 1$; AND and any argument $v_j = 0$; NOT and the argument $v_j = 0$. Otherwise, the value will be 0. The feedforward network will then compute

$$u_1 \leftarrow \begin{cases} -u_1 & \text{if } u_4 = 1 \text{ (} u \text{ is an Arg token),} \\ 1 - \mathbb{I}[u_1 > 0] - u_1 & \text{if } u_4 = 0 \text{ and } u_3 = 1 \text{ (} u \text{ is Type}(i) \text{ with type AND),} \\ \mathbb{I}[u_1 > 0] - u_1 & \text{otherwise.} \end{cases} \quad (12)$$

$$u_{i \neq 1} \leftarrow 0 \quad (13)$$

where u is any token. Combined with the residual connection, this feedforward layer sets u_1 to 1 if $u_1 = 0$ and u corresponds to an AND gate, 0 if u is an Arg token, and $\mathbb{I}[u_1 > 0]$ for all other tokens. We can see that for each token $\text{Type}(i)$ in circuit layers $\leq l + 1$, if it is

- NOT: the first Transformer layer copies the arguments to the correct position and computes NOT ($1 - v_j$), and the second layer copies it to the gate location.
- OR: the first layer copies all arguments to the correct position, and the second layer computes OR by summing the inputs and returning 1 only if the sum is > 0 .
- AND: the first layer copies all arguments to the correct position and computes their NOT ($1 - v_j$), and the second layer computes the negation of the OR of the negated arguments, by returning 1 only if the sum = 0. By De Morgan's laws, this is an AND gate.

Thus, following the evaluation of these two Transformer layers, the residual stream will now contain the value of all the gates in layers $\leq l + 1$ of the AC^0 circuit, the input tokens are unchanged, and all arguments have value 0, as required to compute the next layer. Evaluating all $2l$ layers will result in the final value of C_n being output at the final position. $\square$

Lemma C.2. [27] Let $f : \{0, 1\}^* \rightarrow \{0, 1\}^m$ be a linear space computable function. There exists a Turing machine that, for all $n \in \mathbb{N}$ and $c \in \mathbb{R}^+$, uses at most $c \log n + \log m$ space to map input 1^n to an AC^0 circuit of size at most $nc + c \log n + m$ and depth 3 that computes f on inputs of size at most $c \log n$.

Lemma C.3. (Li et al. [22]) For any fixed $p \in \mathbb{N}$, $\text{sum}_p : (\mathbb{F}_p)^n \rightarrow \mathbb{F}_p$ has uniform AC^0 circuits.

Remark. While Li et al. [22] do not explicitly state that their construction is uniform, their proof hinges on the fact that constant precision saturation arithmetic is computable by a counter-free automata. Such automata recognize star-free languages, which are definable in $\text{FO}[<]$ [25]. Since $\text{FO}[<]$ corresponds to $\text{DLOGTIME-uniform AC}^0$ [19], and $\text{DLOGTIME} \subset L$, the construction is logspace uniform.

Lemma C.4. Exponentiation $[e^x]_p$, multiplication $[xy]_p$, and division $[x/y]_p$ of constant precision numbers $x, y \in \mathbb{F}_p$ has uniform AC^0 circuits.

Proof. These functions are finite and can therefore be implemented as a constant-size lookup table. A logspace TM can trivially hardcode this constant-size circuit description. $\square$

Theorem C.5. $\text{TF}[1, L, P] \subseteq \text{AC}^0$

Proof. We demonstrate that every logspace uniform Transformer with logarithmic-dimensional embeddings, constant precision, and polynomial pause tokens can be simulated by a uniform circuit family. The proof proceeds by decomposing the Transformer’s operations into components computable in AC^0 .

Positional encodings. By construction, our positional encodings can be computed in logarithmic space, with binary inputs i and n of dimension at most $k \log n$, and outputs of dimension $m = k' \log n$. We can use logarithmic space to construct a counter maintaining the index i (from 1 to $b(n)$) and construct constant-size circuits to convert i into binary. By Lemma C.2 we can construct a logspace Turing machine mapping 1^n to the description of an AC^0 circuit computing $\phi(i, n)$.

Attention layers. We will prove that the action of a single head can be computed in uniform AC^0 , and as our model uses a constant number of heads, the action of the entire layer is in uniform AC^0 . Recall that the attention computation is defined as

$$\mathbf{x}_i \leftarrow W^V \sum_{j=1}^i \frac{\exp(f(\mathbf{x}_i, \mathbf{x}_j))}{Z_i} \mathbf{x}_j \quad (14)$$

$$f(i, j) = \mathbf{x}_i^T W^{QK} \mathbf{x}_j \quad (15)$$

$$Z_i = \sum_{j=1}^i \exp(f(\mathbf{x}_i, \mathbf{x}_j)) \quad (16)$$

First, we demonstrate that the attention function f has uniform AC^0 circuits. By definition W^{QK} is logspace computable, and so we can determine any entry W_{ij}^{KQ} by simulating M_1 using logspace. f can therefore be computed in logspace, by using a counters k and l over the length of the vectors, computing each $W_{kl}^{KQ}(\mathbf{x}_j)_k(\mathbf{x}_i)_l$, and storing the summation of these terms in constant space. By Lemma C.2, f is in uniform AC^0 . To do this for all query and key vectors, we can use two counters over the token locations.

Exponentiation of these attention values and multiplication with $\mathbf{x}_j$ can be done in uniform AC^0 by Lemma C.4. Summation over the (up to $\text{poly}(n)$) attention values for calculating the normalizing constant and summation of the value vectors can be done in uniform AC^0 by Lemma C.3.

As W^V is logspace computable, we can construct a logspace computable function g to compute $W^V \mathbf{x}$, again by using counters over the indices of W^V . This function has a logarithmic output dimension, and so we can use Lemma C.2 to prove that g has a uniform AC^0 circuit.

Feedforward layers. Each feedforward layer involves multiplying a token vector with a logspace computable weight matrix. Similarly to g , this is in uniform AC^0 .

As we have demonstrated that all Transformer computations for each layer are in uniform AC^0 , and we have a constant number of layers, $TF[1, L, P] \subseteq AC^0$.

□

Corollary C.6. $TF[1, L, P] = AC^0$

C.2 Logarithmic precision

Theorem C.7. $TC^0 \subseteq TF[L, L, P]$.

Proof. Consider any logspace uniform TC^0 circuit family $\{C_n\}$. Given a Turing machine that generates the description of $\{C_n\}$, we can modify it to instead generate the position encodings and parameters of a constant precision Transformer family $\{T_n\}$ that simulates the circuits, using $b(n) \leq |\text{Desc}(C_n)|$ pause tokens.

Positional encodings. Define $\mathbf{k}(i)$ and $\mathbf{q}(i)$ as in the AC^0 case, though B_p now refers to the largest representable number with logarithmic precision. The positional encodings are:

- Inputs

$$\phi(\text{Inp}(i)) = (0, 0, 0, \mathbf{k}(i), \mathbf{q}(i), \mathbf{k}(i), \mathbf{q}(i)).$$

- Arguments

$$\phi(\text{Arg}(i, j)) = \begin{cases} (0, 0, 1, \mathbf{k}(0), \mathbf{q}(j), \mathbf{k}(i), \mathbf{q}(i)) & \text{if } \text{Dir}(i) \text{ is } > \\ (1, 0, 1, \mathbf{k}(0), \mathbf{q}(j), \mathbf{k}(i), \mathbf{q}(i)) & \text{otherwise.} \end{cases}$$

- Gates

$$\phi(\text{Thresh}(i)) = \begin{cases} (0, \theta/m_i, 0, \mathbf{k}(i), \mathbf{q}(i), \mathbf{k}(0), \mathbf{q}(i)) & \text{if } \text{Dir}(i) \text{ is } > \\ (0, -\theta/m_i, 0, \mathbf{k}(i), \mathbf{q}(i), \mathbf{k}(0), \mathbf{q}(i)) & \text{otherwise,} \end{cases}$$

where θ is the threshold value and m_i is the input arity of the vertex.

$\phi(C_n)$ has the following structure

$$\phi(C_n) = \left[\begin{array}{c} | \\ \phi(\text{Inp}(1)) \dots \phi(\text{Inp}(n)) \phi(\text{Arg}(n+1, j)) \dots \phi(\text{Arg}(n+1, k)) \phi(\text{Thresh}(n+1)) \dots \phi(\text{Threshold}(|C_n|)) \\ | \end{array} \right]$$

Transformer layers. Again, we use two Transformer layers per circuit layer, and proceed by induction, assuming that all l previous circuit layers have been correctly computed.

The first attention layer performs the same operation as in the AC^0 case, copying previously computed vertices to argument locations. The first feedforward layer computes

$$\mathbf{u}_1 \leftarrow \begin{cases} \mathbb{I}[\mathbf{u}_1 > 0] - \mathbf{u}_1 & \text{if } \mathbf{u}_2 = 0 \\ -\mathbb{I}[\mathbf{u}_1 > 0] - \mathbf{u}_1 & \text{otherwise.} \end{cases} \quad (17)$$

$$\mathbf{u}_{i \neq 1} \leftarrow 0. \quad (18)$$

Post feedforward and residual connection, we have

$$\text{Inp}(i) \leftarrow v_i \quad (19)$$

$$\text{Arg}(i, j) \leftarrow \begin{cases} v_j & \text{if Dir}(i) \text{ is } > \text{ and Vertex}(j) \in C_n[\leq l], \\ -v_j & \text{if Dir}(i) \text{ is } < \text{ and Vertex}(j) \in C_n[\leq l], \\ \square & \text{otherwise.} \end{cases} \quad (20)$$

$$\text{Thresh}(i) \leftarrow \begin{cases} v_i & \text{if Vertex}(i) \in C_n[\leq l] \\ \square & \text{otherwise.} \end{cases} \quad (21)$$

The second attention layer also acts in the same way as in the AC^0 case. However, due to the logarithmic precision we no longer have potential overflow when attending to $\text{poly}(n)$ argument locations, and the attention layer will be able to attend to all arguments with attention weights summing to 1, so that averaging of the inputs is performed correctly. After the attention layer and residual connection we have

$$\text{Inp}(i) \leftarrow 2v_i \quad (22)$$

$$\text{Arg}(i, j) \leftarrow \begin{cases} 2v_j & \text{if Dir}(i) \text{ is } > \text{ and Vertex}(j) \in C_n[\leq l], \\ -2v_j & \text{if Dir}(i) \text{ is } < \text{ and Vertex}(j) \in C_n[\leq l], \\ \square & \text{otherwise.} \end{cases} \quad (23)$$

$$\text{Thresh}(i) \leftarrow \begin{cases} \frac{1}{m_i} \sum_{\{j \mid \text{Arg}(i, j)\}} v_j & \text{if Dir}(i) \text{ is } > \text{ and Vertex}(i) \in C_n[\leq l+1], \\ \frac{1}{m_i} \sum_{\{j \mid \text{Arg}(i, j)\}} -v_j & \text{if Dir}(i) \text{ is } < \text{ and Vertex}(i) \in C_n[\leq l+1], \\ \square & \text{otherwise.} \end{cases} \quad (24)$$

The second feedforward layer will compute

$$\mathbf{u}_1 \leftarrow \begin{cases} -\mathbf{u}_1 & \text{if } \mathbf{u}_4 = 1 \text{ (} \mathbf{u} \text{ is an Arg token)} \\ \mathbb{I}[(\mathbf{u}_1 - \mathbf{u}_3) > 0] - \mathbf{u}_1 & \text{otherwise,} \end{cases} \quad (25)$$

$$\mathbf{u}_{i \neq 1} \leftarrow 0 \quad (26)$$

For every token type, we have

$$\text{Inp}(i) \leftarrow v_i \quad (27)$$

$$\text{Arg}(i, j) \leftarrow 0 \quad (28)$$

$$\text{Thresh}(i) \leftarrow \begin{cases} \mathbb{I}\left[\sum_{\{j \mid \text{Arg}(i, j)\}} v_j > k\right] & \text{if Dir}(i) \text{ is } > \text{ and Vertex}(i) \in C_n[\leq l+1], \\ \mathbb{I}\left[\sum_{\{j \mid \text{Arg}(i, j)\}} v_j < k\right] & \text{if Dir}(i) \text{ is } < \text{ and Vertex}(i) \in C_n[\leq l+1], \\ \square & \text{otherwise.} \end{cases} \quad (29)$$

These are the correct output values for the threshold gates in layers $\leq l+1$, and we have ensured that we have the required residual values to compute the next layer. Evaluating all $2l$ layers will result in the final value of C_n being output at the final position. $\square$

Lemma C.8. (Merrill and Sabharwal [27], paraphrased) *For any $p(n) = O(\log n)$, the function $\text{sum}_{p(n)} : (\mathbb{F}_{p(n)})^n \rightarrow \mathbb{F}_{p(n)}$ has uniform TC^0 circuits.*

Lemma C.9. [18] *Multiplication and division of two n -bit numbers is in uniform TC^0 .*

Lemma C.10. *Exponentiation $[e^x]_{p(n)}$ of an $p(n) = O(\log n)$ bit number is in uniform TC^0 .*

Proof. By Hesse [18], iterated multiplication of n n -bit numbers is in uniform TC^0 . As exponentiation of an $p(n) = O(\log n)$ bit number can be seen as iterated multiplication of n constant bit numbers, it is in uniform TC^0 . $\square$

Theorem C.11. $\text{TF}[L, L, P] \subseteq \text{TC}^0$

The only difference between this proof and the constant precision case is that we must be able to perform the required arithmetic with logarithmic precision. By Lemma C.8, Lemma C.9 and Lemma C.10, all the necessary arithmetic operations are in uniform TC^0 .

Corollary C.12. $\text{TC}^0 = \text{TF}[L, L, P]$

C.3 Model parameters and feedforward networks

C.3.1 Attention weights

- First layer:

$$W^{QK} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad W^V = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- Second layer:

$$W^{QK} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & I & 0 & 0 \end{bmatrix} \quad W^V = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

These weights encode a very simple connectivity pattern and can easily be constructed in logspace.

C.3.2 Feedforward networks

Constant precision. We first define a ReLU function

$$g(a) = B_p \left(\text{ReLU}(a) + \text{ReLU} \left(a - \frac{1}{2^p} \right) \right). \quad (30)$$

B_p and 2^{-p} are the largest and smallest representable positive values in $\mathbb{F}_p$, and as $[B_p/2^p]_p = 1$, $g(a) = \mathbb{I}[a > 0]$. The first feedforward network in our Transformer computes $\mathbb{I}[(\mathbf{u}_1 - \mathbf{u}_2) \neq 0] - \mathbf{u}_1$, which can be done by the following ReLU function

$$g(\mathbf{u}_1 - \mathbf{u}_3) + g(\mathbf{u}_3 - \mathbf{u}_1) - \text{ReLU}(\mathbf{u}_1) - \text{ReLU}(-\mathbf{u}_1), \quad (31)$$

implementable by a 3-layer feedforward ReLU network.

The second feedforward network computes

$$\text{ReLU}(1 - g(\mathbf{u}_1) + \mathbf{u}_3 - \mathbf{u}_4 - 1) + \text{ReLU}(g(\mathbf{u}_1) - \mathbf{u}_3 - \mathbf{u}_4) - \text{ReLU}(\mathbf{u}_1) - \text{ReLU}(-\mathbf{u}_1), \quad (32)$$

implementable by a 4-layer feedforward ReLU network.

Logarithmic precision. We use the same g , but now defined over $\mathbb{F}_{p(n)}$, where $p(n) = O(\log n)$. The first feedforward network computes

$$-\text{ReLU}(g(\mathbf{u}_1) + \mathbf{u}_2 - 1) + (\text{ReLU}(g(\mathbf{u}_1) - \mathbf{u}_2) - \text{ReLU}(\mathbf{u}_1) - \text{ReLU}(-\mathbf{u}_1)), \quad (33)$$

implementable by a 4-layer feedforward ReLU network.

The second feedforward network computes

$$\text{ReLU}(g(\mathbf{u}_1 - \mathbf{u}_3) - \mathbf{u}_4) - \text{ReLU}(\mathbf{u}_1) - \text{ReLU}(-\mathbf{u}_1), \quad (34)$$

implementable by a 4-layer feedforward ReLU network.

D Separations in expressivity

We will use the notation $\text{AC}^0[h(n)]$ (and $\text{TC}^0[h(n)]$) to define the class of functions computable by AC^0 (TC^0) circuits with $O(h(n))$ gates/vertices.

D.1 Non-uniform constant precision Transformers

Lemma D.1. [22] $\text{TF}[1, L, 0] \subsetneq \text{AC}^0$

Proof (sketch). The crux of this argument is that when the Transformer has only n tokens and $d(n) = O(\log n)$ embedding dimension, it can be simulated by $\text{AC}^0[n^k]$ for some fixed $k \in \mathbb{N}$, as all positional encoding, attention, and feedforward operations are computable by fixed polysize circuits. We can therefore always find some non-uniform AC^0 circuit with size $n^{k'}$, where $k \leq k'$, such that the function computed by this circuit is not in $\text{TF}[1, L, 0]$. $\square$

This is no longer the case when the Transformer is allowed $\text{poly}(n)$ pause tokens, as an AC^0 circuit simulating this Transformer must be able to add $\text{poly}(n) = \cup_{k \in \mathbb{N}} n^k$ terms to compute the attention layer, and so cannot be of a fixed size n^k .

Corollary D.2. For non-uniform Transformers and AC^0 , $\text{TF}[1, L, 0] \subsetneq \text{TF}[1, L, P]$.

Proof. By Corollary C.6 $\text{TF}[1, L, P] = \text{AC}^0$, and by Lemma D.1 $\text{TF}[1, L, 0] \subsetneq \text{AC}^0$. $\square$

D.2 Uniform constant precision Transformers

In the uniform case, it is more difficult to construct a function f_k for all $k \in \mathbb{N}$ such that $f_k \in \text{AC}^0$ and $f_k \notin \text{AC}^0[n^k]$. However, we can do this for fixed-depth circuits using the tight AC^0 lower bounds for parity [13, 17].

Lemma D.3. (Limaye et al. [23], paraphrased) There exists a fixed depth size hierarchy for uniform AC^0 , such that $\text{AC}_l^0[n^{k\epsilon}] \subsetneq \text{AC}_l^0[n^k]$ for some fixed $0 < \epsilon < 1$.

Proof (sketch). This result of Hastad [17] states that any depth- l AC^0 circuit for parity of n variables must have size $2^{\Omega(n^{1/l-1})}$, which is tight due to the folklore depth- l AC^0 upper bound of $2^{O(n^{1/l-1})}$. These bounds give us a separation between circuits of size $s_0 = 2^{O(n^{1/l-1})}$ and s_0^ϵ for some fixed $0 < \epsilon < 1$. The same separation holds between s^ϵ and s for any $s \leq s_0$, by taking the parity function over some $m \leq n$ variables. If we take $s(n) = O(n^k)$ ($m = \text{polylog}(n)$), we obtain a separation between $\text{AC}_l^0[n^{k\epsilon}]$ and $\text{AC}_l^0[n^k]$.

This separation holds in uniform AC^0 , as parity over $(k \log n)^{l-1}$ bits can be computed via a simple uniform divide-and-conquer approach that computes the parity of $O(\log n)$ bit chunks in parallel at each layer. $\square$

Theorem D.4. $\text{TF}_l[1, L, 0] \subsetneq \text{TF}_l[1, L, Q]$.

Proof. By a similar argument to Lemma D.1 we have that $\text{TF}_l[1, L, 0] \subseteq \text{AC}_{l'}^0[n^{k\epsilon}]$ for some $l', k \in \mathbb{N}$ and $0 < \epsilon < 1$. By Lemma D.3 we know that there is some function f_n over suitably chosen $m = \text{polylog}(n)$ bits such that $f_n \notin \text{AC}_l^0[n^{k\epsilon}]$, and $f_n \in \text{AC}_l^0[n^k]$.

By Theorem C.1, we know that $\text{TF}_l[1, L, |\text{Desc}(C_n)|]$ can compute any function computed by a circuit of depth $l/2$ with $|C_n|$ unbounded fan-in AND, OR, and NOT gates. By the parity bounds, we know that a depth $l/2$ circuit of size $|C_n| = 2^{O(m^{2l'/(l-2)})}$ can compute parity over m input bits. For $m = \text{polylog}(n)$, this is quasi-polynomial in n . $|\text{Desc}(C_n)|$ cannot be more than quasi-polynomial in $|C_n|$, and so the number of pause tokens necessary to compute the function is still quasi-polynomial.

Therefore $f_n \notin \text{TF}_l[1, L, 0]$ and $f_n \in \text{TF}_l[1, L, Q]$. It is intuitively obvious that $\text{TF}_l[1, L, 0] \subseteq \text{TF}_l[1, L, Q]$, and so the proof is complete. $\square$

D.3 Width cannot compensate for depth

It should be clear that the addition of pause tokens does not enable the Transformer class to solve problems that require more serial computation steps, instead increasing the parallel width of computation. For finite-depth, constant precision Transformers with up to $\text{poly}(n)$ pause tokens, we will always be limited to the class AC^0 . A prudent question is whether, within this class, increasing the depth of the Transformer increases its expressivity? This seems intuitively obvious, but we provide a formal proof below.

Theorem D.5. *For any depth l there exists a constant $m > l$ such that $\text{TF}_l[1, L, P] \subsetneq \text{TF}_m[1, L, P]$.*

Proof. By Theorem C.1 and Theorem C.5, we have that $\text{AC}_{l/2}^0 \subseteq \text{TF}_l[1, L, P] \subseteq \text{AC}_{l'}^0$. By the result of Hastad [17], there exist functions in AC_{d+1}^0 that are not in AC_d^0 (with polynomial size). Therefore, $\text{TF}_m[1, L, P]$, where $m \geq 2l'$, is strictly more expressive than $\text{TF}_l[1, L, P]$. $\square$

E 2-layer threshold circuit for parity

Let $x_1, \dots, x_n \in \{0, 1\}$ be the inputs. Define $s = \sum_{i=1}^n x_i$.

First Layer: For $k = 1, 2, \dots, n$, construct a threshold gate

$$G_k = \mathbb{I} \left[\sum_{i=1}^n x_i \geq k \right].$$

In other words,

$$G_k = \begin{cases} 1 & \text{if } s \geq k, \\ 0 & \text{otherwise.} \end{cases}$$

Note that exactly $G_1, G_2, \dots, G_s$ will output 1, and the gates $G_{s+1}, \dots, G_n$ will output 0.

Second Layer: Introduce one additional threshold gate F that takes inputs from $G_1, \dots, G_n$:

$$F = \mathbb{I} \left[\sum_{i=1}^n (-1)^{i+1} G_i \geq 1 \right].$$

Correctness:

$$\sum_{i=1}^n (-1)^{i+1} G_i = \sum_{i=1}^s (-1)^{i+1}$$

This is a simple alternating sum of 1 and -1 for s terms, which is 1 if s is odd, and 0 if it is even. Therefore,

$$F = 1 \iff \sum_{i=1}^n (-1)^{i+1} G_i \geq 1 \iff s \text{ is odd.}$$

Thus F outputs 1 exactly when the number of 1's among $x_1, \dots, x_n$ is odd, i.e. F computes the parity function.

F Experiment details

All experiments were conducted using a single NVIDIA V100 GPU. The model is a custom variant of Huggingface’s GPT-2 implementation, to allow for different input dimensions and position encodings. Our Transformer model uses 2 layers, 4 attention heads, and a hidden dimension of 32. Positional encodings are learned during training. All models are trained for 50 epochs using the Adam optimiser with a learning rate of 5×10^{-4} , $\beta_1 = 0.9$, $\beta_2 = 0.999$, and no weight decay. We disable mixed precision and gradient clipping, as the models are small and training is stable.

Validation accuracy is monitored each epoch, and the model achieving the highest validation accuracy is used for test set evaluation. Training, validation, and test datasets consist of 500,000, 5,000, and 50,000 examples, respectively. The datasets consist of uniformly sampled bitstrings, where the label is the parity. We generate a dataset per random seed for each length in $\{20, 50, 100, 150, 200, 250, 300\}$.

All experiments are logged using Weights & Biases, including batch and epoch metrics. Models are saved every 5 epochs and upon reaching a new maximum validation accuracy. The final test accuracy reported in the plots is averaged over three random seeds.

G Attention pattern analysis

Here we analyse the positional attention patterns of trained models in each class of causal, non-causal, and pause tokens, to try and understand what mechanisms/algorithms they may be using to compute the parity.

To compare attention distributions across positions, we transform raw attention weights $p(k | q)$ for query position q and key position k into a *symmetric normalised form* that measures deviation from a uniform baseline while remaining bounded and zero-centred. Let $K_{\text{eff}}(q)$ denote the number of unmasked keys visible to query q , and define the uniform reference distribution $U(k | q) = 1/K_{\text{eff}}(q)$. We first form the likelihood ratio

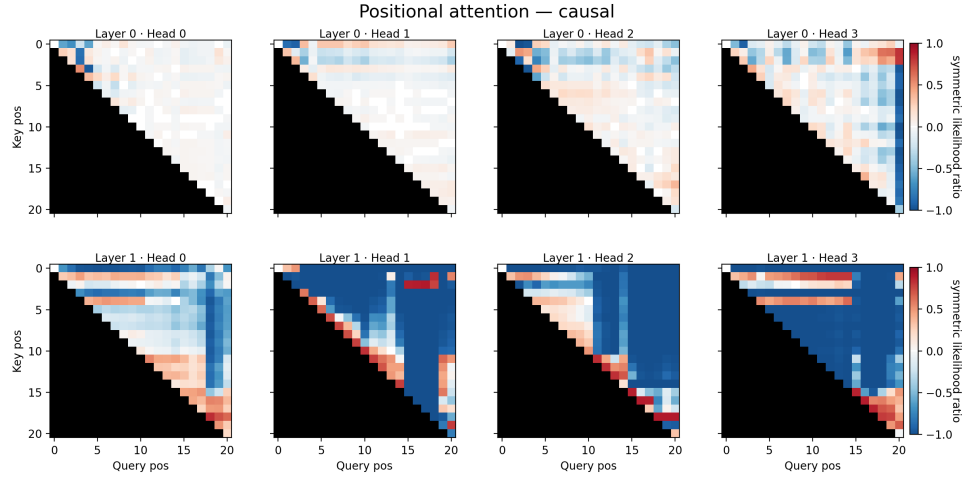
$$r(k | q) = \frac{p(k | q)}{U(k | q)} = K_{\text{eff}}(q) p(k | q),$$

which quantifies how strongly query q attends to key k relative to a uniform baseline. To obtain a symmetric, bounded representation, we define

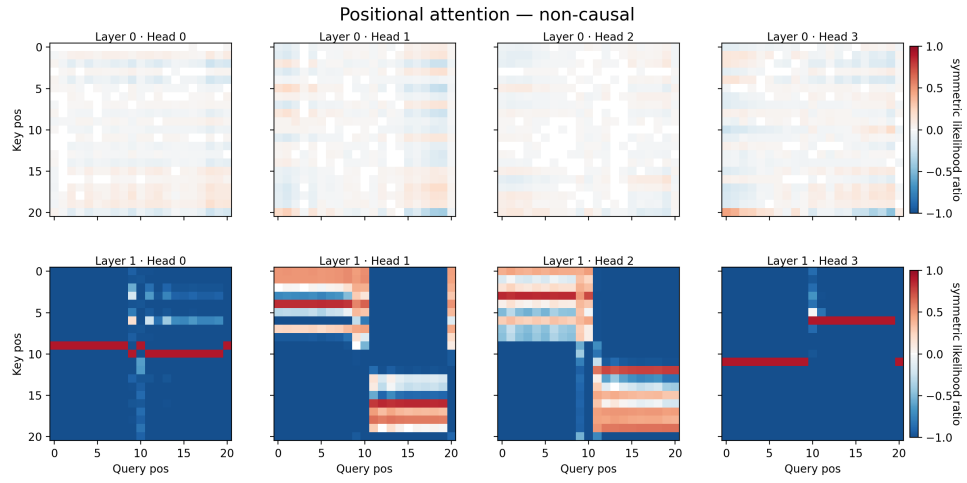
$$\text{SLR}(k | q) = \frac{K_{\text{eff}}(q) p(k | q) - 1}{K_{\text{eff}}(q) p(k | q) + 1} \in (-1, 1).$$

Here $\text{SLR}(k | q) = 0$ corresponds to uniform attention, positive values indicate higher-than-uniform weighting, and negative values indicate suppression relative to the baseline. This *symmetric log-ratio* (SLR) mapping preserves the ordering of the original likelihood ratios, compresses large deviations, and provides a bounded scale suitable for visual comparison across layers and heads.

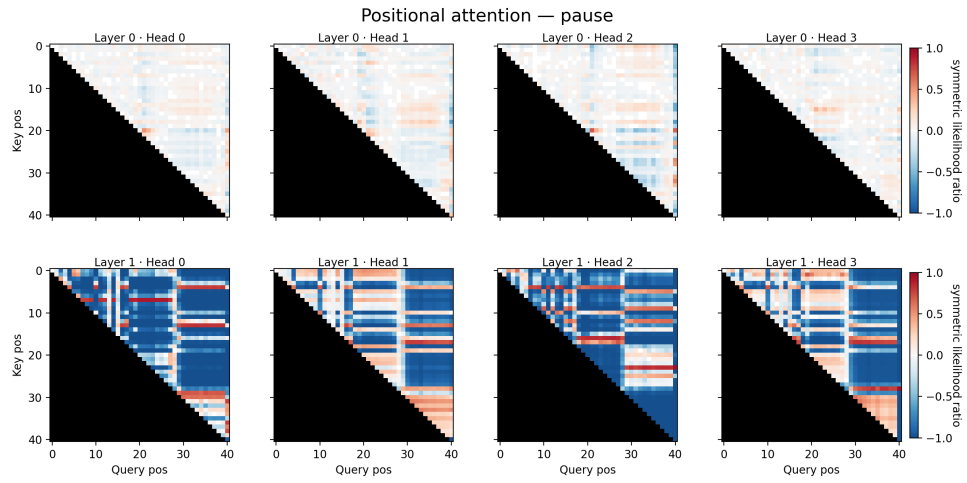
In Figure 3 we can see that the unmasked and pause tokens models attend very uniformly to all tokens in the first layer, as would be expected if it were implementing a method similar to the 2-layer threshold circuit described in Appendix E. In the second layer, the first two heads in the pause token model attend largely to pause token positions (rather than input tokens), again corresponding to the circuit. However, it does not attend heavily to all the pause tokens, suggesting some divergence from this mechanism. The non-causal model attends over all tokens when combining head 1 and head 2 in layer 1, as would be expected if it were using the threshold circuit mechanism. The causal model has less structured attention patterns that elude easy analysis.



(a) Causally-masked model.



(b) Unmasked model.



(c) Causally masked model with pause tokens.

Figure 3: Symmetric attention maps, demonstrating deviation from the uniform distribution for each query position.