

Counting Like Transformers: Compiling Temporal Counting Logic Into Softmax Transformers

Anonymous authors

Paper under double-blind review

Abstract

Deriving formal bounds on the expressivity of transformers, as well as studying transformers that are constructed to implement known algorithms, are both effective methods for better understanding the computational power of transformers. Towards both ends, we introduce the temporal counting logic $\mathbf{K}_t[\#]$ alongside the RASP variant **C-RASP**. We show they are equivalent to each other, and that together they are the best-known lower bound on the formal expressivity of future-masked soft attention transformers with unbounded input size. We prove this by showing all $\mathbf{K}_t[\#]$ formulas can be compiled into these transformers. As a case study, we demonstrate on paper how to use **C-RASP** to construct simple transformer language models that, using greedy decoding, can only generate sentences that have given properties formally specified in $\mathbf{K}_t[\#]$.

1 Introduction

What problems can transformers (Vaswani et al., 2017) solve, what problems can they not solve, and how can we prove it? Formal logic, in connection with programming language theory, formal language theory, and finite model theory, provides a strong framework in which to investigate these questions.

Previous theoretical work, as surveyed by Strobl et al. (2023), has advanced our understanding of transformers immensely. However, it has not fully answered these questions. Much of this work only considers modifications of transformers, like average-hard attention transformers (AHATs) (Barceló et al., 2024) or unique-hard attention transformers (UHATs) (Angluin et al., 2023), which are not known to be either a subset or superset of standard, soft-attention transformers. At the same time, programming languages like RASP (Weiss et al., 2021) propose a human-readable language with which to understand transformer computations. However, current languages compile into transformers that appear to be more powerful than standard transformers (Weiss et al., 2021), or require approximations and restrictions on input length (Lindner et al., 2023) to do so.

In this paper, we target soft attention transformers (that is, transformers as originally defined (Vaswani et al., 2017) and as used in practice). We prove that future-masked soft attention transformer encoders, with no restriction on input length, can recognize all the formal languages defined by formulas of $\mathbf{K}_t[\#]$, a temporal counting logic. Along the way we develop a RASP variant called **C-RASP**, equivalent to $\mathbf{K}_t[\#]$. Both are, to our knowledge, the tightest-known lower bound on the expressivity of soft attention transformer encoders. Our contributions are as follows:

- We define **C-RASP**, the first variant of RASP that provably compiles into future-masked soft attention transformer encoders with no restrictions on the input length.
- We prove that **C-RASP** is equivalent to $\mathbf{K}_t[\#]$.
- We prove that the previously best lower bound, $\text{FOC}[+;\text{MOD}]$ (Chiang et al., 2023), is strictly less expressive than $\mathbf{K}_t[\#]$.
- We prove that transformers which use fixed-precision numbers (as real-world transformers do) can be compiled back to $\mathbf{K}_t[\#]$.
- We explain how **C-RASP** can be used to construct simple transformer decoder language models with formally specifiable behavior.

2 Background

Previous theoretical work on the expressivity of transformers has related them to a variety of automata, circuit classes, and logics, all under varying assumptions (Strobl et al., 2023). Here, we focus on using logics to characterize soft attention transformer encoders with future masking. In particular, these encoders perform the same computations as one step of a transformer decoder. Decoder-only models like GPT (OpenAI, 2023) currently dominate applications of transformers, and empirical work has noted significant limits, and perplexing behavior in transformer models. Thus, we believe theoretical analysis can provide valuable insight into these models.

The previous best upper bound on log-precision transformers (which are argued to closely approximate real-life behavior) is TC^0 (Merrill and Sabharwal, 2023). The previous best lower bound is $\text{FOC}[+; \text{MOD}]$ (Chiang et al., 2023). We strengthen this lower bound using $\mathbf{K}_t[\#]$, and in the process we define **C-RASP**, a new variant of the programming language RASP.

2.1 RASP and Tracr

Implementing algorithms in transformers using human-readable programming languages gives researchers and engineers a deeper understanding of the computations transformers can perform. For example, this perspective has been used by Zhou et al. (2024) to shed light on the length-generalization capabilities of transformers. These programming languages promise to compile into transformers that implement known algorithms, which are therefore interpretable by construction. However, existing examples make several unrealistic assumptions about transformers, rendering them inappropriate for compilation into standard transformers.

The primary example is RASP (Weiss et al., 2021), which makes three strong assumptions. First, the transformers that RASP compiles into use average-hard attention, which are not known to be exactly simulated by soft attention transformers (although average-hard attention behavior has been observed to be learned approximately, in practice (Merrill et al., 2021)). Second, the attention weights (selectors) are not restricted to be dot-products of query and key vectors; this allows compilation of expressions involving arbitrary binary predicates like $x = y$ or $x < y$. Third, they assume position-wise feed-forward networks can implement any computable function, with the rationale that any continuous function can be approximated by the universal approximation theorem (Hornik et al., 1989).

Building on RASP, Tracr (Lindner et al., 2023) compiles a subset of RASP into standard transformers. It compiles RASP selectors to dot-product attention, although this requires a syntactic restriction on selectors and a maximum string length, and it compiles RASP element-wise operations to ReLU FFNs, though only approximately.

Furthermore, neither of these consider layer normalization, which Brody et al. (2023) show contributes to the expressivity of transformers.

Here, we define a variant of RASP that has more restrictions, but that can be compiled exactly into a soft attention transformer encoder. Our variant, **C-RASP**, is based on the temporal counting logic $\mathbf{K}_t[\#]$.

2.2 $\text{FOC}[+; \text{MOD}]$

Counting logics are a rich area (van Benthem and Icard, 2023) of logic, whose connection with transformers has been noted by Chiang et al. (2023) and Barceló et al. (2024). In essence, uniform attention patterns – where attention is spread evenly across positions – can very naturally simulate counting terms. Chiang et al. (2023) define a variant of first-order logic with counting quantifiers, called $\text{FOC}[+; \text{MOD}]$, and prove that, on the one hand, any sentence of $\text{FOC}[+; \text{MOD}]$ can be translated into an equivalent soft attention transformer encoder, and on the other hand, any *fixed-precision* soft attention transformer encoder can be translated into an equivalent sentence of $\text{FOC}[+; \text{MOD}]$.

However, $\text{FOC}[+; \text{MOD}]$ seems somewhat underpowered. It has a normal form that uses only one quantifier alternation ($\exists x. \exists^{=x} p. \dots$) and only one position variable. This means the equivalent transformer only has depth 2, and only uses the output of self-attention at one position. By considering an ordering on positions (and future-masking on the corresponding transformers) we derive a much better lower-bound result.

2.3 Temporal logic

A technical challenge when simulating variants of first-order logic with transformers is that a formula with k free variables, each of which is interpreted as a position in from 1 to n , would seem to correspond to a tensor of n^k values in the corresponding transformer, but transformers only have n values at each layer and n^2 values in the attention weights.

Whereas $\text{FOC}[+; \text{MOD}]$ avoids this difficulty by using a normal form with only one variable, [Angluin et al. \(2023\)](#) and [Barceló et al. \(2024\)](#) avoid it by relying on linear temporal logic.

Temporal logics ([Gabbay et al., 1980](#)) have been widely adopted as tools for the formal verification of state properties during the execution of programs over time. Intuitively, temporal logics can be used to formalize statements such as the following:

Until the first train arrived at the gate, the bar remained lowered.
My arm has been sore since Tuesday.
At no point will the temperature go below zero.

More abstractly, we can also use the syntax of temporal logic to specify the occurrence of symbols in a string w .

Until the first symbol t , w contained only l 's.
Only the symbol s has appeared since position 2 in w .
At no position does w contains a z .

We believe that temporal logics are a very appropriate specification language for thinking about masked self-attention. Firstly, it is quite neat that the temporal accessibility relation – properties at time i can only depend on times $j \leq i$ – corresponds exactly to how future-masking works in transformers. Secondly, temporal logics are highly-utilized in the field of formal methods ([Fisher, 2011](#)), so solidifying this connection with transformers can inspire the sharing of important ideas across disciplines.

3 The Temporal Counting Logic $\mathbf{K}_t[\#]$

In this section, we define the temporal counting logic $\mathbf{K}_t[\#]$.

3.1 Definitions

The temporal logic we target here is the past fragment of the Minimal Tense Logic ([Rescher and Urquhart, 2012](#)), with counting terms ([Barceló et al., 2024](#)). It can also be thought of as a modal logic with arithmetic constraints, like that of [Demri and Lugiez \(2010\)](#), simply restricted to the setting where the structures are strings.

We present the syntax of $\mathbf{K}_t[\#]$ in Backus–Naur form:¹

$$\begin{aligned} F &::= Q_a \mid \neg F \mid F \wedge F \mid C \leq C \mid \top \\ C &::= \#[F] \mid C + C \mid C - C \mid 1 \end{aligned}$$

¹We pronounce $\mathbf{K}_t[\#]$ as “K-t-sharp”. The logic \mathbf{K}_t is E.J. Lemmon’s minimal tense logic ([Rescher and Urquhart, 2012](#)), where \mathbf{K} is the minimal modal logic, and the t refers to “tense”, indicating that our structures are linear, like timelines or strings. Additionally, $\#$ is the modal counting operator, which is fairly standard notation in counting logics.

In temporal logics, formulas and terms are written without arguments because they are always interpreted with respect to a structure at a specified position. In our setting, they are interpreted with respect to a string w at a position i , where $w = w_1w_2 \cdots w_n$ and $i \in [1, n]$. A count term C is interpreted as an integer, written $C^{w,i}$ and defined as follows.

$$\begin{aligned} \#[F]^{w,i} &= |\{j \in [1, i] \mid w, j \models F\}| \\ (C_1 + C_2)^{w,i} &= C_1^{w,i} + C_2^{w,i} \\ (C_1 - C_2)^{w,i} &= C_1^{w,i} - C_2^{w,i} \\ 1^{w,i} &= 1 \end{aligned}$$

As syntactic sugar, we allow any natural number, which implicitly is defined as a sum of 1's. Similarly, 0 can be defined as $\#[\neg\top]$, and i as $\#[\top]$. Next, the interpretation of a formula F at position i , written $w, i \models F$, defined as follows:

$$\begin{aligned} w, i \models Q_a &\iff w_i = a \\ w, i \models \neg F &\iff w, i \not\models F \\ w, i \models F_1 \wedge F_2 &\iff w, i \models F_1 \text{ and } w, i \models F_2 \\ w, i \models C_1 \leq C_2 &\iff C_1^{w,i} \leq C_2^{w,i} \\ w, i \models \top &\text{is always the case} \end{aligned}$$

We say that a string w with length n *end-satisfies* ϕ a formula of $\mathbf{K}_t[\#]$ whenever $w, n \models \phi$ (Maler and Pnueli, 1990). The language defined by ϕ is the set of all strings end-satisfied by ϕ . As a final note, whenever w is implicit, we can write $F(i)$ which is True iff $w, i \models F$ and also write $C(i)$ which is equal to $C^{w,i}$.

3.2 Examples

Although an exact characterization of $\mathbf{K}_t[\#]$ is not currently known, we can see that it can define a variety of regular, context-free, and non-context-free languages.

Language	Formula
a^*b^*	$\#[Q_a \wedge (\#[Q_b] \geq 1)] = 0$
$a^*b^*a^*$	$\#[Q_b \wedge \#[Q_a \wedge (\#[Q_b] \geq 1)] \geq 1] = 0$
$a^n b^n c^n$	$\#[Q_b \wedge (\#[Q_c] = 0)] = \#[Q_a \wedge (\#[Q_b \vee Q_c] = 0)]$ $= \#[Q_a] = \#[Q_b] = \#[Q_c]$
Dyck-1	$(\#[Q_{\lceil}] = \#[Q_{\rceil}]) \wedge (\#[\#[Q_{\rceil}]] > \#[Q_{\lceil}]) = 0$
hello	$\#[\top] = 5 \wedge Q_o \wedge \#[Q_l \wedge \#[Q_e \wedge \#[Q_h] = 1]] = 1] = 2$

3.3 Modal Depth

Definition 3.1. The *modal depth* of a formula ϕ or term C , which we notate as $md(\phi)$, is the maximum level of nesting of $\#$ terms. That is,

$$\begin{aligned} md(Q_a) &= 0 & md(1) &= 0 \\ md(\neg\phi) &= md(\phi) & md(\#[\phi]) &= 1 + md(\phi) \\ md(\phi_1 \wedge \phi_2) &= \max(md(\phi_1), md(\phi_2)) & md(C_1 + C_2) &= \max(md(C_1), md(C_2)) \\ md(C_1 \leq C_2) &= \max(md(C_1), md(C_2)) \end{aligned}$$

The following construction gives some intuition on the effect of modal depth.

Lemma 3.2. For every string s of length n , there exists a formula ϕ_a of modal depth n such that $w \models \phi_a$ if and only if w contains s as a subsequence.

Proof. Let $s = s_1 s_2 \cdots s_n$. Then define $\phi_s := \tau_n \geq 1$ where

$$\begin{aligned} \tau_1 &:= \#[Q_{s_1}] \\ \tau_{k+1} &:= \begin{cases} \#[Q_{s_{k+1}} \wedge \tau_k \geq 1] & s_k \neq s_{k+1} \\ \#[Q_{s_{k+1}} \wedge \tau_k \geq 2] & s_k = s_{k+1} \end{cases} \end{aligned}$$

Verification is left as an exercise for the reader. \square

As a consequence of the above and [Theorem 5.7](#), we see that masked soft attention transformers can recognize all the piecewise testable languages ([Klíma and Polák, 2010](#)), a subset of the star-free languages. Recall, however, that $\mathbf{K}_t[\#]$ can express much more: for example, the context sensitive language $a^n b^n c^n$.

4 C-RASP

Before translating $\mathbf{K}_t[\#]$ formulas into transformers, we, follow in the footsteps of [Weiss et al. \(2021\)](#) to define a variant of RASP called **C-RASP**. In proofs, we generally prefer the more compact syntax of $\mathbf{K}_t[\#]$, but for writing programs, we use **C-RASP**.

4.1 Definitions

Definition 4.1 (C-RASP). A **C-RASP** program is defined as a sequence P_1, \dots, P_n of **C-RASP** operations. There are two types of operations:

Boolean-Valued Operations		Count-Valued Operations	
Initial	$P(i) := Q_a(i) \text{ for } a \in \Sigma$	Counting	$C(i) := \#[j \leq i] P(j)$
Boolean	$P(i) := \neg P_1(i)$ $P(i) := P_1(i) \wedge P_2(i)$	Conditional	$C(i) := P(i) ? C_1(i) : C_2(i)$
Comparison	$P(i) := C_1(i) \leq C_2(i)$	Addition	$C(i) := C_1(i) + C_2(i)$
Constant	$P(i) := 1$	Subtraction	$C(i) := C_1(i) - C_2(i)$
		Min/Max	$C(i) := \min(C_1(i), C_2(i))$ $C(i) := \max(C_1(i), C_2(i))$
		Constant	$C(i) := 1$

Counting operations count the positions $j \leq i$ such that $P(j)$ holds, returning the sum. We could also extend the syntax to use an expression $P(i, j)$ which depends on both i and j ; this can be transformed to $P(j)$ since our logic has only unary predicates, as shown in [Lemma B.1](#). Conditional operations return C_1 if P holds, and C_2 otherwise.

By convention, when using a **C-RASP** program to recognize languages, we use the value of the *last* operation, which must be Boolean-valued, at the last position, to determine acceptance. That is, if the program is run on input w with length n , and the last operation is D , then we accept w if and only if $D(n)$ is true.

Example 4.2. We present a program to recognize Dyck-1 as an example. More annotated examples can be found in [Appendix B.2](#)

$C_\ell(i) := \#[j \leq i] Q_\ell(j)$	The number of (up to position i
$C_\gamma(i) := \#[j \leq i] Q_\gamma(j)$	The number of) up to position i
$V(i) := C_\ell(i) < C_\gamma(i)$	Violation: there are more) than (
$C_V(i) := \#[j \leq i] V(j)$	The number of Violations
$M(i) := C_V(i) = 0$	Matched: zero Violations
$B(i) := C_\ell(i) = C_\gamma(i)$	Balanced: same number of (and)
$D(i) := M(i) \wedge B(i)$	String is Matched and Balanced

4.2 C-RASP and $K_t[\#]$

C-RASP may have a more convenient syntax to write programs in, but **C-RASP** programs and $K_t[\#]$ formulas are exactly equivalent in expressivity.

Theorem 4.3. *A C-RASP program recognizes language L iff a $K_t[\#]$ formula defines L . More precisely, given alphabet Σ , for any $K_t[\#]$ formula ϕ there is a C-RASP program P such that $w \in \Sigma^*$ end-satisfies ϕ iff w is accepted by P , and vice versa.*

Proof. See [Appendix B.3](#). □

5 From $K_t[\#]$ to Masked-Attention Transformers

In this section, we show how to compile $K_t[\#]$ into transformers. It would also be possible to translate directly from **C-RASP** to transformers; this would use fewer dimensions, but more layers.

5.1 Transformers

We assume familiarity with the transformer architecture ([Vaswani et al., 2017](#)), and only review the basic definitions particular to our setting. To simplify our analysis, we do not consider positional encodings at first, deferring them to [Section 6](#).

Definition 5.1. (Word Embeddings) Let w be an input string of length n over a finite alphabet Σ . We prepend to w a special symbol BOS, which we assume is not in Σ . We abbreviate $n + 1$ as n' . A word embedding with dimension d over Σ is a function $WE: \Sigma \cup \{\text{BOS}\} \rightarrow \mathbb{R}^d$ applied position-wise to a string of length n to form a tensor in $\mathbb{R}^{d \times n'}$.

Definition 5.2 (Transformer Block). A transformer block B , defined with a dimension d , specifies a function $B: \mathbb{R}^{d \times n'} \rightarrow \mathbb{R}^{d \times n'}$ that computes

$$\begin{aligned} B(A) &= \text{LN}_2(\text{FFN}(A') + A') \\ A' &= \text{LN}_1(\text{SA}(A) + A) \end{aligned}$$

where SA denotes a self-attention layer, by the standard definition ([Vaswani et al., 2017](#)), FFN denotes a two layer feed-forward neural network with ReLU activations between the layers, and LN_1 and LN_2 denote position-wise applications of LayerNorm. This setup is commonly referred to as a “post-norm” block.

5.2 Overview of the translation

The input and output of a transformer block are tensors in $\mathbb{R}^{d \times n'}$. The resulting sequence of tensors across transformer blocks is sometimes referred to as the “residual stream” ([Elhage et al., 2021](#)). We store the values of each subformula or count term of a $K_t[\#]$ formula in a different dimension of the residual stream.

Let $A \in \mathbb{R}^{d \times n'}$ be a tensor in the residual stream. Formulas ϕ_k are stored as two rows of A :

$$A_{2k-1:2k,*} = \begin{bmatrix} 1 & -2\phi_k(1) + 1 & -2\phi_k(2) + 1 & \cdots & -2\phi_k(n) + 1 \\ -1 & 2\phi_k(1) - 1 & 2\phi_k(2) - 1 & \cdots & 2\phi_k(n) - 1 \end{bmatrix}.$$

Similarly, count terms C_k are stored as:

$$A_{2k-1:2k,*} = \begin{bmatrix} 0 & -\frac{C_k(1)}{2} & -\frac{C_k(2)}{3} & \cdots & -\frac{C_k(n)}{n'} \\ 0 & +\frac{C_k(1)}{2} & +\frac{C_k(2)}{3} & \cdots & +\frac{C_k(n)}{n'} \end{bmatrix}.$$

The division of $C(i)$ by $(i + 1)$ is a consequence of the fact that attention computes an average rather than a sum. Dealing with these divisions is a common feature of many transformer constructions. In contrast to other constructions that undo the divisions using

nonstandard embeddings (Pérez et al., 2021; Barceló et al., 2024) or nonstandard versions of LayerNorm (Merrill and Sabharwal, 2024), our construction uses no position embeddings and only standard LayerNorm. A minor consequence of our handling of comparison (Appendix C.3) is that while Boolean values are preserved throughout the computation, integer values can get overwritten.

The reason for representing every value as two transformer activation values is to account for LayerNorm. It ensures that all feature vectors have zero mean, so LayerNorm only applies a position-wise scaling factor. When necessary, we describe how to use LayerNorm to remove this scaling factor in Appendix C.3.

Observe that for subformulas, our convention states the BOS position is always false. Consequently, for count terms, the BOS position is always 0. We can ensure this with a feed-forward layer.

Lemma 5.3. *Using the word embedding and a single feed-forward layer, we can set the BOS position to False, without disturbing the Boolean value at any other position.*

Proof. See Appendix C.1. □

5.3 Counting using masked uniform attention

Count terms in $K_t[\#]$ can be simulated by uniform self-attention layers.

Lemma 5.4. *Let $A_{2k-1:2k'}$ store a Boolean vector as defined above. For any i , let $C_{k,i}$ be the number of positions $j \leq i$ such that $A_{2k-1:2k',j}$ is True. Then there is a transformer block that computes, at each position i , and in two other dimensions $2k' - 1, 2k'$, the values $-\frac{C_{k,i}}{i+1}$ and $\frac{C_{k,i}}{i+1}$.*

Proof. See Appendix C.2 for proof and pictures. □

5.4 Other operations using position-wise feed-forward networks

All other $K_t[\#]$ formulas and terms can be simulated by feed-forward layers.

Lemma 5.5. *The following position-wise operations can be simulated by a single transformer block, using existing dimensions as input and a fresh dimension as output: addition (+), subtraction (-), comparison (\leq), and Boolean operations (\wedge , \neg).*

Proof. See Appendix C.3. □

5.5 Compiling $K_t[\#]$ formulas into masked uniform attention transformers

Definition 5.6. Fix an alphabet Σ , and assume that the symbol BOS is not in Σ . We say a masked soft attention transformer T (as a composition of blocks $T = B_b \circ \dots \circ B_1 \circ WE$) with d dimensions *simulates* a $K_t[\#]$ formula ϕ if for every input $w \in \Sigma^*$ with length n and every subformula ψ_k of ϕ , there is some dimension d_k such that

$$[T(\text{BOS} \cdot w)]_{2d_k-1:2d_k,i+1} = \begin{cases} \begin{bmatrix} -1 \\ +1 \end{bmatrix} & \text{if } w, i \models \psi_k \\ \begin{bmatrix} +1 \\ -1 \end{bmatrix} & \text{otherwise.} \end{cases}$$

Theorem 5.7. *For every $K_t[\#]$ formula ϕ , there exists a soft attention transformer which simulates ϕ . Moreover, the transformer will have at most $4nd(\phi) + 1$ blocks.*

Proof. We induct on the modal depth of ϕ . If ϕ is of modal depth 0, it must be a Boolean combination of Q_a formulas. Then, each Q_a can be simulated in some pair of dimensions by WE, and by Lemma 5.5 (Boolean) we can append one block which computes any given Boolean combination of Q_a formulas.

Assume for formulas ψ of modal depth m we can construct a transformer which simulates ψ , consisting of m sequences of 4 blocks each:

1. One to ensure the BOS position is false in every Boolean vector
2. One to simulate $\#$ terms over existing formulas
3. One to simulate comparisons of linear combinations of $\#$ terms
4. One to compute Boolean combinations of existing formulas.

Additionally, each comparison block only references the $\#$ terms computed in its immediately preceding block, and no others.

Now, let ϕ be a $\mathbf{K}_t[\#]$ formula of modal depth $m + 1$. By the inductive hypothesis, for each subformula ψ_k of modal depth m there is a transformer T_k which simulates it. Moreover, T_k has depth at most $4x - 1$. By definition, ϕ is a Boolean combination of:

- Subformulas of modal depth at most m
- Subformulas of the form $\sum_{k \in K_1} c_k \cdot \#[\psi_k] \leq \sum_{k \in K_2} c_k \cdot \#[\psi_k]$, where K_1 and K_2 are two disjoint sets of indices, c_k are integers, and ψ_k are subformulas of modal depth m .

We claim that all transformers satisfying the inductive hypothesis can be composed in parallel. We show how to do this in [Appendix C.4](#). Thus we compose all T_k into a single transformer, T_m . This allows us to reference all ψ_k in later steps.

1. Add one block that ensures the BOS position is false, as described in [Appendix C.1](#)
2. Add a block to compute $\#[\psi_k]$ for all relevant ψ_k , as described in [Lemma 5.4](#)
3. Add a block to compute all formulas of the form $\sum_{k \in K_1} c_k \cdot \#[\psi_k] \leq \sum_{k \in K_2} c_k \cdot \#[\psi_k]$, as described in [Lemma 5.5](#) (Arithmetic Comparison)
4. add one block to compute all Boolean combinations of the above subformulas as necessary, as described in [Lemma 5.5](#) (Boolean)

This constructs a transformer which simulates ϕ of modal depth $m + 1$ and satisfies the inductive hypothesis. Moreover, since we have added 4 blocks, the new transformer has depth at most $4(m + 1) + 1$. \square

6 Relationship to other formalisms

The expressivity of $\mathbf{K}_t[\#]$ can be characterized in relation to other bounds on transformer formal expressivity. To begin with, the previous best lower bound on soft attention transformer encoders was $\text{FOC}[+; \text{MOD}]$, by [Chiang et al. \(2023\)](#). However, $\mathbf{K}_t[\#]$ forms a tighter lower bound due to its ability to model order.

Lemma 6.1. *$\mathbf{K}_t[\#]$ formulas can simulate $\text{FOC}[+]$ formulas, and $\mathbf{K}_t[\#; \text{MOD}]$ can simulate $\text{FOC}[+; \text{MOD}]$. In general, for any set \mathcal{P} of unary predicates, the extension $\text{FOC}[+; \mathcal{P}]$ is contained in $\mathbf{K}_t[\#; \mathcal{P}]$.*

Proof. See [Appendix B.3](#). \square

We’ve seen that **C-RASP** can define Dyck-1 ([Example 4.2](#)), and [Bhattachamishra et al. \(2020\)](#) prove that transformers, under the same assumption as ours, can express Dyck-1 via a very similar construction. However, it’s easy to show that $\text{FOC}[+; \text{MOD}]$ cannot define Dyck-1. Thus, $\mathbf{K}_t[\#]$ is a more realistic lower bound for transformers than $\text{FOC}[+; \text{MOD}]$:

Proposition 6.2. *There is no sentence of $\text{FOC}[+; \text{MOD}]$ that defines Dyck-1.*

Proof. Suppose that such a sentence σ exists. [Chiang et al. \(2023\)](#) show that there exists an M such that σ cannot distinguish between two strings that differ only by swapping symbols exactly M positions apart. Since $(^M)^M \in \text{Dyck-1}$ and $(^{M-1})^{M-1} \notin \text{Dyck-1}$ but σ cannot distinguish them, we have a contradiction. \square

On the other hand, $\mathbf{K}_t[\#]$ is a strict subset of $\text{LTL}(\mathbf{C}, +)$ ([Barceló et al., 2024](#)). However, $\text{LTL}(\mathbf{C}, +)$ has only been shown to be a lower bound on AHATs, which are not known to be either a subset or superset of standard, soft-attention transformers. Furthermore, they allows some layers to be future-masked and some to not be, which does not reflect the standard setup. Understanding the precise connection between AHATs and soft attention transformers is left for future exploration, and is expected to require modifications to the standard architecture in order to derive an exact inclusion.

7 From Fixed-Precision Transformers to $\mathbf{K}_t[\#]$

In practice, transformers use fixed-precision numbers. We adapt the proof that $\text{FOC}[+]$ can simulate fixed-precision soft attention transformers (Chiang et al., 2023) to show that $\mathbf{K}_t[\#]$ can simulate fixed-precision, masked soft attention transformers. If $\mathbf{K}_t[\#]$ is extended with modular predicates, it can also simulate sinusoidal positional encodings.

Theorem 7.1. $\mathbf{K}_t[\#]$ can simulate fixed-precision masked soft attention transformers without positional encodings, and $\mathbf{K}_t[\#; \text{MOD}]$ can simulate fixed-precision masked soft attention transformers with sinusoidal positional encodings.

Proof. See Appendix D. □

8 Autoregressive C-RASP Programs

With a simple extension, we can use **C-RASP** to construct decoder language models.

Definition 8.1 (C-RASP Language Model). Assume a **C-RASP** program over alphabet Σ has vectors $N_a(i)$ defined for each $a \in \Sigma \cup \{\text{EOS}\}$. Then a **C-RASP** program can be converted to an autoregressive language model with greedy decoding in the following manner

1. Translate the program into a transformer.
2. Append a linear layer, which selects the dimensions which simulate $N_a(i)$
3. Consider last position as a uniform probability distribution: dimensions with $N_a(|w|) = 1$ will have the same maximum probability score, and all those with $N_a(|w|) = 0$ will have the minimum probability.
4. Run the transformer on w , and then select an a such that the N_a dimension holds the maximum probability, with arbitrary tie-breaking. Append a to w , and repeat until EOS is selected.

We say a **C-RASP** language model assigns nonzero probability p to word $w = w_0 \cdots w_{n-1}$ iff $w, i \models N_{w_{i+1}}$ for all $0 \leq i < n-1$, and $w, n-1 \models N_{\text{EOS}}$. We say a **C-RASP** Language Model recognizes language L whenever it assigns nonzero probability to w iff $w \in L$.

In logical terms, we can construct a **C-RASP** language model to recognize L iff there exists a formula ϕ of $\mathbf{K}_t[\#]$ which recognizes L and for all $a \in \Sigma$ we can define formulas N_a such that $w \models N_a \iff \exists w'. waw' \models \phi$. In this case, $N_{\text{EOS}} = \phi$

Example 8.2. Append to the end of Dyck program in Example 4.2 the following operations:

$$\begin{aligned} N_{\langle}(i) &:= \neg Q_{\text{EOS}}(i) \\ N_{\rangle}(i) &:= \neg Q_{\text{EOS}}(i) \wedge C_{\rangle}(i) < C_{\langle}(i) \\ N_{\text{EOS}}(i) &:= D(i) \end{aligned}$$

Corollary 8.3. For every piecewise testable language L , there exists a **C-RASP** Language Model which recognizes L .

Proof. This is an immediate consequence of Lemma 3.2. For any piecewise testable language L we can define a formula ϕ of $\mathbf{K}_t[\#]$ which recognizes L . Next, observe that it is always the case that $\exists w'. waw' \models \phi$. Finally, define $N_{\text{EOS}} = \phi$ and $N_a = \top$ for all $a \in \Sigma$. □

9 Concluding Remarks

We have introduced the temporal counting logic $\mathbf{K}_t[\#]$ alongside the RASP variant **C-RASP** and proved that they are the best-known lower bound on the expressivity of future-masked soft attention transformers, with unbounded input size. Unlike previous results, we have made minimal extra assumptions about transformers, so all formulas in $\mathbf{K}_t[\#]$ can compile directly into standard transformers. As such, an implementation of **C-RASP** should prove appropriate for constructing transformers to run experiments on, and further theoretical analysis of $\mathbf{K}_t[\#]$ and its extensions should shed light on the expressive power of transformers.

References

- Dana Angluin, David Chiang, and Andy Yang. Masked hard-attention transformers and Boolean RASP recognize exactly the star-free languages, 2023. URL <https://arxiv.org/abs/2310.13897>. arXiv:2310.13897.
- Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *Proceedings of ICLR*, 2018. URL https://openreview.net/forum?id=B1J_rgWRW.
- Pablo Barceló, Alexander Kozachinskiy, Anthony Widjaja Lin, and Vladimir Podolskii. Logical languages accepted by transformer encoders with hard attention. In *Proc. ICLR*, 2024. URL <https://openreview.net/forum?id=gbrHZq07mq>.
- Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of Transformers to recognize formal languages. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7096–7116, 2020. doi: 10.18653/v1/2020.emnlp-main.576. URL <https://aclanthology.org/2020.emnlp-main.576>.
- Shaked Brody, Uri Alon, and Eran Yahav. On the expressivity role of LayerNorm in Transformers’ attention. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14211–14221, 2023. doi: 10.18653/v1/2023.findings-acl.895. URL <https://aclanthology.org/2023.findings-acl.895>.
- David Chiang, Peter Cholak, and Anand Pillay. Tighter bounds on the expressivity of transformer encoders. In *Proc. ICML*, 2023. URL <https://arxiv.org/abs/2301.10743>.
- Stéphane Demri and Denis Lugiez. Complexity of modal logics with Presburger constraints. *Journal of Applied Logic*, 8(3):233–252, 2010. doi: 10.1016/j.jal.2010.03.001.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL <https://transformer-circuits.pub/2021/framework/index.html>.
- Michael Fisher. *An introduction to practical formal methods using temporal logic*. John Wiley & Sons, 2011.
- Dov Gabbay, Amir Pnueli, Saharon Shelah, and Jonathan Stavi. On the temporal analysis of fairness. In *Proceedings of the 7th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL)*, pages 163–173, 1980. doi: 10.1109/FSCS.1990.89589.
- Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. URL 10.1016/0893-6080(89)90020-8.
- Ondřej Klíma and Libor Polák. Hierarchies of piecewise testable languages. *International Journal of Foundations of Computer Science*, 21(4):517–533, 2010. doi: 10.1142/S0129054110007404.
- David Lindner, János Kramár, Matthew Rahtz, Thomas McGrath, and Vladimir Mikulik. Tracr: Compiled transformers as a laboratory for interpretability. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 37876–37899, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/771155abaae744e08576f1f3b4b7ac0d-Abstract-Conference.html.
- Oded Maler and Amir Pnueli. Tight bounds on the complexity of cascaded decomposition of automata. In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 672–682, 1990. doi: 10.1109/FSCS.1990.89589.

- William Merrill and Ashish Sabharwal. A logic for expressing log-precision transformers. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 52453–52463, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/a48e5877c7bf86a513950ab23b360498-Abstract-Conference.html.
- William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL <https://openreview.net/forum?id=NjNGlPh8Wh>.
- William Merrill, Vivek Ramanujan, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. Effects of parameter norm growth during transformer training: Inductive bias from gradient descent. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1766–1781, 2021. doi: 10.18653/v1/2021.emnlp-main.133. URL <https://aclanthology.org/2021.emnlp-main.133>.
- OpenAI. GPT-4 technical report, 2023. URL <https://arxiv.org/abs/2303.08774>. arXiv:2303.08774.
- Jorge Pérez, Pablo Barceló, and Javier Marinkovic. Attention is Turing-complete. *Journal of Machine Learning Research*, 22:75:1–75:35, 2021. URL <http://jmlr.org/papers/v22/20-302.html>.
- Nicholas Rescher and Alasdair Urquhart. *Temporal Logic*, volume 3. Springer Science & Business Media, 2012.
- Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. Transformers as recognizers of formal languages: A survey on expressivity, 2023. URL <https://arxiv.org/abs/2311.00208>. arXiv:2311.00208.
- Johan van Benthem and Thomas Icard. Interleaving logic and counting. *Bulletin of Symbolic Logic*, 29(4):503–587, 2023. doi: 10.1017/bsl.2023.30.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like Transformers. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 11080–11090, 2021. URL <https://proceedings.mlr.press/v139/weiss21a.html>.
- Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Josh Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can Transformers learn? A study in length generalization. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL <https://openreview.net/forum?id=AssIuHnmHX>.

A Acknowledgements

B Proofs Related to C-RASP

B.1 Extensions

Lemma B.1. Consider an extension of **C-RASP** in which the counting operation has the form

$$C(i) := \#_2 [j \leq i] F(i, j)$$

which allows F to be a Boolean combination of **C-RASP** operations $P(i)$ and $P(j)$ evaluated at both i and j . Any program with this extended operation is actually equivalent to a **C-RASP** program with only the normal counting operation.

Proof. Let F be a Boolean combination of P_1, \dots, P_k evaluated at either i or j . First, observe that $C(i) = C_1(i) + C_2(i) - C_3(i)$ where

$$\begin{aligned} C(i) &:= \#_2 [j \leq i] F_1(i, j) \vee F_2(i, j) \\ C_1(i) &:= \#_2 [j \leq i] F_1(i, j) \\ C_2(i) &:= \#_2 [j \leq i] F_2(i, j) \\ C_3(i) &:= \#_2 [j \leq i] F_1(i, j) \wedge F_2(i, j) \end{aligned}$$

Now, write F in DNF and split using the above. Now every single counting operation is of the form

$$C(i) := \#_2 [j \leq i] \left(\bigwedge_{x \in I} P_x(i) \wedge \bigwedge_{x \in J} P_x(j) \right)$$

Where I and J store the indices of **C-RASP** operations which depend on i and j , respectively, within the counting operation. Observe then that this is equivalent to

$$C(i) := \# [j \leq i] \left(\bigwedge_{x \in I \cap J} P_x(j) \right)$$

Thus, every $\#_2$ operation can be factored out as a sequence of normal $\#$ operations. \square

B.2 More C-RASP Examples

We sometimes put multiple **C-RASP** operations on the same line for brevity.

a^*b^* over $\Sigma = \{a, b\}$

$C_a(i) := \# [j \leq i] Q_a(j)$	# positions with a 's
$C_b(i) := \# [j \leq i] Q_b(j)$	# positions with b 's
$V(i) := Q_a(i) \wedge C_b(i) \geq 1$	Violation: an a has b 's preceding it
$C_V(i) := \# [j \leq i] V(j)$	# Violations
$Y(i) := C_V(i) = 0$	Zero Violations

$a^*b^*a^*$ over $\Sigma = \{a, b\}$

$C_a(i) := \# [j \leq i] Q_a(j)$	# positions with a 's
$C_b(i) := \# [j \leq i] Q_b(j)$	# positions with b 's
$BA(i) := Q_a(i) \wedge C_b(i) \geq 1$	A subsequence ba ends at i
$C_{ba}(i) := \# [j \leq i] BA(j)$	# ends of subsequence ba
$BAB(i) := Q_b(i) \wedge C_{ba}(i) \geq 1$	the subsequence bab ends at i
$C_{bab}(i) := \# [j \leq i] BAB(j)$	# ends of subsequence bab
$Y(i) := C_{bab}(i) = 0$	There are no subsequences bab

$a^n b^n c^n$ over $\Sigma = \{a, b, c\}$

$C_a(i) := \# [j \leq i] Q_a(j)$	# positions with a 's
$C_b(i) := \# [j \leq i] Q_b(j)$	# positions with b 's
$C_c(i) := \# [j \leq i] Q_c(j)$	# positions with c 's
$A(i) := C_b(i) + C_c(i) = 0$	No preceding b 's or c 's
$B(i) := C_c(i) = 0$	No preceding c 's
$C_A(i) := \# [j \leq i] Q_a(j) \wedge A(j)$	# a 's with no preceding b 's or c 's
$C_B(i) := \# [j \leq i] Q_b(j) \wedge B(j)$	# b 's with no preceding c 's
$G_a := C_A(i) = C_a(i)$	no a 's have preceding b 's or c 's
$G_b := C_B(i) = C_b(i)$	no b 's have preceding c 's
$G_{abc}(i) := C_a(i) = C_b(i) = C_c(i)$	same number of a 's, b 's, c 's
$Y(i) := G_a(i) \wedge G_b(i) \wedge G_{abc}(i)$	Correct order and number of symbols

$hello$ over $\Sigma = \{e, h, l, o\}$

$C_e(i) := \# [j \leq i] Q_e(j)$	# positions with e 's
$C_h(i) := \# [j \leq i] Q_h(j)$	# positions with h 's
$C_l(i) := \# [j \leq i] Q_l(j)$	# positions with l 's
$C_o(i) := \# [j \leq i] Q_o(j)$	# positions with o 's
$C_\Sigma(i) := \# [j \leq i] 1$	# symbols in string
$HE(i) := Q_e(i) \wedge C_h(i) = 1$	A subsequence he ends at i
$C_{he}(i) := \# [j \leq i] HE(j)$	# ends of subsequence he
$HEL(i) := Q_l(i) \wedge C_{he}(i) = 1$	A subsequence hel ends at i
$C_{hel}(i) := \# [j \leq i] HEL(j)$	# ends of subsequence hel
$HELLO(i) := Q_o(i) \wedge C_{hel}(i) = 2$	A subsequence $hello$ ends at i
$Y(i) := HELLO(i) \wedge C_\Sigma(i) = 5$	Length 5 and contains subsequence $hello$

As a potential point of clarification for the $HELLO(i)$ line, observe that if a string contains two positions that are the end of a subsequence hel , then that string must contain the subsequence $hell$.

B.3 Equivalence with $K_t[\#]$

Theorem 4.3. A **C-RASP** program recognizes language L iff a $K_t[\#]$ formula defines L . More precisely, given alphabet Σ , for any $K_t[\#]$ formula ϕ there is a **C-RASP** program P such that $w \in \Sigma^*$ end-satisfies ϕ iff w is accepted by P , and vice versa.

Proof. It is straightforward to translate $K_t[\#]$ formulas into **C-RASP** programs.

In the other direction, we induct on the length of **C-RASP** program $\mathcal{P} = P_1, \dots, P_n$. Assume that Boolean operations $P_k(i)$ are simulated by formulas \hat{P}_k , and count operations $C_k(i) = \# [j \leq i] V(j)$ are simulated by terms \hat{C}_k . This is straightforward, but there are many cases. The main idea is that whenever we have a conditional or min/max operation, we divide the formula into two cases depending on what the result of the operation should be.

- If P_{k+1} is a count-valued vector, it is not used for string acceptance as defined in 4.1. Thus the formula for this is the formula for the last Boolean-valued vector, by the IH.
- If $P_{k+1}(i) = Q_a(i)$, let $\hat{P}_{k+1} = Q_a$.

- If $P_{k+1}(i) = \neg P_\ell(i)$ let $\hat{P}_{k+1} = \neg \hat{P}_\ell$.
- If $P_{k+1}(i) = P_\ell(i) \wedge P_m(i)$ let $\hat{P}_{k+1} = \hat{P}_\ell \wedge \hat{P}_m$.
- If $P_{k+1}(i) = C_\ell(i) \leq C_m(i)$ we need to divide on cases to handle if $C_\ell(i)$ is a conditional or min/max, as $\mathbf{K}_t[\#]$ does not have these terms built in. This is straightforward, but there are many cases.
 - If $C_\ell(i) = \# [j \leq i] P_1(i)$, let $\hat{C}_\ell = \# [\hat{P}_1]$. Then:
 - If $C_m(i) = \# [j \leq i] P_2(i)$, let $\hat{P}_{k+1} = \hat{C}_\ell \leq \# [\hat{P}_2]$.
 - If $C_m(i) = P_2(i) ? C_3(i) : C_4(i)$, let $\hat{P}_{k+1} = (\hat{P}_2 \wedge \hat{C}_\ell \leq \hat{C}_3) \vee (\neg \hat{P}_2 \wedge \hat{C}_\ell \leq \hat{C}_4)$.
 - If $C_m(i) = \min(C_3(i), C_4(i))$, let $\hat{P}_{k+1} = (\hat{C}_3 \leq \hat{C}_4 \wedge \hat{C}_\ell \leq \hat{C}_3) \vee (\hat{C}_4 \leq \hat{C}_3 \wedge \hat{C}_\ell \leq \hat{C}_4)$.
 - If $C_m(i) = \max(C_3(i), C_4(i))$, let $\hat{P}_{k+1} = (\hat{C}_3 \geq \hat{C}_4 \wedge \hat{C}_\ell \leq \hat{C}_3) \vee (\hat{C}_4 \geq \hat{C}_3 \wedge \hat{C}_\ell \leq \hat{C}_4)$.
 - If $C_m(i) = c$, for $c \in \mathbb{N}$ let $\hat{P}_{k+1} = \hat{C}_\ell \leq c$.
 - If $C_\ell(i) = P_1(i) ? C_1(i) : C_2(i)$, then:
 - If $C_m(i) = \# [j \leq i] P_2(i)$, let $\hat{P}_{k+1} = (\hat{P}_1 \wedge \hat{C}_1 \leq \# [\hat{P}_2]) \vee (\neg \hat{P}_1 \wedge \hat{C}_2 \leq \# [\hat{P}_2])$.
 - If $C_m(i) = P(i) ? C_3(i) : C_4(i)$, let $\hat{P}_{k+1} = (\hat{P}_1 \wedge \hat{P}_2 \wedge \hat{C}_1 \leq \hat{C}_3) \vee (\hat{P}_1 \wedge \neg \hat{P}_2 \wedge \hat{C}_1 \leq \hat{C}_4) \vee (\neg \hat{P}_1 \wedge \hat{P}_2 \wedge \hat{C}_2 \leq \hat{C}_3) \vee (\neg \hat{P}_1 \wedge \neg \hat{P}_2 \wedge \hat{C}_2 \leq \hat{C}_4)$.
 - If $C_m(i) = \min(C_3(i), C_4(i))$, let $\hat{P}_{k+1} = (\hat{P}_1 \wedge \hat{C}_3 \leq \hat{C}_4 \wedge \hat{C}_1 \leq \hat{C}_3) \vee (\hat{P}_1 \wedge \hat{C}_3 \geq \hat{C}_4 \wedge \hat{C}_1 \leq \hat{C}_4) \vee (\neg \hat{P}_1 \wedge \hat{C}_3 \leq \hat{C}_4 \wedge \hat{C}_2 \leq \hat{C}_3) \vee (\neg \hat{P}_1 \wedge \hat{C}_3 \geq \hat{C}_4 \wedge \hat{C}_2 \leq \hat{C}_4)$.
 - If $C_m(i) = \max(C_3(i), C_4(i))$, let $\hat{P}_{k+1} = (\hat{P}_1 \wedge \hat{C}_3 \geq \hat{C}_4 \wedge \hat{C}_1 \leq \hat{C}_3) \vee (\hat{P}_1 \wedge \hat{C}_3 \leq \hat{C}_4 \wedge \hat{C}_1 \leq \hat{C}_4) \vee (\neg \hat{P}_1 \wedge \hat{C}_3 \geq \hat{C}_4 \wedge \hat{C}_2 \leq \hat{C}_3) \vee (\neg \hat{P}_1 \wedge \hat{C}_3 \leq \hat{C}_4 \wedge \hat{C}_2 \leq \hat{C}_4)$.
 - If $C_m(i) = c$, for $c \in \mathbb{N}$ let $\hat{P}_{k+1} = (\hat{P}_1 \wedge \hat{C}_1 \leq c) \vee (\neg \hat{P}_1 \wedge \hat{C}_2 \leq c)$.
 - If $C_\ell(i) = \min(C_1(i), C_2(i))$, then:
 - If $C_m(i) = \# [j \leq i] P_2(i)$, let $\hat{P}_{k+1} = (\hat{C}_1 \leq \hat{C}_2 \wedge \hat{C}_1 \leq \# [\hat{P}_2]) \vee (\hat{C}_1 \geq \hat{C}_2 \wedge \hat{C}_2 \leq \# [\hat{P}_2])$.
 - If $C_m(i) = P(i) ? C_3(i) : C_4(i)$, let $\hat{P}_{k+1} = (\hat{C}_1 \leq \hat{C}_2 \wedge \hat{P}_2 \wedge \hat{C}_1 \leq \hat{C}_3) \vee (\hat{C}_1 \leq \hat{C}_2 \wedge \neg \hat{P}_2 \wedge \hat{C}_1 \leq \hat{C}_4) \vee (\hat{C}_1 \geq \hat{C}_2 \wedge \hat{P}_2 \wedge \hat{C}_2 \leq \hat{C}_3) \vee (\hat{C}_1 \geq \hat{C}_2 \wedge \neg \hat{P}_2 \wedge \hat{C}_2 \leq \hat{C}_4)$.
 - If $C_m(i) = \min(C_3(i), C_4(i))$, let $\hat{P}_{k+1} = (\hat{C}_1 \leq \hat{C}_2 \wedge \hat{C}_3 \leq \hat{C}_4 \wedge \hat{C}_1 \leq \hat{C}_3) \vee (\hat{C}_1 \leq \hat{C}_2 \wedge \hat{C}_3 \geq \hat{C}_4 \wedge \hat{C}_1 \leq \hat{C}_4) \vee (\hat{C}_1 \geq \hat{C}_2 \wedge \hat{C}_3 \leq \hat{C}_4 \wedge \hat{C}_2 \leq \hat{C}_3) \vee (\hat{C}_1 \geq \hat{C}_2 \wedge \hat{C}_3 \geq \hat{C}_4 \wedge \hat{C}_2 \leq \hat{C}_4)$.
 - If $C_m(i) = \max(C_3(i), C_4(i))$, let $\hat{P}_{k+1} = (\hat{C}_1 \leq \hat{C}_2 \wedge \hat{C}_3 \geq \hat{C}_4 \wedge \hat{C}_1 \leq \hat{C}_3) \vee (\hat{C}_1 \leq \hat{C}_2 \wedge \hat{C}_3 \leq \hat{C}_4 \wedge \hat{C}_1 \leq \hat{C}_4) \vee (\hat{C}_1 \geq \hat{C}_2 \wedge \hat{C}_3 \geq \hat{C}_4 \wedge \hat{C}_2 \leq \hat{C}_3) \vee (\hat{C}_1 \geq \hat{C}_2 \wedge \hat{C}_3 \leq \hat{C}_4 \wedge \hat{C}_2 \leq \hat{C}_4)$.
 - If $C_m(i) = c$, for $c \in \mathbb{N}$ let $\hat{P}_{k+1} = (\hat{C}_1 \leq \hat{C}_2 \wedge \hat{C}_1 \leq c) \vee (\hat{C}_1 \geq \hat{C}_2 \wedge \hat{C}_2 \leq c)$.
 - If $C_\ell(i) = \max(C_1(i), C_2(i))$, this identical to the previous case but just switch $\hat{C}_1 \leq \hat{C}_2$ with $\hat{C}_1 \geq \hat{C}_2$.
 - If $C_\ell(i) = c$ for $c \in \mathbb{N}$, this should be straightforward given the above cases written out.

□

Lemma 6.1. $\mathbf{K}_t[\#]$ contains $\text{FOC}[+]$, and $\mathbf{K}_t[\#; \text{MOD}]$ contains $\text{FOC}[+; \text{MOD}]$. In general, for any set \mathcal{P} of unary predicates, the extension $\text{FOC}[+; \mathcal{P}]$ is contained in $\mathbf{K}_t[\#; \mathcal{P}]$, when defined as expected.

Proof. All $\text{FOC}[+]$ sentences can be rewritten in the following normal form (Chiang et al., 2023, Theorem 1):

$$\exists x_1 \dots \exists x_n. \left(\bigwedge_i \exists^{=x_i} p. \psi_i(p) \wedge \chi(x_1, \dots, x_n) \right)$$

where all ψ_i and χ are quantifier-free, and χ is a set of linear constraints on $x_1 \dots x_n$.

With respect to end-satisfiability, this is simply equivalent to the $\mathbf{K}_t[\#]$ formula.

$$\chi(\#[\psi_1], \dots, \#[\psi_n])$$

The same is true for $\text{FOC}[+; \text{MOD}]$, if we extend $\mathbf{K}_t[\#]$ with modular predicates to $\mathbf{K}_t[\#; \text{MOD}]$ where $(w, i) \models \text{MOD}_m^k \iff i \equiv k \pmod m$.

In general, from this normal form we can see that for any set \mathcal{P} of unary predicates, the extension $\text{FOC}[+; \mathcal{P}]$ is contained in $\mathbf{K}_t[\#; \mathcal{P}]$, when defined appropriately. \square

C Proofs: From $\mathbf{K}_t[\#]$ to Transformers

In all proofs, we omit the scaling factor applied by LayerNorm, because it can be confusing and also does not affect the end result. However, whenever relevant, we mention how to account for this scaling factor.

Here, we recall the definition of LayerNorm

Definition C.1. LayerNorm is a function $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that

$$f(x) = \frac{x - \mu}{\sigma} \quad \text{where} \quad \mu = \frac{1}{d} \sum_{i=1}^d x_i \quad \sigma = \sqrt{\frac{1}{d} \sum_{i=1}^d (x_i - \mu)^2}$$

Observe that if $\mu = 0$, LayerNorm only applies a scaling factor to all values in a position. Observe furthermore if the absolute value of all values in a position are equal, LayerNorm scales all of them to be ± 1 , which is important in [Appendix C.3](#).

An essential part of our construction is access to a Boolean vector that is True at the BOS position and False everywhere else. We call this a *Start-Separating Vector*, adapting terminology from [Merrill and Sabharwal \(2024\)](#). Using the word embedding for BOS, we can construct a dimension that holds such a vector (ensuring $\text{WE}(w_i)$ is true at dimensions $2k_s - 1, 2k_s$ only when $w_i = \text{BOS}$).

$$\begin{array}{cc} & \begin{matrix} 1 & 2 & & n' \end{matrix} \\ \begin{matrix} 2k_s - 1 \\ 2k_s \end{matrix} & \begin{bmatrix} \vdots & \vdots & & \vdots \\ -1 & +1 & \cdots & +1 \\ +1 & -1 & \cdots & -1 \\ \vdots & \vdots & & \vdots \end{bmatrix} \end{array}$$

C.1 BOS Handling Lemma

Lemma 5.3. Using the word embedding and a single feed-forward layer, we can set the BOS position to False, without disturbing the Boolean value at any other position.

Proof. Construct a feed-forward layer which computes the min of every value in an odd dimension with the value in dimension $2k_s - 1$ of the Start-Separating Vector and the max

of every value in an even dimension with the value in $2k_s$, as described in [Lemma 5.5](#). Since we've maintained that all values are ± 1 in our Boolean representation, this sets the BOS position to be False, while the others are unmodified. \square

C.2 Counting Lemma

Lemma 5.4. Let $A_{2k-1:2k,*}$ store a Boolean vector as defined above. For any i , let $C_{k,i}$ be the number of positions $j \leq i$ such that $A_{2k-1:2k,j}$ is True. Then there is a transformer block that computes, at each position i , and in two other dimensions $2k' - 1, 2k'$, the values $-\frac{C_{k,i}}{i+1}$ and $\frac{C_{k,i}}{i+1}$.

Proof. First, recall the definition of a masked self attention layer $SA: \mathbb{R}^{d \times n} \rightarrow \mathbb{R}^{d \times n}$:

$$\begin{aligned} SA(A) &= [c_0 \quad \cdots \quad c_n] \text{ where} \\ W^{(Q)}: \mathbb{R}^d &\rightarrow \mathbb{R}^{d_k} \\ W^{(K)}: \mathbb{R}^d &\rightarrow \mathbb{R}^{d_k} \\ W^{(V)}: \mathbb{R}^d &\rightarrow \mathbb{R}^d \\ s_{ij} &= \frac{W^{(Q)} A_{*,i} \cdot W^{(K)} A_{*,j}}{\sqrt{d}} \\ c_i &= \frac{\sum_{j=0}^i \exp(s_{ij}) W^{(V)} A_{*,j}}{\sum_{j=0}^i \exp(s_{ij})} \end{aligned}$$

To simulate a count term $\#[\phi] = C(i)$ of $\mathbf{K}_t[\#]$, we need to construct a transformer block such that if the Boolean values $\phi(i)$ are stored in some dimension $2k - 1, k$, we can compute $\frac{C(i)}{i+1}$ in some other dimensions $2k' - 1, k'$.

To achieve this, we only need *uniform attention* – that is, at each position i , we set $W^{(Q)} = W^{(K)} = \mathbf{0}$, which makes $s_{ij} = 0$ for all i, j . This spreads attention weight evenly across all positions $j \leq i$.

Furthermore, by setting $W^{(V)}$ as follows, we can add the value from position i in dimensions $2k - 1, 2k$ to position i in dimensions $2k' - 1, 2k'$.

$$W^{(V)} = \begin{matrix} & \begin{matrix} 2k-1 & 2k & & 2k'-1 & 2k' \end{matrix} \\ \begin{matrix} 2k-1 \\ 2k \\ 2k'-1 \\ 2k' \end{matrix} & \begin{bmatrix} \vdots & \vdots & & \vdots & \vdots \\ \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots \\ \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 1 & 1 & \cdots & 0 & 0 & \cdots \\ \cdots & 1 & 1 & \cdots & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix} \end{matrix}$$

As such, we get the attention layer to compute the average of all unmasked positions $j \leq i$. Recalling the definition once more, the general expression reads:

$$c_{i,k} = \frac{\sum_{j=0}^i \exp(s_{ij}) [W^{(V)} A_{*,j}]_k}{\sum_{j=0}^i \exp(s_{ij})}$$

Then, with the use of uniform attention and our carefully constructed $W^{(V)}$, we compute in position i of dimension k , the value $c_{i,k}$, which is the average of all values up to position i in dimension k . The expression reduces to:

$$c_{i,k} = \frac{\sum_{j=0}^i \exp(s_{ij}) [W^{(V)} A_{*,j}]_k}{\sum_{j=0}^i \exp(s_{ij})} = \frac{\sum_{j=0}^i [W^{(V)} A_{*,j}]_k}{\sum_{j=0}^i 1} = \frac{\sum_{j=0}^i A_{k,j}}{\sum_{j=0}^i 1} = \frac{\sum_{j=0}^i A_{k,j}}{i+1}$$

Before moving on, note that we represent Booleans as $-1, +1$ instead of $0, 1$, and we'll also write the sum of positions $j \leq i$ that hold a True as $C_{i,k}$. Here we write out the resulting tensor after the described self-attention layer showing the relevant dimensions at positions $0, 1, \dots, n'$. We write $B_i \in \{0, 1\}$ be the Boolean value at position i .

$$\begin{array}{c} \begin{matrix} 2k_0 - 1 \\ 2k_0 \\ \\ 2k - 1 \\ 2k \\ \\ 2k' - 1 \\ 2k' \\ \vdots \end{matrix} \begin{bmatrix} 1 & 2 & & n' \\ \vdots & \vdots & & \vdots \\ -1 & -1 & \cdots & -1 \\ +1 & +1 & \cdots & +1 \\ & & & \vdots \\ +1 & -2B_1 + 1 & & -2B_n + 1 \\ -1 & +2B_1 - 1 & & +2B_n - 1 \\ & & & \vdots \\ 0 & -2\frac{C_{1,d'}}{2} + 1 & \cdots & -2\frac{C_{n,d'}}{n+1} + 1 \\ 0 & +2\frac{C_{1,d'}}{2} - 1 & \cdots & +2\frac{C_{n,d'}}{n+1} - 1 \\ \vdots & \vdots & & \vdots \end{bmatrix} \end{array}$$

Note, however, that instead of the desired value $\frac{C_{i,k}}{i+1}$, we have actually computed $2\frac{C_{i,k}}{i+1} - 1$. We use a feed-forward layer to undo this transformation by subtracting (or adding) 1 and then dividing by 2 in dimension d (or $d+1$). We cannot achieve the ± 1 using the bias of a FFN, as a scaling factor may be applied to the tensor by LayerNorm. Thus, we store a constant $[-1, 1]$ in dimensions $2k_0 - 1, k_0$, which will always have the same scaling factor applied to it. Then, it is straightforward to construct a feed-forward layer that simply adds dimension $2k_0 - 1$ to $2k - 1$ and $2k_0$ to $2k$ in order to remove the \pm , and then divides by 2 to get the result:

$$\begin{array}{c} \begin{matrix} 2k' - 1 \\ 2k' \\ \vdots \end{matrix} \begin{bmatrix} 1 & 2 & & n' \\ \vdots & \vdots & & \vdots \\ 0 & -\frac{C_{1,d'}}{2} & \cdots & -\frac{C_{n,d'}}{n+1} \\ 0 & +\frac{C_{1,d'}}{2} & \cdots & +\frac{C_{n,d'}}{n+1} \\ \vdots & \vdots & & \vdots \end{bmatrix} \end{array}$$

As a final note, which is relevant for the later subsection on arithmetic comparison, we will make every self-attention layer compute a counting operation over the start-separating vector ([Appendix C.1](#)) in order to compute the value $\frac{1}{i+1}$ at every position i in some dimension. \square

C.3 Feed-Forward Lemma

Lemma 5.5. The following position-wise operations can be simulated by a single transformer block, using existing dimensions as input and a fresh dimension as output: addition (+), subtraction (−), comparison (\leq), and Boolean operations (\wedge , \neg). Arbitrary Boolean expressions in DNF can be simulated using two blocks.

Proof. A two-layer feed-forward network with ReLU activations on the first layer and none on the second layer can compute any continuous piecewise linear function with a finite number of pieces, or CPWL (Arora et al., 2018). All of these operations are CPWL, but we give explicit constructions for concreteness, and the case of comparisons requires special care.

We consider each type of operation in turn. All constructions only use the feed-forward layer. We can force the self-attention layer to perform no changes to the residual stream by setting all weights to 0; the residual connection then adds the original values back in.

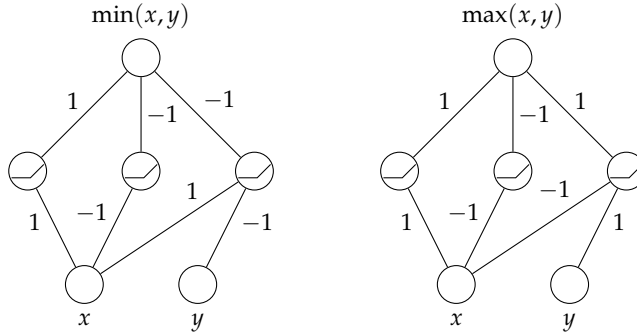
Addition and subtraction: These operations are straightforward: put 1’s in W_1 such that we add the appropriate values to the fresh dimension, and negate using W_2 if needed.

Min/Max: Recall that $\text{ReLU}(x) = \max(0, x)$. Then $\min(x, y) = x - \text{ReLU}(x - y)$.

- If $x < y$, then $\text{ReLU}(x - y) = 0$, so $x - \text{ReLU}(x - y) = x = \min(x, y)$.
- If $x \geq y$, then $\text{ReLU}(x - y) = x - y$, so $x - \text{ReLU}(x - y) = x - (x - y) = y = \min(x, y)$.

Similarly, $\max(x, y) = x + \text{ReLU}(y - x)$.

Therefore there exist FFNs to compute the min or max of two real numbers:



To compute the min of two count terms, we compute the min of their positive components and the max of their negative components. Similarly for the max of two count terms.

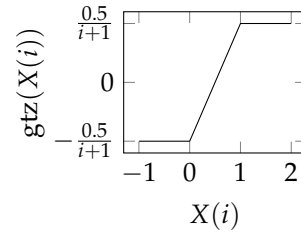
Comparison: This requires access to $\pm \frac{1}{i+1}$ in some dimensions $2k_0 - 1, 2k_0$. This is easily achieved by requiring every self-attention layer to perform a counting operation over the start-separating vector (Appendix C.1).

First we explain how to simulate a comparison of two count terms $C_1(i) \leq C_2(i)$. Then, we describe how to extend this to compare linear combinations of count terms.

Suppose that we want to compare C_1 and C_2 in dimensions $2k_1 - 1, 2k_1$ and $2k_2 - 1, 2k_2$, and put the result in dimension $2k_3 - 1, 2k_3$. Initially, the residual stream looks like this:

$$\begin{array}{c} i \\ \vdots \\ 2k_0 - 1 \\ 2k_0 \\ \vdots \\ 2k_1 - 1 \\ 2k_1 \\ \vdots \\ 2k_2 - 1 \\ 2k_2 \\ \vdots \\ 2k_3 - 1 \\ 2k_3 \\ \vdots \end{array} \begin{bmatrix} \vdots \\ \dots - \frac{1}{i+1} \dots \\ \dots + \frac{1}{i+1} \dots \\ \vdots \\ \dots - \frac{C_1(i)}{i+1} \dots \\ \dots + \frac{C_1(i)}{i+1} \dots \\ \vdots \\ \dots - \frac{C_2(i)}{i+1} \dots \\ \dots + \frac{C_2(i)}{i+1} \dots \\ \vdots \\ \dots 0 \dots \\ \dots 0 \dots \\ \vdots \end{bmatrix}$$

We construct a feed-forward layer that computes the function:

$$\text{gtz}(X(i)) = \min\left(\frac{0.5}{i+1}, \frac{X(i)}{i+1} - \frac{0.5}{i+1}\right) - \min\left(0, \frac{X(i)}{i+1}\right).$$


Observe that $\text{gtz}(C_2(i) - C_1(i) + 0.5)$ equals $\frac{0.5}{i+1}$ if $C_1(i) \leq C_2(i)$, and $-\frac{0.5}{i+1}$ otherwise. This is because the counts $C_1(i), C_2(i)$ must be integers, so if $C_1(i) \leq C_2(i)$, then $C_2(i) - C_1(i) + 0.5 \geq 0.5$, and the expression will evaluate to $\frac{0.5}{i+1}$. Otherwise, $C_2(i) - C_1(i) + 0.5 < 0.5$, and the expression will evaluate to $-\frac{0.5}{i+1}$.

It is straightforward, then, to use the construction for min/max from above to produce a feed-forward layer that computes $\text{gtz}(C_2(i) - C_1(i))$. Essentially, we use W_1 to compute the values (using the pre-existing values from the residual stream)

$$\frac{0.5}{i+1}, \frac{C_2(i) - C_1(i) + 0.5}{i+1}, -\frac{C_2(i) - C_1(i) + 0.5}{i+1}$$

Then we use W_2 to compute

$$\frac{0.5}{i+1} - \text{ReLU}\left(\frac{0.5}{i+1} - \frac{C_2(i) - C_1(i) + 0.5}{i+1}\right) - \text{ReLU}\left(-\frac{0.5}{i+1} - \frac{C_2(i) - C_1(i) + 0.5}{i+1}\right)$$

which equals $\text{gtz}(C_2(i) - C_1(i))$ as desired.

Similarly, it is straightforward to construct a feed-forward layer to compare linear combinations of count terms. That is, for disjoint sets of indices K_1 and K_2 , to compute

$$\text{gtz} \left(\sum_{k \in K_2} c_k \cdot C_k(i) - \sum_{k \in K_1} c_k \cdot C_k(i) \right).$$

So we can construct a feed-forward layer $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ that computes in each dimension i the following

$$f \left(\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{2k_3-1} \\ v_{2k_3} \\ \vdots \\ v_{d-1} \\ v_d \end{bmatrix} \right) = \begin{bmatrix} \text{gtz}(v_1) \\ \text{gtz}(v_2) \\ \vdots \\ \text{gtz} \left(\sum_{k \in K_2} c_k \cdot C_k(i) - \sum_{k \in K_1} c_k \cdot C_k(i) \right) \\ \text{gtz} \left(\sum_{k \in K_2} c_k \cdot C_k(i) - \sum_{k \in K_1} c_k \cdot C_k(i) \right) \\ \vdots \\ \text{gtz}(v_{d-1}) \\ \text{gtz}(v_d) \end{bmatrix}.$$

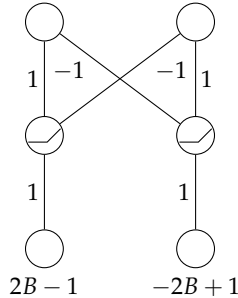
This truncates all positive values in the residual stream at this point to be $\frac{0.5}{i+1}$ at position i , and all nonpositive values to be $-\frac{0.5}{i+1}$. As a result, the next application of LayerNorm (with appropriate parameter settings) scales every single value to ± 1 . In particular, all previously-computed Boolean values are preserved, and the newly-computed dimensions $2k_3 - 1, 2k_3$ hold the correct Boolean value based on the desired comparison

As a side effect, all previously-computed counts also get changed to ± 1 , but we do not need these counts any longer.

Boolean operations: The Boolean operations \wedge and \neg can be computed by FFNNs with ReLU activations. Conjunction (\wedge) is equivalent to min/max:

$$\begin{bmatrix} \vdots \\ -2B_1 + 1 \\ 2B_1 - 1 \\ \vdots \end{bmatrix} \wedge \begin{bmatrix} \vdots \\ -2B_2 + 1 \\ 2B_2 - 1 \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ \max(-2B_1 + 1, -2B_2 + 1) \\ \min(2B_1 - 1, 2B_2 - 1) \\ \vdots \end{bmatrix}.$$

Logical negation (\neg) is equivalent to arithmetic negation, or swapping the positive and negative components:



For an arbitrary Boolean formula, convert it to *canonical* disjunctive normal form, which is a disjunction $\phi_1 \vee \dots \vee \phi_n$ of clauses, at most one of which can be true for any value of the inputs.

Each clause is of the form $\phi_m = \bigwedge_{k \in K_m} \psi_k$, where each ψ_k is an input or a negated input and K_m is a set of indices for each clause. A slightly different construction can be used to compute \wedge over inputs in K_m . Observe that:

$$\begin{aligned}\bigwedge_{k \in K_m} B_k &= \text{ReLU} \left(\left(\sum_{k \in K_m} (B_k) \right) - (|K_m| - 1) \right) \\ &= \text{ReLU} \left(\left(\sum_{k \in K_m} \frac{1}{2} (2B_k - 1) \right) - \frac{3|K_m|}{2} + 1 \right).\end{aligned}$$

And this can be computed for each clause using the first layer and ReLU of a feed-forward layer. Recall again that if the constant m is fixed, we can retrieve it by multiplying the constant ± 1 from dimensions $2k_0 - 1, 2k_0$ as described in [Appendix C.2](#). Then, because at most one clause can be true, the sum of all clauses will either be 1 or 0. Then, we convert back to the ± 1 representation of truth values.

$$\bigvee_{m=1}^n \left(\bigwedge_{k \in K_m} B_k \right) = 2 \cdot \left(\sum_{m=1}^n \text{ReLU} \left(\left(\sum_{k \in K_m} \frac{1}{2} (2B_k - 1) \right) - \frac{3|K_m|}{2} + 1 \right) \right) - 1.$$

This can all be done in a single feed-forward layer. □

C.4 Parallel Composition of Transformers

Lemma C.2. *Recall by the inductive hypothesis: Assume for formulas ψ of modal depth m we can construct a transformer which simulates ψ , consisting of m sequences of 4 blocks each:*

1. One to ensure the BOS position is false in every Boolean vector
2. One to simulate $\#$ terms over existing formulas
3. One to simulate comparisons of linear combinations of $\#$ terms
4. One to compute Boolean combinations of existing formulas.

Additionally, each comparison block only references the $\#$ terms computed in its immediately preceding block, and no others. We want to show that if T_1 with $4b_1$ blocks in d_1 dimensions simulates ϕ_1 and T_2 with $4b_2$ blocks in d_2 dimensions simulates ϕ_2 , both satisfying the inductive hypothesis, we can construct T_3 with $\max(4b_1, 4b_2)$ blocks in $d_1 + d_2$ dimensions which simulates both ϕ_1 and ϕ_2 , while also satisfying the inductive hypothesis.

Proof. Here we detail the claim in [Theorem 5.7](#) that we can compose many transformers (which satisfy the inductive hypothesis), into a larger transformer which simulates all the formulas the smaller transformers do, in parallel.

Let $1 \leq k \leq \min(b_1, b_2)$. We provide a high-level description of how to compose two sequences of 4 blocks $(B_{1,b})_{b \in [4k-3, 4k]}$ and $(B_{2,b})_{b \in [4k-3, 4k]}$ from T_1 and T_2 to create a sequence of blocks $(B_{3,b})_{b \in [4k-3, 4k]}$ for T_3 .

- Block $B_{3,4k-3}$ will ensure the BOS position is False at every Boolean vector
- Block $B_{3,4k-2}$ will set $W^{(V)}$ in its self-attention layer carefully so as to compute the same counting terms as $B_{1,4k-2}$ does in its first d_1 dimensions. It does this similarly with the weights from $B_{2,4k-2}$, in its next d_2 dimensions.
- Block $B_{3,4k-1}$ will copy over weights from $B_{1,4k-1}$ and $B_{2,4k-1}$ to compute the relevant comparisons in the right dimensions. Observe that the LayerNorm scaling trick, as described in [Appendix C.3](#), still applies to this layer
- Block $B_{3,4k}$ will compute Boolean combinations of the relevant dimensions.

Notice that since $B_{1,4k-1}$ and $B_{2,4k-1}$ only reference $B_{1,4k-2}$ and $B_{2,4k-2}$, so too does $B_{3,4k-1}$ only reference $B_{3,4k-2}$. This procedure can be used to compose every level of 4 blocks from T_1 and T_2 .

For $\min(b_1, b_2) \leq k \leq \max(b_1, b_2)$, the construction is straightforward. If T_1 has more blocks than T_2 , after the $4b_1$ -th block we can zero out the weights that output into the first

d_1 dimensions of T_3 , so that the rest of T_2 's blocks can be simulated while the part of the residual stream corresponding to T_1 remains unchanged.

As such, T_3 will satisfy the inductive hypothesis, with $\max(b_1, b_2)$ blocks in $d_1 + d_2$ dimensions. This procedure can be iterated to compose any finite collection of transformers T_k satisfying the inductive hypothesis. \square

D Fixed Precision Masked Transformers to $\mathbf{K}_t[\#; \text{MOD}]$

Definition D.1. A fixed-precision number with r integer bits and s fractional bits is a number in $\mathbb{F}_{r,s} = \{i/2^s \mid -2^{r+s} \leq i < 2^{r+s}\}$. For any value $a \in \mathbb{F}_{r,s}$, we write $\langle a \rangle_b$ for the b -th bit of the two's complement representation of a . That is,

$$\langle a \rangle_b = \lfloor a \cdot 2^{-b} \rfloor - \lfloor a \cdot 2^{-b-1} \rfloor \cdot 2.$$

This is a two's complement representation.

It helps to access each individual bit of x .

Proposition D.2. We write x^b for the b -th bit of x , whenever x is a fixed-precision number. Then, observe that we can write a formula $F_m(x) \iff x = F_m$ iff we can write formulas $F^b(x) \iff x^b = 1$.

Proof. This should be clear by an example: 1001 in $\mathbb{F}_{5,0}$. We write $F_m(x) \iff F^0(x) \wedge \neg F^1(x) \wedge \neg F^2(x) \wedge F^3(x)$. \square

Let us first define what it means for a formula of $\mathbf{K}_t[\#; \text{MOD}]$ to simulate a Fixed Precision transformer.

Definition D.3. Let $T : \Sigma^n \rightarrow \mathbb{F}_{r,s}^{d \times n'}$ be a fixed-precision masked soft attention transformer defined exactly the same except we use $\mathbb{F}_{r,s}$ instead of \mathbb{R} . We say T can be simulated in $\mathbf{K}_t[\#; \text{MOD}]$ if for every $F_m \in \mathbb{F}_{r,s}$ and every dimension k of T we can write a formula Φ_m^k such that

$$[T(\text{BOS} \cdot w)]_{k,i+1} = F_m \iff w, i \models \Phi_m^k$$

Similarly, defining predicates $\beta_m^k(i)$ for the BOS position. Essentially this means that we can write a formula that tells us what value the transformer must output, given any input.

Theorem 7.1. $\mathbf{K}_t[\#]$ can simulate fixed-precision masked soft attention transformers without positional encodings, and $\mathbf{K}_t[\#; \text{MOD}]$ can simulate fixed-precision masked soft attention transformers with sinusoidal positional encodings.

Proposition D.4. Assume we have Boolean functions $F_m^k(i)$ which return true iff the value at position i in dimension k is $F_m \in \mathbb{F}_{r,s}$. Then any function of the form $f_1(x_1, \dots, x_d) = f_2(x_1, \dots, x_d)$, where the x_k is the value at dimension k in position i , can be written as a Boolean combination of $F_m^k(i)$.

Proof. Essentially, this is $\mathbb{F}^n \times \mathbb{F}^n \rightarrow \{0, 1\}$, which only takes on finitely many values. Thus it is tedious, but straightforward, to enumerate all tuples of x which should return true, and write a formula that returns the correct answer. Let

$$\mathcal{K} = \{(m_1, \dots, m_d) \mid f_1(F_{m_1}, \dots, F_{m_d}) = f_2(F_{m_1}, \dots, F_{m_d})\}$$

That is, \mathcal{K} stores all n -tuples K of indices such that given an n -tuple of fixed-precision numbers which each have the corresponding index in K , the equality holds. Then write

$$\bigvee_{K \in \mathcal{K}} \left(\bigwedge_{k \leq d} F_{K_k}^k(i) \right)$$

This formula will return 1 iff x_k in dimensions k at position i have the correct fixed-precision values in order to satisfy the equality. \square

As a result, for any function $f: \mathbb{F}_{r,s}^d \rightarrow \mathbb{F}_{r,s}^d$ we can write formulas $\phi(i)$ to check which fixed-precision value the output of the function is at position i , given the formulas that check the values of the inputs at position i . This means

Lemma D.5. *We can write formulas $\text{FFN}_m^k(i)$, for each feed-forward layer FFN to check whether the output at position i in dimension k is $F_m \in \mathbb{F}_{r,s}$. Same for LayerNorm $\text{LN}_m^k(i)$.*

Proof. This is a direct consequence of [Proposition D.4](#) because these functions all take a finite number of fixed-precision inputs, and have a finite number of outputs. \square

The same is not the case for self-attention layers, as we have no bound on the input length. However, count terms help us out here.

Lemma D.6. *Assume we are using $\mathbb{F}_{r,s}$ as our fixed-precision representation. Assume we have access to predicates $F^b(i)$ which tell us if the value at position i has a 1 in the b -th bit of its fixed-precision representation. Then we can compute the following summation as a counting term*

$$2^s \cdot \sum_{j \leq i} X_j$$

where X_j is a value at position j .

Proof. Observe this summation is equivalent to

$$\begin{aligned} \sum_{j \leq i} X_j &= 2^0 \cdot \sum_{j \leq i} F^0(j) + 2^1 \cdot \sum_{j \leq i} F^1(j) \dots + 2^{r+s-1} \cdot \sum_{j \leq i} F^{r+s-1}(j) - 2^{r+s} \cdot \sum_{j \leq i} F^{r+s}(j) \\ &= \# [F^0] + 2\# [F^1] + \dots + 2^{r+s-1}\# [F^{r+s-1}] - 2^{r+s}\# [F^{r+s}]. \end{aligned} \quad \square$$

As a clarifying note, recall we are using two's complement, which is why the most significant bit is subtracted.

Lemma D.7. *We can write formulas $C_m^k(i)$ which are true iff the output of a self-attention layer at dimension k in position i is $F_m \in \mathbb{F}_{r,s}$*

Proof. Recall the definition of a self-attention layer in a fixed-precision masked soft attention transformer, $SA: \mathbb{F}_{r,s}^{d \times n} \rightarrow \mathbb{F}_{r,s}^{d \times n}$:

$$\begin{aligned} SA(A) &= [c_0 \quad \dots \quad c_n] \text{ where} \\ W^{(Q)}: \mathbb{F}_{r,s}^d &\rightarrow \mathbb{F}_{r,s}^{d_k} \\ W^{(K)}: \mathbb{F}_{r,s}^d &\rightarrow \mathbb{F}_{r,s}^{d_k} \\ W^{(V)}: \mathbb{F}_{r,s}^d &\rightarrow \mathbb{F}_{r,s}^d \\ s_{ij} &= \frac{W^{(Q)} A_{*,i} \cdot W^{(K)} A_{*,j}}{\sqrt{d}} \\ c_i &= \frac{\sum_{j=0}^i \exp(s_{ij}) W^{(V)} A_{*,j}}{\sum_{j=0}^i \exp(s_{ij})} \end{aligned}$$

More specifically, we want to define a formula $c_m^k(k)$ such that

$$c_m^k(i) \iff F_m \leq \frac{\sum_{j \leq i} e^{Q_i \cdot K_j} V_j^k}{\sum_{j \leq i} e^{Q_i \cdot K_j}} < F_m + 2^{-s}$$

where V_j^k is the k -th component of the $W^{(V)} A_{*,j}$. The bounds are because division in $\mathbb{F}_{r,s}$ must perform some sort of rounding to the nearest number. We rearrange the equation to the following:

$$\phi_m^k(i) \iff \sum_{j \leq i} (F_m \cdot e^{Q_i \cdot K_j}) \leq \sum_{j \leq i} (e^{Q_i \cdot K_j} V_j^k) < \sum_{j \leq i} ((F_m + 2^{-s}) \cdot e^{Q_i \cdot K_j}).$$

We only write out how to evaluate the left inequality to save space, but the right inequality works exactly the same. Now observe that we can enumerate all the (finitely many) values that the \mathbb{F}^d vector Q_i as $(Q_x)_{x \leq pd}$ in order to get an expression that only depends on j :

$$c_m^k(i) \iff \begin{cases} \sum_{j \leq i} (F_m \cdot e^{Q_1 \cdot K_j}) \leq \sum_{j \leq i} (e^{Q_1 \cdot K_j} V_j^k) & Q_i = Q_1 \\ \sum_{j \leq i} (F_m \cdot e^{Q_2 \cdot K_j}) \leq \sum_{j \leq i} (e^{Q_2 \cdot K_j} V_j^k) & Q_i = Q_2 \\ \vdots & \vdots \\ \sum_{j \leq i} (F_m \cdot e^{Q_{pd} \cdot K_j}) \leq \sum_{j \leq i} (e^{Q_{pd} \cdot K_j} V_j^k) & Q_i = Q_{pd} \end{cases}$$

This can straightforwardly be translated into a Boolean combination of formulas $c_{m,Q_x(i)}^k$, etc., which by [Proposition D.4](#) are definable. Furthermore, by the same proposition it is also straightforward to construct a formula $QKV(j)$ that specifies the value of $e^{Q_x \cdot K_j} V_j^k$ at position j , as well as a formula $F_m QK(j)$ that specifies the value of $F_m \cdot e^{Q_{pd} \cdot K_j}$. This allows us to write the above expression in a form such that [Lemma D.6](#) can be applied directly to both sides of the equation, thus computing both as count terms and allowing their comparison within $\mathbf{K}_t[\#; \text{MOD}]$. Notice that by applying [Lemma D.6](#) we are implicitly scaling both sides up by 2^s , but this is fine, as equality would still be preserved under this scaling factor. □

Finally, to complete the simulation it remains to show that we can write formulas that simulate the word embedding and positional encoding, which will be similar to the construction of [Chiang et al. \(2023\)](#).

Lemma D.8. *We can write $\mathbf{K}_t[\#]$ formulas $WE_m^k(i)$ that check that the k -th dimension of the embedding at position i has the value $F_m \in \mathbb{F}_{r,s}$*

Proof. For $a \in \Sigma \cup \{\text{BOS}\}$, let $WE(a) = [F_{m_1,a}, \dots, F_{m_d,a}]$ denote the fixed-precision word embedding for a . Then, using the same notation as earlier, we can write

$$WE_m^k(i) = \bigvee_{a \in \{a \mid WE(a)^k = F_m\}} Q_a(i).$$

Similarly, we define $\beta_m^k(i)$, which checks the word embedding for BOS has value F_m at dimension k . Sinusoidal positional encodings can be described in the same manner, using modular predicates. □

Finally,

Proof. As a direct consequence of the above lemmas, we can write a formula of $\mathbf{K}_t[\#]$, which simulates a fixed-precision masked soft attention transformer without positional encodings. Adding modular predicates allows $\mathbf{K}_t[\#; \text{MOD}]$ to simulate fixed-precision masked soft attention transformer with sinusoidal positional encodings. \square