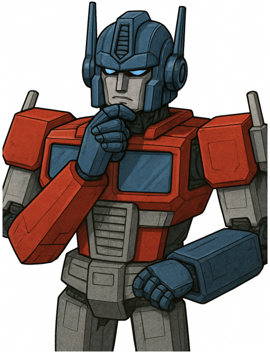
Marco Sälzer[1] Chris Köcher[2] Alexander Kozachinskiy[3]
Georg Zetzsche[2] Anthony Widjaja Lin[1,2]

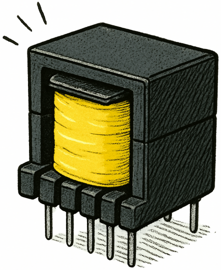
[1] Rheinland-Pfälzische Technische Universität, Kaiserslautern & Landau, Germany
[2] Max Planck Institute for Software Systems, Kaiserslautern, Germany
[3] Centro Nacional de Inteligencia Artificial, Santiago, Chile

October 2, 2025

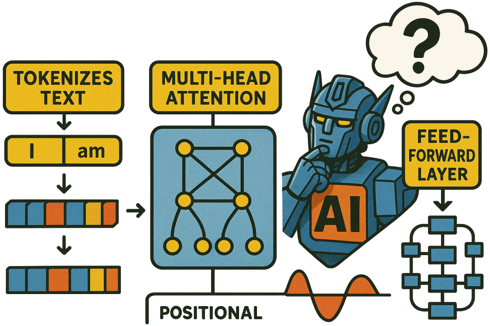
# Transformers?



✗



✗




✓

- Transformers are the basic model in machine learning used in recent LLMs

ChatGPT Gemini Meta AI

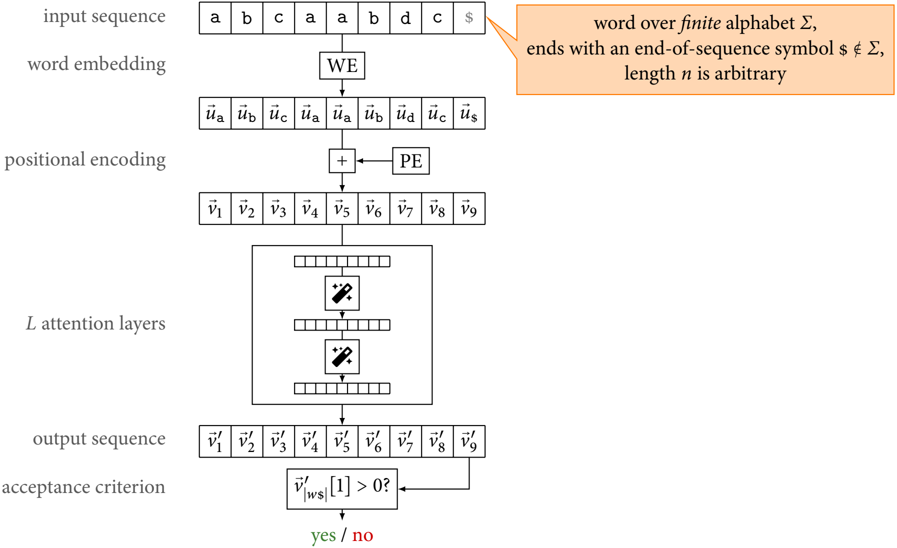
# Motivation

- Artificial intelligence is often not intelligent:

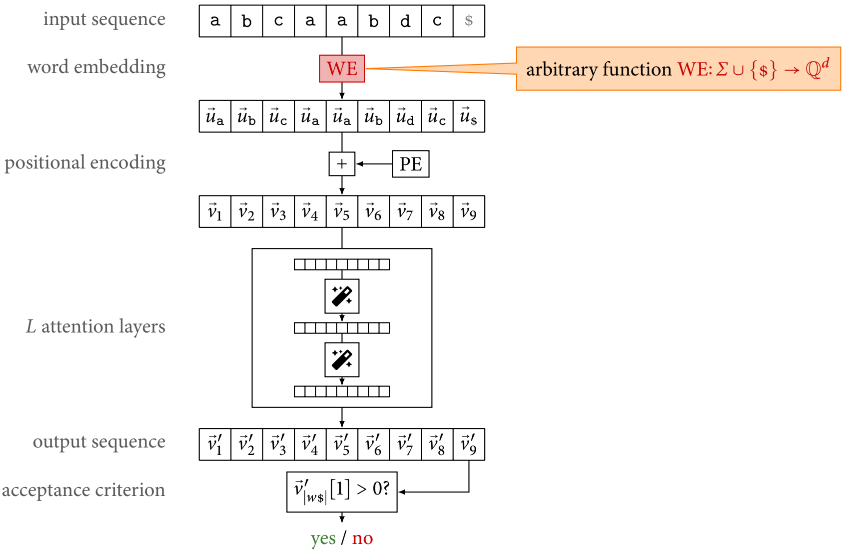

Source: Facebook, "15 most powerful passports in the world".

- We try to verify transformers.
  - A first step: study expressibility

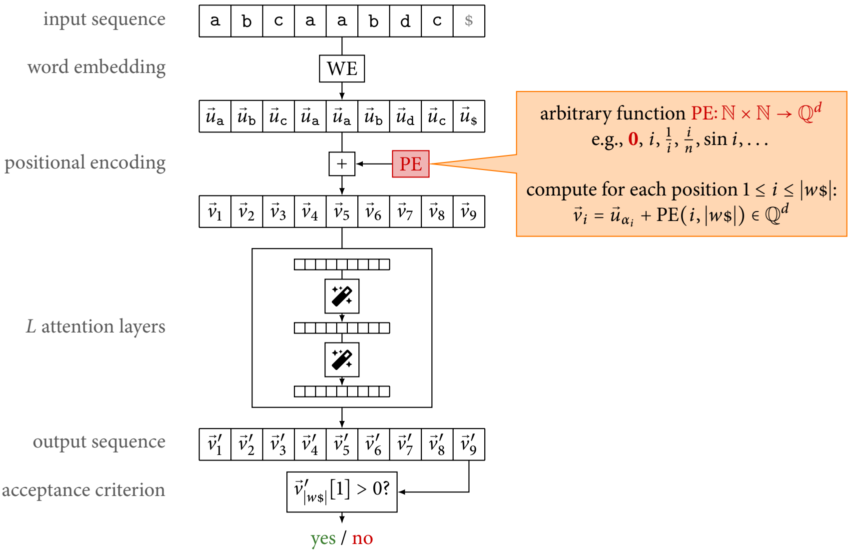
# Transformers: Scheme



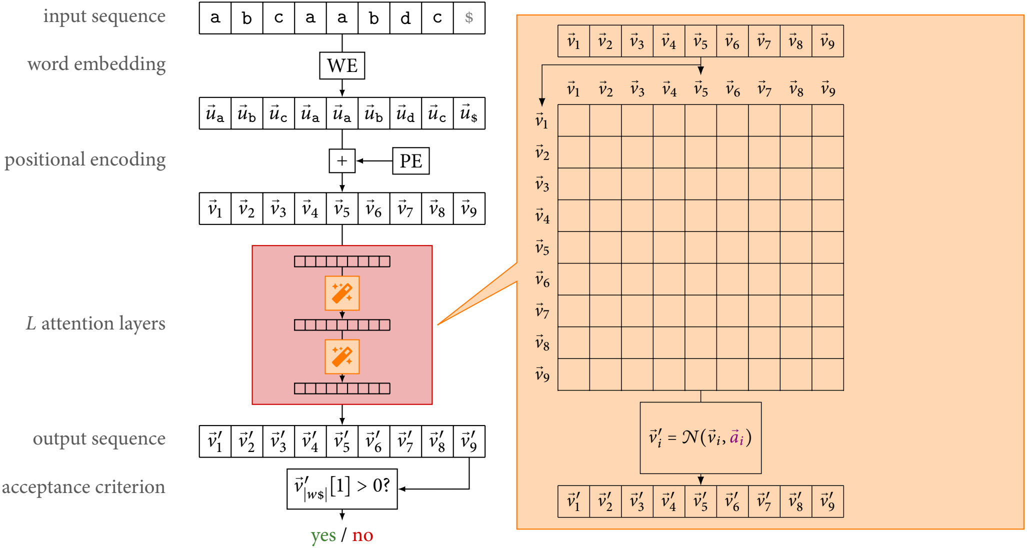
# Transformers: Scheme



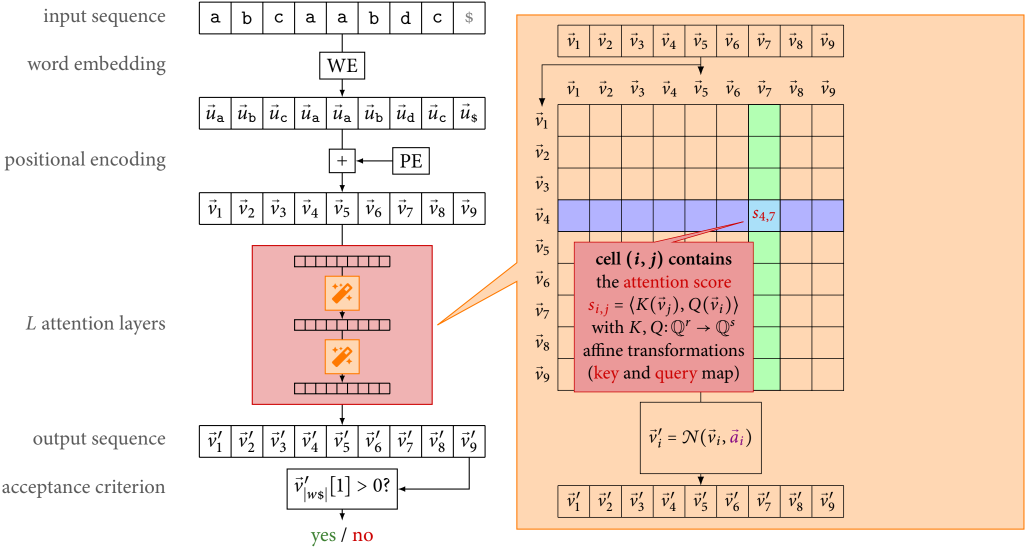
# Transformers: Scheme



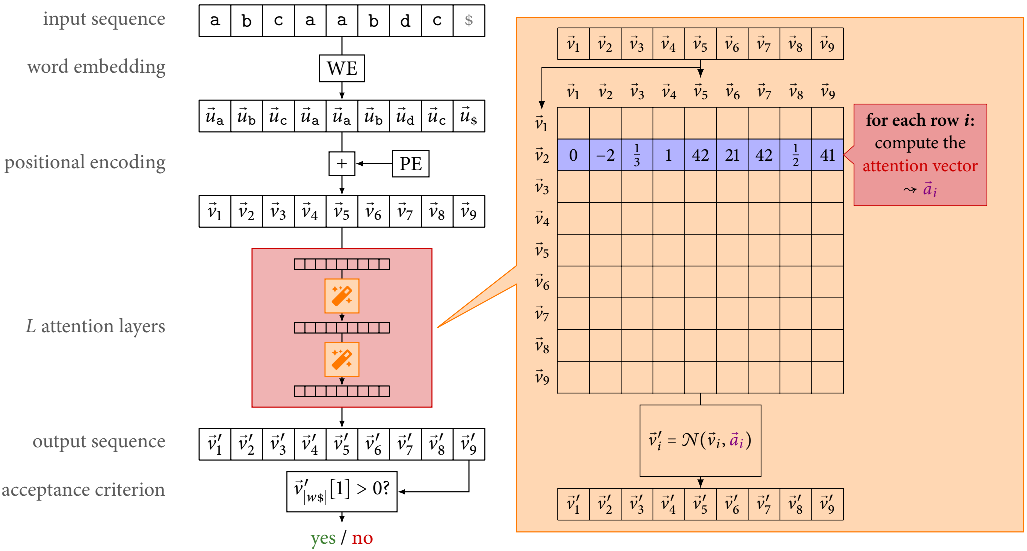
# Transformers: Scheme



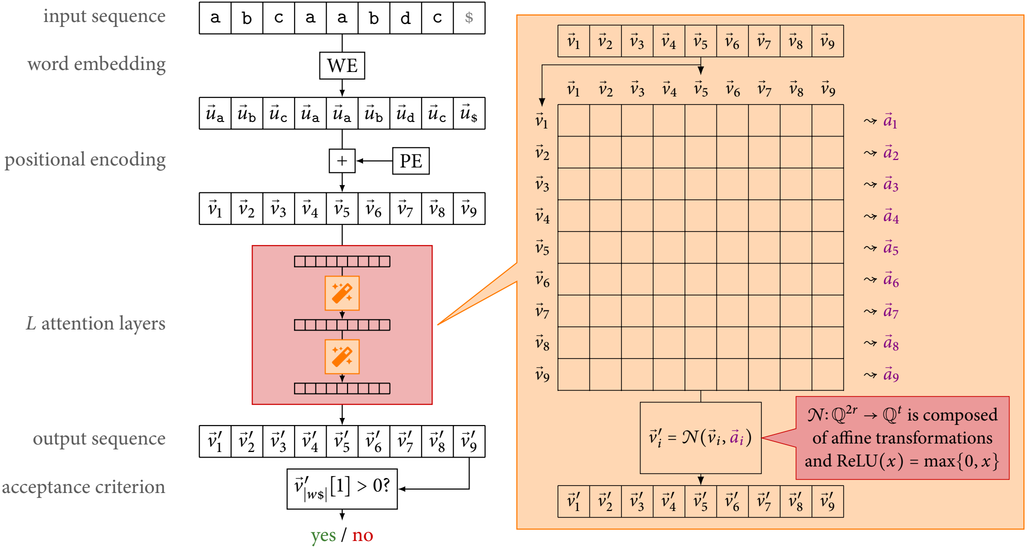
# Transformers: Scheme



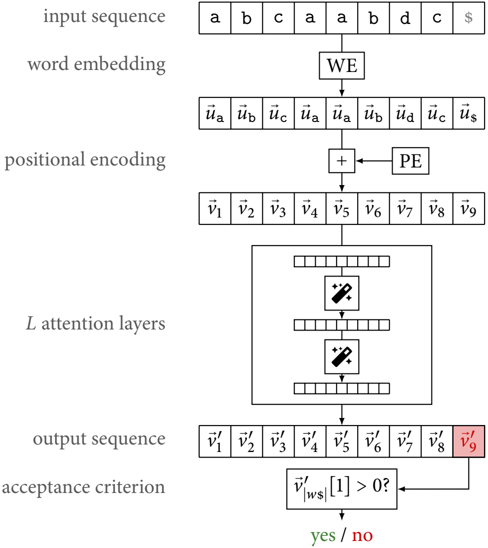
# Transformers: Scheme



# Transformers: Scheme



# Transformers: Scheme



# Transformers: Attention Mechanisms

## 1 Unique Hard Attention (UHA)

| | $\vec{v}_1$ | $\vec{v}_2$ | $\vec{v}_3$ | $\vec{v}_4$ | $\vec{v}_5$ | $\vec{v}_6$ | $\vec{v}_7$ | $\vec{v}_8$ | $\vec{v}_9$ |
|---|---|---|---|---|---|---|---|---|---|
| $\vec{v}_i$ | 0 | $-2$ | $\frac{1}{3}$ | 1 | 42 | 21 | 42 | $\frac{1}{2}$ | 41 |

$$\rightsquigarrow \vec{a}_i = \vec{v}_5$$

## 2 Average Hard Attention (AHA)

| | $\vec{v}_1$ | $\vec{v}_2$ | $\vec{v}_3$ | $\vec{v}_4$ | $\vec{v}_5$ | $\vec{v}_6$ | $\vec{v}_7$ | $\vec{v}_8$ | $\vec{v}_9$ |
|---|---|---|---|---|---|---|---|---|---|
| $\vec{v}_i$ | 0 | $-2$ | $\frac{1}{3}$ | 1 | 42 | 21 | 42 | $\frac{1}{2}$ | 41 |

$$\rightsquigarrow \vec{a}_i = \frac{1}{2}\vec{v}_5 + \frac{1}{2}\vec{v}_7$$

## 3 Softmax Attention (SMA)

| | $\vec{v}_1$ | $\vec{v}_2$ | $\vec{v}_3$ | $\vec{v}_4$ | $\vec{v}_5$ | $\vec{v}_6$ | $\vec{v}_7$ | $\vec{v}_8$ | $\vec{v}_9$ |
|---|---|---|---|---|---|---|---|---|---|
| $\vec{v}_i$ | 0 | $-2$ | $\frac{1}{3}$ | 1 | 42 | 21 | 42 | $\frac{1}{2}$ | 41 |

$$\rightsquigarrow \vec{a}_i = \sum_{j=1}^{|w\$|} \frac{e^{s_{i,j}}}{\sum_{k=1}^{|w\$|} e^{s_{i,k}}} \cdot \vec{v}_j$$

# Transformers: Languages

- accepted language of a transformer $\mathfrak{T}$: $L(\mathfrak{T}) = \{w \in \Sigma^* \mid \mathfrak{T}(w\$) = \text{yes}\}$.
- UHAT / AHAT / SMAT: all languages accepted by a transformer with UHA / AHA / SMA mechanism

- NoPE-C: all languages accepted by a C-transformer without positional encoding (i.e., $\mathrm{PE}(i, n) = \vec{0}$)
- C[U]: all languages accepted by a C-transformer such that each layer is uniform (i.e., if the key and query maps $K$ and $Q$ are constant).
- C[≤L]: all languages accepted by a C-transformer with at most $L$ attention layers.

**Theorem (Yang, Chiang, Angluin @ NeurIPS 2024)**

NoPE-UHAT[Masking] = Star-Free

# Overview

| | | | | |
|---|---|---|---|---|
| NoPE-AHAT[$\leq$1] = QFPA | $\subsetneq$ | NoPE-AHAT = NoPE-AHAT[U] = SemiAlg | $\subsetneq$ | AHAT |
| | | | $\subsetneq$ | SMAT |

## Definition

Let $\Sigma = \{\mathtt{a}_1, \mathtt{a}_2, \ldots, \mathtt{a}_d\}$ be an alphabet. The Parikh map is defined as

$$\Psi \colon \Sigma^* \to \mathbb{N}^d \colon w \mapsto \left(|w|_{\mathtt{a}_1}, |w|_{\mathtt{a}_2}, \ldots, |w|_{\mathtt{a}_d}\right)$$

where $|w|_{\mathtt{a}_i}$ is the number of occurrences of $\mathtt{a}_i$ in the word $w$.

NoF

QFPA

=

SemiAlg

$\subsetneq$

SMAT

$L \subseteq \Sigma^*$ is semi-algebraic (in SemiAlg) if there are polynomials $p_1, p_2, \ldots, p_k \in \mathbb{Z}[\vec{x}]$ such that $L = L_{p_1} \cup \cdots \cup L_{p_k}$ where $L_{p_i} = \{w \in \Sigma^* \mid p_i(\Psi(w)) > 0\}$.

$L \subseteq \Sigma^*$ is semilinear (in QFPA) if it is semi-algebraic such that all polynomials are linear (i.e., have degree $\leq 1$).

$\Updownarrow$

$L$ is described by a Presburger formula.

## Definition

Let $\Sigma = \{a_1, a_2,$

$$\Psi : \Sigma^* \to \mathbb{N}^d : w \mapsto (|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_d})$$

where $|w|_{a_i}$ is the number of occurrences of $a_i$ in the word $w$.

# Overview

NoPE-A trivial ]

=

QFPA

? $\subsetneq$ ?

NoPE-AHAT

=

NoPE-AHAT[U]

=

SemiAlg

$\subsetneq$ AHAT

?

$\subsetneq$ SMAT

# Overview

$\mathsf{NoPE\text{-}AHAT}[\leq 1] = \mathsf{QFPA}$

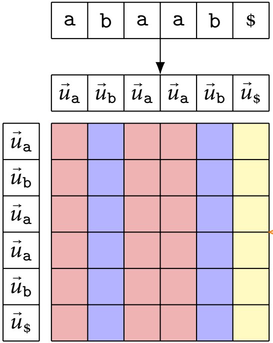
$\subsetneq$

$\mathsf{NoPE\text{-}AHAT} = \mathsf{NoPE\text{-}AHAT}[U] = \mathsf{SemiAlg}$

$\subsetneq$ $\mathsf{AHAT}$

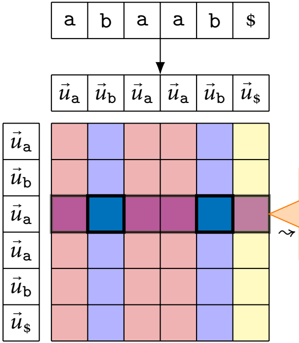
$\subsetneq$ $\mathsf{SMAT}$

# NoPE-AHAT ⊆ SemiAlg (1)



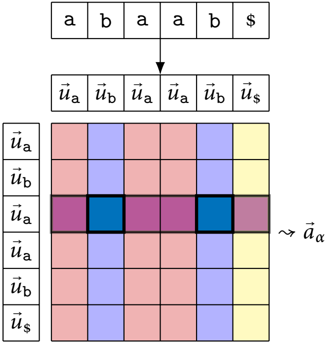

All a-columns (resp. a-rows) are the same.

# NoPE-AHAT $\subseteq$ SemiAlg (1)




If one b-position has maximal score, then all b-positions attend.

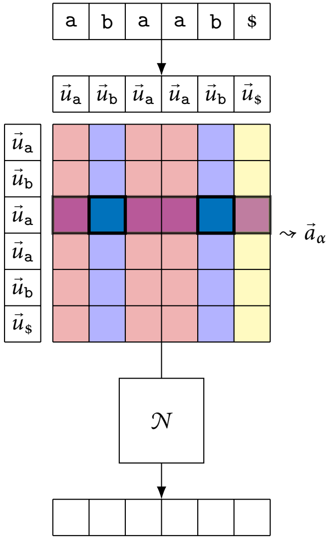
# NoPE-AHAT $\subseteq$ SemiAlg (1)




$\Rightarrow$ attention corresponds to choosing a $\Gamma \subseteq \Sigma \cup \{\$\}$

$\Rightarrow$ $O(2^{|\Sigma|})$ many cases for each attention vector

# NoPE-AHAT ⊆ SemiAlg (1)




⇒ attention corresponds to choosing a $\Gamma \subseteq \Sigma \cup \{\$\}$

⇒ $O(2^{|\Sigma|})$ many cases for each attention vector

⇒ two cases per letter and ReLU,
$O(2^{|\Sigma|k})$ many cases if $\mathcal{N}$ contains $k$ ReLUs

⇒ $O(2^{|\Sigma|^2 k})$ many different outcomes

⇒ $O(2^{|\Sigma|^2 kL})$ many different outcomes after $L$ layers

# NoPE-AHAT ⊆ SemiAlg (2)

- For each of the $O(2^{|\Sigma|^2 kL})$ many choices construct a conjunction of polynomial inequalities that verify:
  - i the choices maximize the attention scores
  - ii the resulting vector $\vec{v}_\$^{(L)}$ satisfies the accepting condition
- Choice of maximal score can be done via equations of the form

$$\langle K(\vec{v}_\alpha), Q(\vec{v}_\beta)\rangle \geq \langle K(\vec{v}_\gamma), Q(\vec{v}_\beta)\rangle$$

- Attention vectors can be expressed via

$$\vec{a}_\alpha[i] = \frac{\sum_{\beta\in\Gamma} x_\beta \cdot \vec{v}_\beta[i]}{\sum_{\beta\in\Gamma} x_\beta}$$

- Finally, take a disjunction over all $O(2^{|\Sigma|^2 kL})$ conjunctions. □

# NoPE-AHAT ⊆ SemiAlg (2)

- For each of the $O(2^{|\Sigma|^2 kL})$ many choices construct a conjunction of polynomial inequalities that verify:
  - i the choices maximize the attention scores
  - ii the resulting vector $\vec{v}_\$^{(L)}$ satisfies the accepting condition
- Choice of maximal score can be done via equations of the form

$$\langle K(\vec{v}_\alpha), Q(\vec{v}_\beta)\rangle \geq \langle K(\vec{v}_\gamma), Q(\vec{v}_\beta)\rangle$$

$K$ are $Q$ are affine transformations
$\Rightarrow$ polynomials

- Attention vecto

$$\vec{a}_\alpha[i] = \frac{\cdots [i]}{\sum_{\beta \in \Gamma} x_\beta}$$

- Finally, take a disjunction over all $O(2^{|\Sigma|^2 kL})$ conjunctions. □

# NoPE-AHAT $\subseteq$ SemiAlg (2)

- For each of the $O(2^{|\Sigma|^2 kL})$ many choices construct a conjunction of polynomial inequalities that verify:
  - i the choices maximize the attention scores
  - ii the resulting vector $\vec{v}_{\$}^{(L)}$ satisfies the accepting condition
- Choice of maximal score can be done via equations of the form

$$\langle K(\vec{v}$$

Equivalently:
$\vec{v}'_\alpha[i] \cdot \sum_{\beta \in \Gamma} x_\beta = \sum_{\beta \in \Gamma} x_\beta \cdot \vec{v}_\beta[i]$

- Attention vectors can be expres

$$\vec{a}_\alpha[i] = \frac{\sum_{\beta \in \Gamma} x_\beta \cdot \vec{v}_\beta[i]}{\sum_{\beta \in \Gamma} x_\beta}$$

- Finally, take a disjunction over all $O(2^{|\Sigma|^2 kL})$ conjunctions. $\square$

# Overview

| NoPE-AHAT[$\leq$1] = QFPA | $\subsetneq$ | NoPE-AHAT = NoPE-AHAT[U] = SemiAlg | $\subsetneq$ | AHAT |
| --- | --- | --- | --- | --- |
| | ? | | $\subsetneq$ | SMAT |

# Monomials into AHAT

## Lemma

*Let $p \in \mathbb{Z}[\vec{x}]$ be a monomial of degree $L$. Then*

$$L_p = \{w \in \Sigma^* \mid p(\Psi(w)) > 0\} \in \mathsf{NoPE\text{-}AHAT}[\leq \mathsf{L}, \mathsf{U}].$$

0. Word embedding: map each letter $\alpha \in \Sigma \cup \{\$\}$ to unit vector $\vec{e}_\alpha \in \mathbb{N}^{\Sigma \cup \{\$\}}$
1. Compute frequencies: compute $|\Sigma| + 1$ many new components holding values $\frac{x_\alpha}{|w\$|}$
   - One uniform attention layer $\Rightarrow$ attention vector $\vec{a}_i = \sum_{\alpha \in \Sigma \cup \{\$\}} \frac{|w\$|_\alpha}{|w\$|} \cdot \vec{e}_\alpha = \left(\frac{x_\alpha}{|w\$|}\right)_{\alpha \in \Sigma \cup \{\$\}}$
   - Append the attention vector
2. Multiply: compute $\frac{x_\alpha}{|w\$|} \cdot y_i$ for some $\alpha \in \Sigma \cup \{\$\}$ and a component $y_i$
   - Let $u_\alpha \in \{0,1\}$ be the value in component $\alpha$.
   - Compute $u_\alpha \cdot y_i$ via a neural network:
   $$u_\alpha \cdot y_i = \mathrm{ReLU}(y_i - (1 - u_\alpha)) \in \{0, y_i\}$$
   - Accumulate the $u_\alpha \cdot y_i$ to $\frac{x_\alpha \cdot y_i}{|w\$|}$ in a uniform attention layer.
3. Iterate: compute $\frac{p(\vec{x})}{|w\$|^L}$ by iterated multiplication. □

# Monomials into AHAT

**Lemma**

*Let $p \in \mathbb{Z}[\vec{x}]$ be a monomial of degree L. Then*

$$L_p = \{w \in \Sigma^* \mid p(\Psi(w)) > 0\} \in \text{NoPE-AHAT}[\leq \mathsf{L}, \mathsf{U}].$$

0. Word embedding: map each letter $\alpha \in \Sigma \cup \{\$\}$ to unit vector $\vec{e}_\alpha \in \mathbb{N}^{\Sigma \cup \{\$\}}$
1. Compute frequencies: compute $|\Sigma| + 1$ many new components holding values $\frac{x_\alpha}{|w\$|}$
   - One uniform attention layer $\Rightarrow$ attention vector $\vec{a}_i = \sum_{\alpha \in \Sigma \cup \{\$\}} \frac{|w\$|_\alpha}{|w\$|} \cdot \vec{e}_\alpha = \left(\frac{x_\alpha}{|w\$|}\right)_{\alpha \in \Sigma \cup \{\$\}}$
   - Append the attention vector
2. Multiply: compute $\frac{x_\alpha}{|w\$|} \cdot y_i$ for some $\alpha \in \Sigma \cup \{\$\}$ and a component $y_i$
   - Let $u_\alpha \in \{0,1\}$ be the value in component $\alpha$.
   - Compute $u_\alpha \cdot y_i$ via a neural network:

$$u_\alpha \cdot y_i = \text{ReLU}(y_i - (1 - u_\alpha)) \in \{0, y_i\}$$

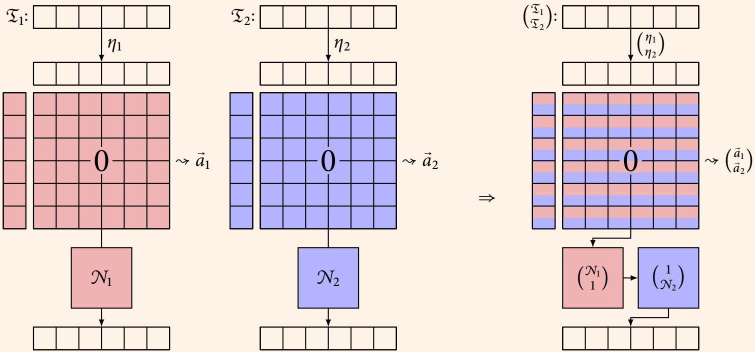
   - Accumulate the … to $\frac{x_\alpha \cdot y_i}{|w\$|}$ in … uniform attention layer.

$\in \{0,1\}$ $\in [0,1]$

3. Iterate: compute $\frac{p(\vec{x}}{|w\$|^L}$ … ted … cation. □

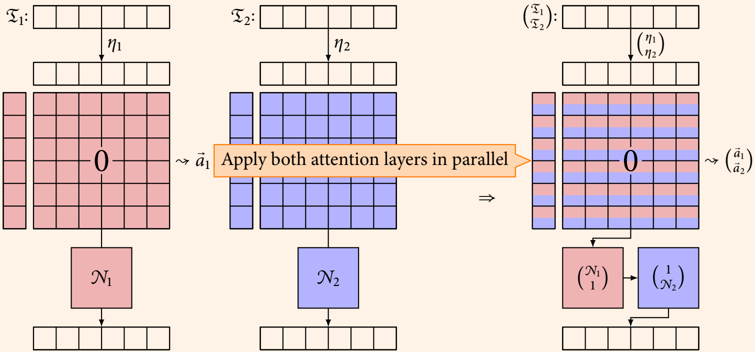
# Parallelize Uniform NoPE-AHATs

## Observation



# Parallelize Uniform NoPE-AHATs

## Observation



# Polynomials into AHAT

**Lemma**

*Let $p \in \mathbb{Z}[\vec{x}]$ be a polynomial of degree $L$. Then*

$$L_p = \{w \in \Sigma^* \mid p(\Psi(w)) > 0\} \in \mathsf{NoPE\text{-}AHAT}[\leq \mathsf{L}, \mathsf{U}].$$

- We may assume that $p$ is homogeneous, i.e. all monomials of $p$ have degree $L$:

$$p'(x_0, \vec{x}) := x_0^L \cdot p\left(\frac{x_1}{x_0}, \ldots, \frac{x_m}{x_0}\right)$$

  - $p(\vec{x}) > 0 \iff p'(1, \vec{x}) > 0$
  - We will represent $x_0$ via the end marker \$.
- Construct AHATs for all monomials, parallelize them, and compute the sum in the neural network in the last layer. □

# Semi-Algebraic Sets into AHAT

**Proposition**

$\mathsf{SemiAlg} \subseteq \mathsf{NoPE\text{-}AHAT}[\mathsf{U}]$

- Let $L \in \mathsf{SemiAlg}$ and $p_1, \ldots, p_k \in \mathbb{Z}[\vec{x}]$ be polynomials with $L = L_{p_1} \cup \cdots \cup L_{p_k}$.
- Compute AHATs for all $L_{p_i}$, parallelize them, and compute the maximum of their outputs in the neural network in the last layer.
  - Note: $\max\{x, y\} = \max\{0, x - y\} + y = \mathrm{ReLU}(x - y) + y$ $\square$

# Overview

$$\text{NoPE-AHAT}[\leq 1] = \text{QFPA} \;\subsetneq\; \text{NoPE-AHAT} = \text{NoPE-AHAT}[\text{U}] = \text{SemiAlg}$$

$$\text{NoPE-AHAT} = \text{NoPE-AHAT}[\text{U}] = \text{SemiAlg} \;\subsetneq\; \text{AHAT}$$

$$\text{NoPE-AHAT} = \text{NoPE-AHAT}[\text{U}] = \text{SemiAlg} \;\subsetneq\; \text{SMAT}$$

# Overview

$$\text{NoPE-AHAT}[\leq 1] = \text{QFPA} \subsetneq \text{NoPE-AHAT} = \text{NoPE-AHAT}[\text{U}] = \text{SemiAlg}$$

$$\text{NoPE-AHAT} = \text{NoPE-AHAT}[\text{U}] = \text{SemiAlg} \subsetneq \text{AHAT}$$

$$\text{NoPE-AHAT} = \text{NoPE-AHAT}[\text{U}] = \text{SemiAlg} \subsetneq \text{SMAT}$$

# Related Results

**Corollary**

1. $\textsc{Parity} = \{w \in \{\mathsf{a}, \mathsf{b}\}^* : |w|_{\mathsf{a}} \textit{ is even}\} \notin \mathsf{NoPE\text{-}AHAT}$
2. $\mathsf{Proj}(\mathsf{NoPE\text{-}AHAT}) = \mathsf{RE} \cap \mathsf{PI}$

**Theorem**

$\mathsf{Proj}(\mathsf{NoPE\text{-}AHAT}[\leq 2, \mathsf{U}]) = \mathsf{RE} \cap \mathsf{PI}$

**Corollary**

$\mathsf{NoPE\text{-}AHAT}[\leq 2, \mathsf{U}]$ *can express languages not recognized by higher-order recursion schemes (HORS), Petri nets, simplified multi-counter machines, or* LTL *with counting.*

# Related Results

## Corollary

1. Parity $= \{w \in \{a,b\}^* \mid |w|_a \text{ is even}\} \notin$ NoPE-AHAT
2. Proj(NoPE-AHAT) = RE ∩ PI

recursively enumerable languages

permutation-invariant languages

languages $\pi(L)$ for $L \in$ NoPE-AHAT and projections $\pi\colon \Sigma \to \Gamma \cup \{\varepsilon\}$

## Corollary

NoPE-AHAT[≤2, U] *can express languages not recognized by higher-order recursion schemes (HORS), Petri nets, simplified multi-counter machines, or* LTL *with counting.*

# Overview

NoPE-AHAT[$\leq$1] = QFPA $\subsetneq$ NoPE-AHAT = NoPE-AHAT[U] = SemiAlg

NoPE-AHAT = NoPE-AHAT[U] = SemiAlg $\subsetneq$ AHAT

NoPE-AHAT = NoPE-AHAT[U] = SemiAlg $\subsetneq$ SMAT

Needs a closer look on
the proof of NoPE-AHAT = SemiAlg

# Overview

$$\text{NoPE-AHAT}[\leq 1] = \text{QFPA} \subsetneq \text{NoPE-AHAT} = \text{NoPE-AHAT}[U] = \text{SemiAlg}$$

$$\text{NoPE-AHAT} \subsetneq \text{AHAT}$$

$$\text{NoPE-AHAT} \subsetneq \text{SMAT}$$

# Overview

NoPE-AHAT[≤1]

=

QFPA

?

$\subsetneq$

NoPE-AHAT

=

NoPE-AHAT[U]

=

$\subsetneq$ AHAT

$\subsetneq$ SMAT

$$\textsc{Sqrt} = \left\{ w \in \{\mathsf{a}, \mathsf{b}\}^* \;\middle|\; |w|_{\mathsf{a}} < \frac{|w|}{\sqrt{2}} \right\}$$

$\in$ SemiAlg $\setminus$ QFPA

# Overview

NoPE-AHAT[$\leq$1] = QFPA

$\subsetneq$

NoPE-AHAT = NoPE-AHAT[U] = SemiAlg

$\subsetneq$ AHAT

$\subsetneq$ SMAT

# Emptiness Problem

## Corollary

*The emptiness problem is*

1. *decidable for* NoPE-AHAT$[\leq 1]$.
2. *undecidable for* NoPE-AHAT$[\leq 2]$.

## Theorem

*The following two problems are inter-reducible:*

1. *The emptiness problem for AHAT without positional encoding and without end marker*
2. *The solvability problem of Diophantine equations over rationals*

# Emptiness Problem

Since Presburger arithmetic is decidable
[Presburger 1929]

**Co**

*The emptiness problem*

1. *decidable for* NoPE-AHAT[$\leq 1$].
2. *undecidable for* NoPE-AHAT[$\leq 2$].

Since Proj(NoPE-AHAT[$\leq 2$]) = RE $\cap$ PI

**Theo**

*The following two problems are inter-reducible:*

1. *The emptiness problem for AHAT without positional encoding and without end marker*
2. *The solvability problem of Diophantine equations over rationals*

(Un-)decidability of this problem is still open!

# Overview

NoPE-AHAT$[\leq 1]$

$=$

QFPA

$\subsetneq$

NoPE-AHAT

$=$

SemiAlg

?

$\subsetneq$

AHAT

$\textsc{Parity} \in \text{AHAT} \setminus \text{NoPE-AHAT}$

[Barceló, Kozachinskiy, Lin, Podolskii 2024]

# Overview

NoPE-AHAT[$\leq$1]
=
QFPA

$\subsetneq$

NoPE-AHAT
=
NoPE-AHAT[U]
=
SemiAlg

$\subsetneq$ AHAT

$\subsetneq$ SMAT

# Overview

$\mathsf{NoPE\text{-}AHAT}[\leq 1] = \mathsf{QFPA}$

$\subsetneq$

$\mathsf{NoPE\text{-}AHAT} = \mathsf{NoPE\text{-}AHAT}[\mathsf{U}] = \mathsf{SemiAlg}$

$\subsetneq$ $\mathsf{AHAT}$

$\subsetneq$ $\mathsf{SMAT}$

"$\subseteq$": Uniform AHA = uniform SMA

"$\neq$": Parity $\in \mathsf{SMAT} \setminus \mathsf{NoPE\text{-}AHAT}$

# Overview

$\text{NoPE-AHAT}[\leq 1] = \text{QFPA} \subsetneq \text{NoPE-AHAT} = \text{NoPE-AHAT}[\text{U}] = \text{SemiAlg}$

$\text{NoPE-AHAT} = \text{NoPE-AHAT}[\text{U}] = \text{SemiAlg} \subsetneq \text{AHAT}$

$\text{NoPE-AHAT} = \text{NoPE-AHAT}[\text{U}] = \text{SemiAlg} \subsetneq \text{SMAT}$

- Open Problems:
  - Do $\text{AHAT}[\text{U}] = \text{AHAT}$ or $\text{NoPE-SMAT}[\text{U}] = \text{NoPE-SMAT}$ hold?
  - Is $\text{NoPE-AHAT}[\leq \text{L}, \text{U}]$ equal to semi-algebraic sets with polynomials of degree $\leq L$?
  - Are intersection, separability, … decidable?

Thank you!

arXiv: 2505.11199