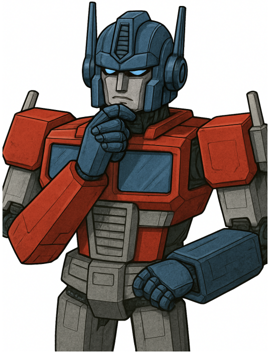
Chris Köcher[1] Alexander Kozachinskiy[2] Anthony Widjaja Lin[1,3]
Marco Sälzer[3] Georg Zetzsche[1]

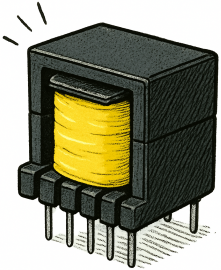
[1] Max Planck Institute for Software Systems, Kaiserslautern, Germany
[2] Centro Nacional de Inteligencia Artificial, Santiago, Chile
[3] Rheinland-Pfälzische Technische Universität, Kaiserslautern & Landau, Germany

September 4, 2025

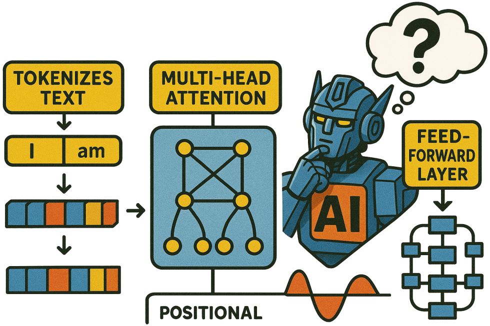
Transformers?



✗



✗



✓

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

ChatGPT Gemini Meta AI

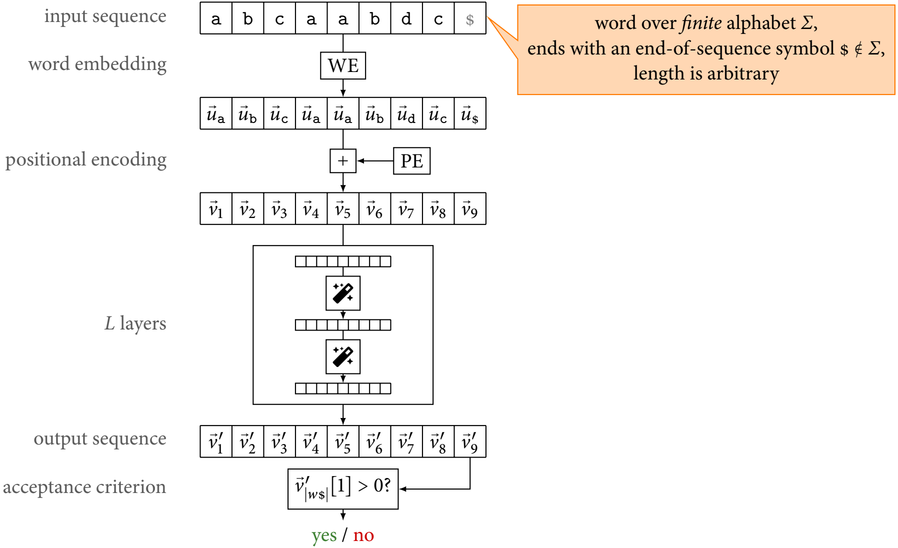
Motivation

- Artificial intelligence is often not intelligent:

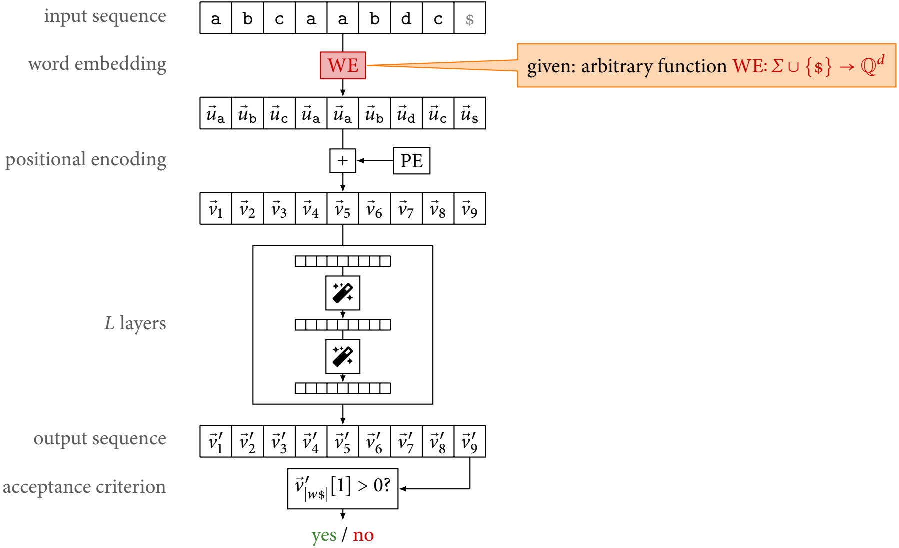
Source: Facebook, "15 most powerful passports in the world".

- We try to understand what transformers actually can do.
  - A first step: study expressibility

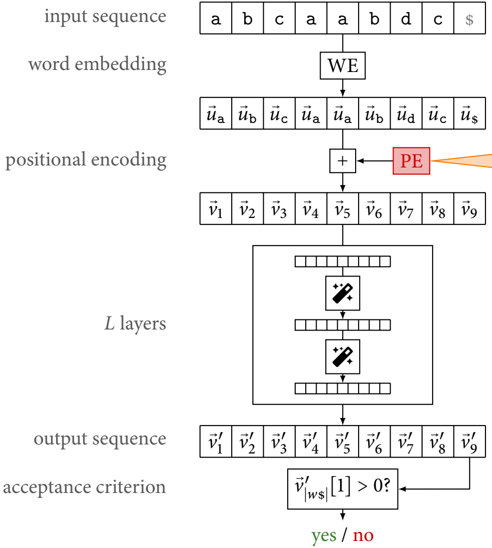
Transformers: Scheme



Transformers: Scheme



Transformers: Scheme



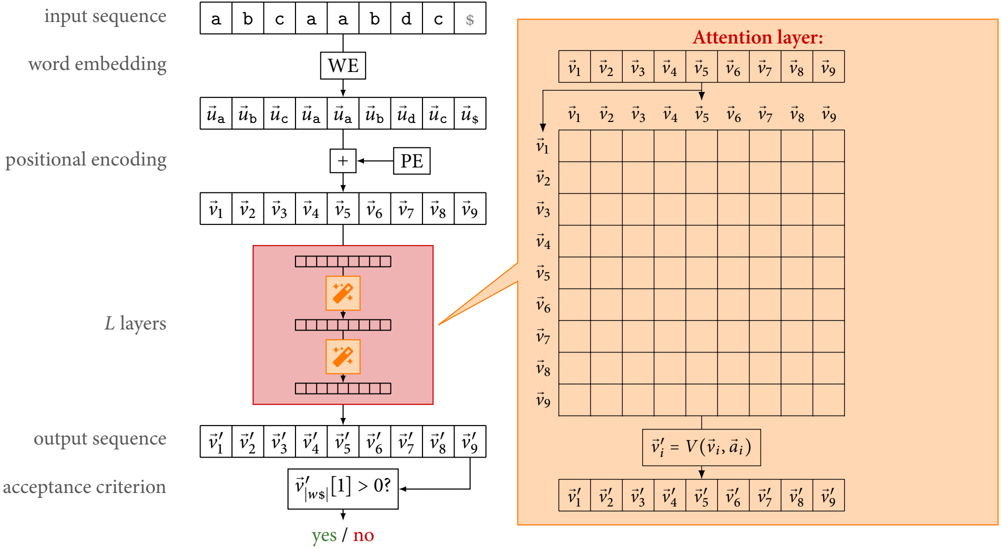
given: arbitrary function $\mathrm{PE}\colon \mathbb{N} \times \mathbb{N} \to \mathbb{Q}^d$

e.g., $\mathbf{0}$, i, $\frac{1}{i}$, $\frac{i}{|w|}$, $\sin i, \ldots$

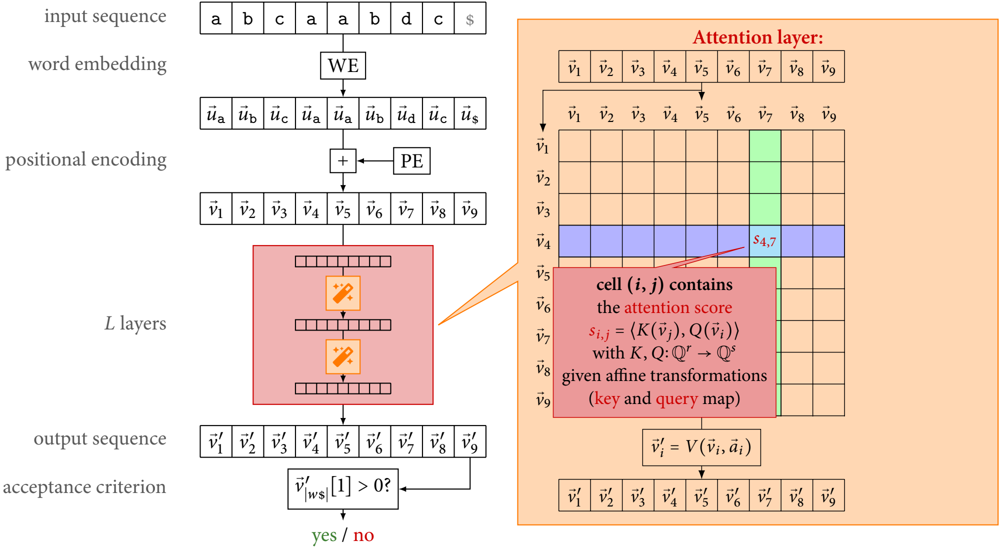
compute for each position $1 \le i \le |w\$|$:

$\vec{v}_i = \vec{u}_{\alpha_i} + \mathrm{PE}(i, |w\$|) \in \mathbb{Q}^d$

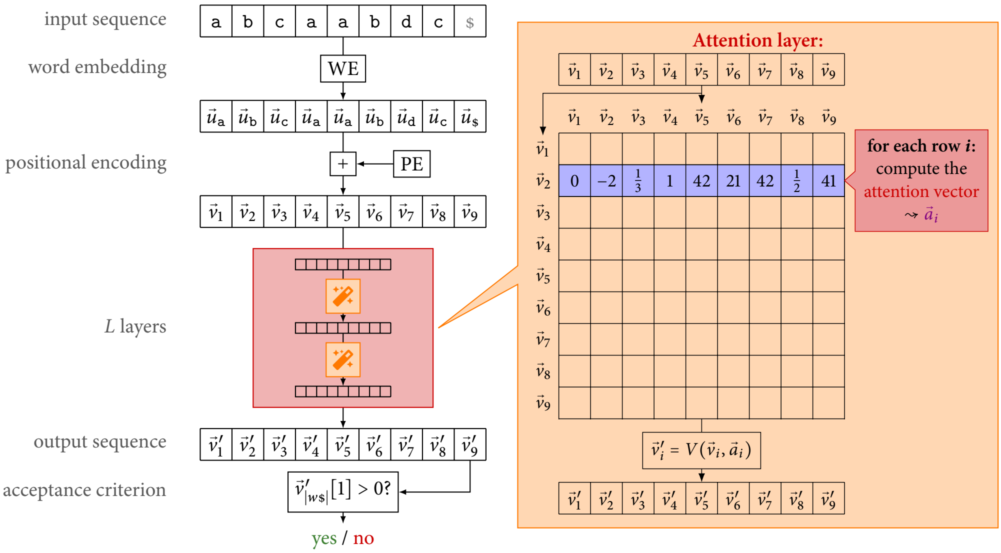
Transformers: Scheme



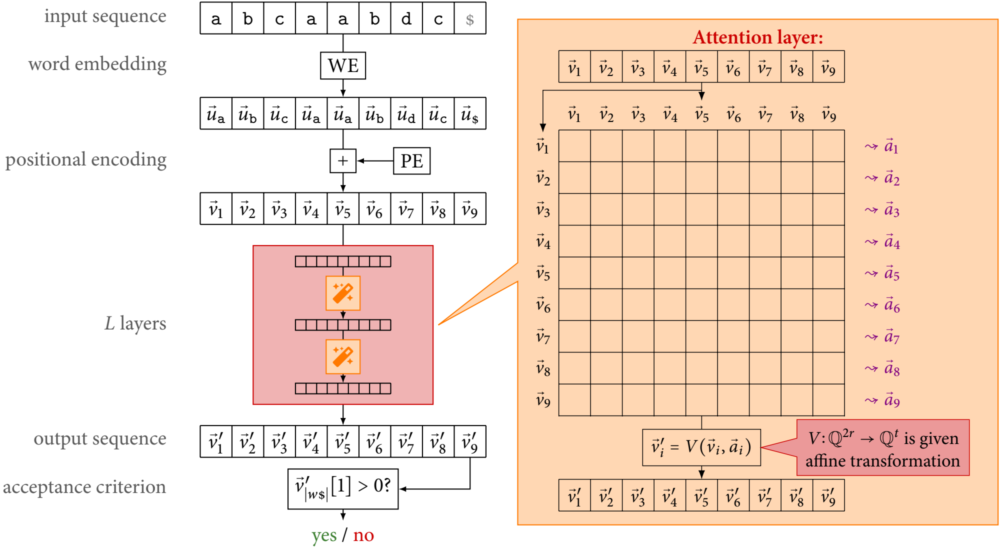
Transformers: Scheme



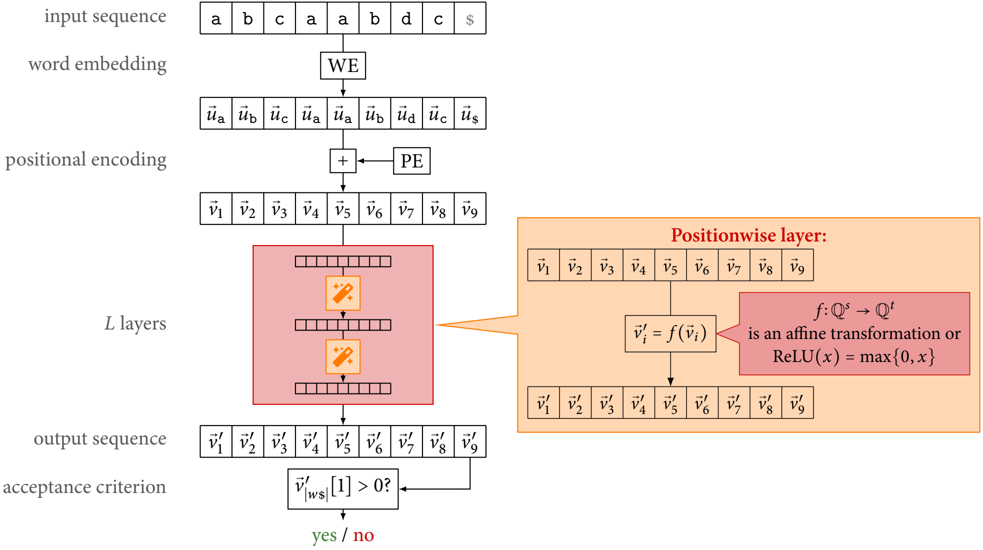
Transformers: Scheme



Transformers: Scheme



Transformers: Scheme



1 Unique Hard Attention Transformers (UHAT)

$$\begin{array}{c|c|c|c|c|c|c|c|c|c|}
 & \vec{v}_1 & \vec{v}_2 & \vec{v}_3 & \vec{v}_4 & \vec{v}_5 & \vec{v}_6 & \vec{v}_7 & \vec{v}_8 & \vec{v}_9 \\
\hline
\vec{v}_i & 0 & -2 & \frac{1}{3} & 1 & 42 & 21 & 42 & \frac{1}{2} & 41 \\
\hline
\end{array} \rightsquigarrow \vec{a}_i = \vec{v}_5$$

2 Average Hard Attention Transformers (AHAT)

$$\begin{array}{c|c|c|c|c|c|c|c|c|c|}
 & \vec{v}_1 & \vec{v}_2 & \vec{v}_3 & \vec{v}_4 & \vec{v}_5 & \vec{v}_6 & \vec{v}_7 & \vec{v}_8 & \vec{v}_9 \\
\hline
\vec{v}_i & 0 & -2 & \frac{1}{3} & 1 & 42 & 21 & 42 & \frac{1}{2} & 41 \\
\hline
\end{array} \rightsquigarrow \vec{a}_i = \frac{1}{2}\vec{v}_5 + \frac{1}{2}\vec{v}_7$$

3 Softmax Attention Transformers with temperature scaling $\tau > 0$ (τ-SMAT)

$$\begin{array}{c|c|c|c|c|c|c|c|c|c|}
 & \vec{v}_1 & \vec{v}_2 & \vec{v}_3 & \vec{v}_4 & \vec{v}_5 & \vec{v}_6 & \vec{v}_7 & \vec{v}_8 & \vec{v}_9 \\
\hline
\vec{v}_i & 0 & -2 & \frac{1}{3} & 1 & 42 & 21 & 42 & \frac{1}{2} & 41 \\
\hline
\end{array} \rightsquigarrow \vec{a}_i = \sum_{j=1}^{|w\$|} \frac{e^{s_{i,j}/\tau}}{\sum_{k=1}^{|w\$|} e^{s_{i,k}/\tau}} \cdot \vec{v}_j$$

Transformers: Languages

- accepted language of a transformer $\mathfrak{T}$: $L(\mathfrak{T}) = \{w \in \Sigma^* \mid \mathfrak{T}(w\$) = \text{yes}\}$.
- AHAT / τ-SMAT / SMAT: all languages accepted by an AHAT / τ-SMAT / 1-SMAT

- 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.

Results

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

NoPE-AHAT = NoPE-AHAT[U] = SemiAlg $\subsetneq$ AHAT $\subsetneq$ TC^0

AHAT $\subseteq$ τ-SMAT

NoPE-AHAT = NoPE-AHAT[U] = SemiAlg $\subseteq$ SMAT $\subseteq$ τ-SMAT

Definition

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

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

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

Results

TC^0

N

QFPA

SemiAlg $\subseteq$ SMAT $\subseteq$ τ-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 ≤ 1).

$\Updownarrow$

L is described by a Presburger formula.

defined as

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

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

Results

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

$$\mathsf{NoPE\text{-}AHAT} \subsetneq \mathsf{AHAT} \subsetneq \mathsf{TC}^0 \qquad \mathsf{AHAT} \subseteq \tau\text{-}\mathsf{SMAT}$$

$$\mathsf{NoPE\text{-}AHAT} \subseteq \mathsf{SMAT} \subseteq \tau\text{-}\mathsf{SMAT}$$

Definition

Let $\Sigma = \{\mathsf{a}_1, \mathsf{a}_2, \ldots$ ... ned as

... $|_{\mathsf{a}_d})$

where $|w|_{\mathsf{a}_i}$ is the r...

Corollary

$\mathsf{Proj}(\mathsf{NoPE\text{-}AHAT}) = \mathsf{RE} \cap \mathsf{PI}$

Corollary

The emptiness problem is undecidable for (NoPE-)AHAT with two (uniform) layers.

Results

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

$\mathsf{NoPE\text{-}AHAT} \subsetneq \mathsf{AHAT} \subsetneq \mathsf{TC}^0$

$\mathsf{NoPE\text{-}AHAT} \subseteq \mathsf{SMAT} \subseteq \tau\text{-}\mathsf{SMAT}$

$\mathsf{AHAT} \subseteq \tau\text{-}\mathsf{SMAT}$

Definition

Let $\Sigma = \{\mathsf{a}_1, \mathsf{a}_2, \ldots$

$|_{\mathsf{a}_d})$

wh

Corol

recursively enumerable languages

$\mathsf{Proj}(\mathsf{NoPE\text{-}AHAT}) = \mathsf{RE} \cap \mathsf{PI}$

permutation-invariant languages

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

is undecidable for (NoPE-)AHAT with two (uniform) layers.

$\mathsf{NoPE\text{-}AHAT}[\leq 1] = \mathsf{QFPA} \subsetneq \mathsf{NoPE\text{-}AHAT} = \mathsf{NoPE\text{-}AHAT}[\mathsf{U}] = \mathsf{SemiAlg} \subsetneq \mathsf{AHAT} \subsetneq \mathsf{TC}^0$

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

$\mathsf{AHAT} \subseteq \tau\text{-}\mathsf{SMAT}$

Definition

Let $\Sigma = \{\mathsf{a}_1, \mathsf{a}_2, \ldots$ ... ned as

... $|_{\mathsf{a}_d})$

where $|w|_{\mathsf{a}_i}$ is the n...

Corollary

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

Corollary

The emptiness problem is undecidable for (NoPE-)AHAT with two (uniform) layers.

Results

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

$\mathsf{NoPE\text{-}AHAT}[\mathsf{U}] \subsetneq \mathsf{AHAT} \subsetneq \mathsf{TC}^0$

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

$\mathsf{AHAT} \subseteq \tau\text{-}\mathsf{SMAT}$

- Open Problems:
  - Are $\mathsf{NoPE\text{-}AHAT} \subseteq \mathsf{SMAT}$ and $\mathsf{AHAT}, \mathsf{SMAT} \subseteq \tau\text{-}\mathsf{SMAT}$ proper inclusions?
  - Is $\mathsf{NoPE\text{-}AHAT}[\leq \mathsf{L}, \mathsf{U}]$ equal to semi-algebraic sets with polynomials of degree $\leq L$?

Thank you!

arXiv: 2505.11199