Rational Transductors

Mehryar Mohri

Google Research & Courant Institute of Mathematical Sciences

ICML 2026 | **Oral Presentation**

Transformers: Remarkable Success, Fundamental Limitations

Remarkable success [Vaswani et al., 2017]**:**

- Self-attention models long-range *semantic* dependencies.
- De facto standard for NLP, code, vision.

Established theoretical limitation:

Circuit complexity bounds

Without intermediate recurrence or CoT:

- Hard attention $\Rightarrow$ AC^0 [Hahn, 2020]
- Soft attention $\Rightarrow$ TC^0 [Merrill & Sabharwal, 2023; Chiang 2024]

TC^0 models can in principle approximate parity, but **lack the inductive bias to learn solutions that generalize to unseen lengths** [Merrill et al., 2022].

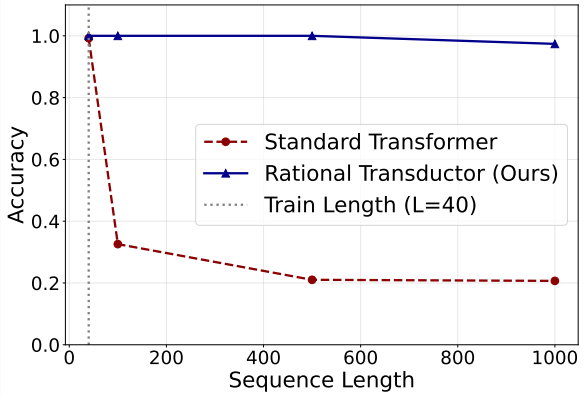
Concrete limitations:

Task	Transformer
Parity	✗
Modular counting (MOD_k)	✗
State tracking	✗
Length generalization	✗
Carry-bit propagation	✗

Diagnosis

Transformers excel at **semantics** but lack a **syntactic** inductive bias.

Illustration: The Length Generalization Failure



Setup:
Modulo-5 counting.
Trained on $L = 40$,
tested up to $L = 1{,}000$ ($25\times$).

Transformer (red):
Catastrophic positional drift.
Collapses to random chance (20%).

Rational Transductor (blue):
$>99\%$ accuracy at all test lengths.

The problem

The Transformer has **no mechanism** to extrapolate positional information beyond its training horizon.

This Talk

1. Can we augment Transformers with a **syntactic co-processor** that addresses all these limitations, while **preserving** $O(\log T)$ **parallel efficiency**?
2. Can we provide **strong theoretical guarantees** for these new models? (expressivity, optimization, generalization)
3. Can we **empirically validate** the benefits of these models on algorithmic and length generalization tasks?

Roadmap

1. **Weighted Finite Automata:** the syntactic co-processor
2. **Architecture:** the sidecar design
3. **Expressivity:** closing the Regular Gap
4. **Optimization:** Unified Scaled Cayley
5. **Learning guarantees:** error bounds & generalization
6. **Experiments:** validating the theory

Weighted Finite Automata (WFA)

Definition

A WFA $\mathcal{A} = (\Sigma, d, \boldsymbol{\alpha}, \{\mathrm{M}_\sigma\}_{\sigma\in\Sigma})$:

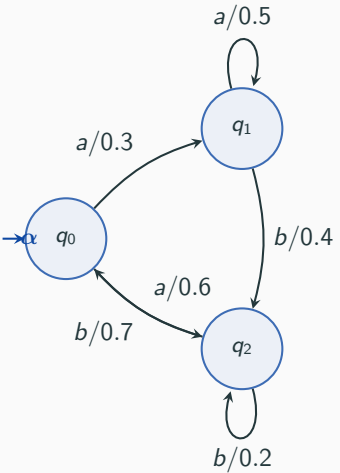
- Σ: finite alphabet
- d: state dimension
- $\boldsymbol{\alpha} \in \mathbb{R}^d$: initial state vector
- $\mathrm{M}_\sigma \in \mathbb{R}^{d\times d}$: transition matrix for token σ

Input: $x = x_1\, x_2 \cdots x_T \in \Sigma^*$.

Linear state update:

$$\mathbf{h}_t = \mathrm{M}_{x_t}\, \mathbf{h}_{t-1}, \qquad \mathbf{h}_0 = \boldsymbol{\alpha}$$

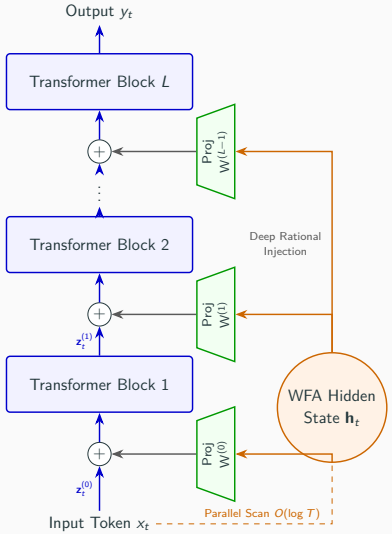
- Each input token x_t selects its own matrix M_{x_t}.
- **Linear** recurrence $\Rightarrow$ parallel associative scan in $O(\log T)$.
- State dynamics correspond *exactly* to Rational Power Series (Schützenberger, 1961).



Each arc label: **token / weight**.

Matrices $\mathrm{M}_a, \mathrm{M}_b \in \mathbb{R}^{3\times 3}$ encode transition weights.

Architecture: The Sidecar Design



Two-stream decoupling:

Transformer: semantic content.

Rational Head: syntactic state tracking.

Deep Rational Injection at every layer l:

$$\tilde{\mathbf{z}}_t^{(l)} = \mathbf{z}_t^{(l)} + \mathrm{W}_{\mathrm{proj}}^{(l)}\,\mathbf{h}_t$$

Fresh, uncorrupted state at *every* depth.

Key properties

- WFA computed from inputs via parallel scan: $O(\log T)$.
- No sequential dependency added to the Transformer.
- Unlike deep SSMs (Mamba): no layer-wise sequential bottleneck.

Expressivity: Closing the Regular Gap

Theorem: Parity Separation

No fixed-depth Transformer can *uniformly* recognize parity (i.e., with a single model for all lengths). A Rational Transductor with $d{=}2$ recognizes it for *any* length T.

Proof: Swap matrix $\mathrm{M}_1 = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ flips state between Even/Odd exactly.

Theorem: Exact MOD_k Counting

$\forall\, k{\geq}2$: $\exists$ a $d{=}k$ Transductor that recognizes MOD_k for any length.
Finite-depth Transformers cannot uniformly.

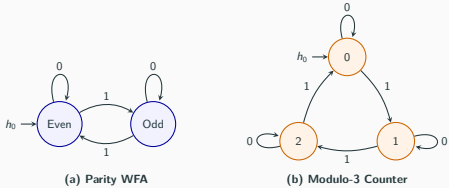
Proof: Cyclic permutation matrix; orthogonal $\Rightarrow$ lossless for any T.

Theorem: Krohn-Rhodes Completeness

A Stacked Rational Transductor exactly simulates *any* DFA.

Proof: Groups $\rightarrow$ Rational Head; aperiodic monoids $\rightarrow$ Transformer FFNs.

WFA constructions:



(a) Parity WFA (b) Modulo-3 Counter

Task	Transf.	Transd.
Parity	✗	✓
MOD_k	✗	✓
All Reg. Lang.	✗	✓
Length gen.	✗	✓

Circuit complexity

$$\mathrm{AC}^0 \subsetneq \mathrm{TC}^0 \subseteq \mathrm{NC}^1 \subseteq \mathcal{F}_{\mathrm{RT}} \subseteq \mathrm{PNC}^1$$

Contains all of NC^1.
Strictly below P-complete RNNs.

Optimization: Unified Scaled Cayley Parameterization

Problem: How do we parameterize the transition matrices M_t so that gradient descent remains well-behaved over arbitrarily long sequences?

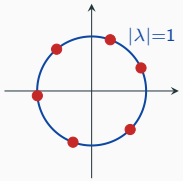
Solution:

$$M_t = \underbrace{g_t}_{\text{scalar gain}} \cdot \underbrace{(I + A_t)(I - A_t)^{-1}}_{\text{Cayley transform} \in SO(d)}$$

- $A_t = W_t - W_t^\top$ is skew-symmetric (learnable W_t).
- The Cayley transform maps A_t to a **rotation** matrix: all eigenvalues on the unit circle, $\|\mathcal{C}(A_t)\|_2 = 1$.
- $g_t \in (0, 1]$ is a learnable scalar gain controlling energy:

Regime	g_t	**Effect on** M_t
Conservation	$g_t = 1$	$\|M_t\|_2 = 1$: exact countin
Decay	$g_t = \text{sigmoid}(\theta_t)$	$\|M_t\|_2 < 1$: gating

Since $\|M_t\|_2 = g_t$, $\gamma = \max_t g_t$ controls stability.



Eigenvalues on the unit circle when $g_t = 1$ (conservation).

Theorem: Gradient Preservation

Let $\boldsymbol{\delta}_t = \nabla_{\mathbf{h}_t} \mathcal{L}$ be the gradient of the loss w.r.t. the state $\mathbf{h}_t$.

At γ=1: gradient norm **exactly preserved**.

At γ<1: $\|\boldsymbol{\delta}_{t-k}\|_2 \leq \gamma^k \|\boldsymbol{\delta}_t\|_2 + C$ (controlled decay, no explosion).

Stability is **structural**, not enforced by gradient clipping.

Learning Guarantee I: Length Generalization

Why does it generalize to unseen lengths?

The model learns a *time-invariant* transition rule: the same M_σ at every position.

Theorem: Time-Invariant Error Bound

Let M^* be the **true** (target) transition matrices and $\hat{\mathrm{M}}$ the **learned** ones.

If $\left\|\hat{\mathrm{M}}\right\|_2 \leq \gamma < 1$ and $\left\|\hat{\mathrm{M}} - \mathrm{M}^*\right\| \leq \varepsilon$:

$$\sup_{t \geq 1} \left\|\mathbf{h}_t^* - \tilde{\mathbf{h}}_t\right\| \leq \frac{\varepsilon\, C}{1-\gamma}$$

Bound independent of sequence length T.

Proof idea: Error recurrence $\|e_t\| \leq \gamma \|e_{t-1}\| + \varepsilon C$ $\Rightarrow$ geometric series $\rightarrow \frac{\varepsilon C}{1-\gamma}$.

Error *saturates*: does not compound.

Contractive regime ($\gamma < 1$)

- Error bounded by $\frac{\varepsilon C}{1-\gamma}$, uniformly.
- No positional drift.
- Justifies disabling positional encodings entirely.

Conservation regime ($\gamma = 1$)

Orthogonal M $\Rightarrow$ **algebraic exactness**.

Learned WFA isomorphic to target automaton.

State tracking is **exact**, not approximate.

Both regimes guarantee length generalization, through different mechanisms.

Learning Guarantee II: T-Independent Generalization

Standard RNN generalization bounds scale with T. Ours do not.

Theorem: Generalization Bound

Let H_f be the **Hankel matrix** of a rational function f: an infinite matrix indexed by $(u, v) \in \Sigma^* \times \Sigma^*$ with entries (f, uv).

The Rademacher complexity of rational functions with $\|H_f\|_{\text{nuc}} \leq r$ is:

$$\hat{\mathfrak{R}}_S(\mathcal{F}_{\text{RT}}) = \tilde{O}\left(\frac{r}{\sqrt{N}}\right)$$

No dependence on sequence length T.

The key point

State dimension d

controls generalization,

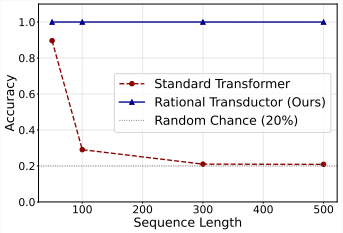
not sequence length T

Why? By Fliess' theorem, the rank of the Hankel matrix H_f equals the minimal automaton dimension $d_{\min}$.

Bounding its nuclear norm $\|H_f\|_{\text{nuc}} \leq r$ is equivalent to regularizing the

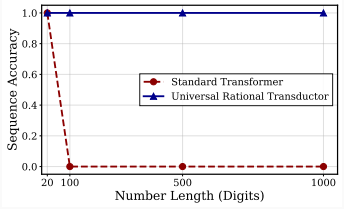
effective state dimension.

Experiments: Validating the Theory



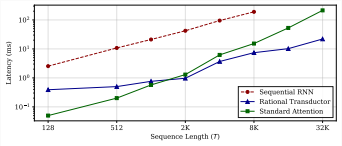
Modulo-5 Counting.
Transformer (red): 20%.

Transductor (blue): **100%**.



1000-Digit Addition.
Carry-bit propagation.
Transformer: 0% at L=100.

Transductor: **100% at** L=1,000.



Latency Scaling.
Parallel scan: $O(\log T)$.
Overtakes RNN at T=512.

Avoids $O(T^2)$ attention wall.

Universal Rational Transductor

Orthogonal head (exact counting) $\oplus$ Stochastic head (discrete state transitions): mirrors the Krohn-Rhodes decomposition at the architectural level.

Summary: Key Properties of Rational Transductors

- ✓ **Efficiency:** $O(\log T)$ parallel training via associative scans. No sequential bottleneck.
- ✓ **Expressivity:** $\mathrm{NC}^1 \subseteq \mathcal{F}_{\mathrm{RT}} \subseteq \mathrm{PNC}^1$. Contains all of NC^1. Krohn-Rhodes complete.
- ✓ **Optimization:** Unified Scaled Cayley guarantees gradient stability structurally. No gradient explosion at $\gamma{=}1$.
- ✓ **Generalization:** Time-invariant error bound: $\sup_t \|e_t\| \leq \frac{\varepsilon C}{1-\gamma}$. Hankel-Rademacher: $\tilde{O}(r/\sqrt{N})$, no T.
- ✓ **Experiments:** 25× length generalization. 100% on 1000-digit addition.

Complexity landscape

$$\mathrm{AC}^0 \subsetneq \mathrm{TC}^0 \subseteq \mathrm{NC}^1 \subseteq \mathcal{F}_{\mathrm{RT}} \subseteq \mathrm{PNC}^1$$

Future directions:

- Large-scale pretraining
- Program synthesis & verification
- Pushdown (context-free) extensions

Rational Transductors provide a **theoretical backbone** for next-generation sequence models: augmenting Transformers with WFAs yields both **semantics** and **syntax**, with rigorous theoretical guarantees.

Thank You

Questions?

Backup: How Does This Differ from Mamba / SSMs?

Deep SSMs (Mamba, H3):

- Recurrence *inside* the backbone.
- Layer-wise sequential dependency.
- Training depth: $O(L \log T)$.

Rational Transductors:

- Recurrence *outside* (sidecar).
- Deep Injection to all layers.
- Training depth: $O(\log T)$ exactly.
- First proof of NC^1 completeness.

Modularity

The Rational Head can be bolted onto any existing Transformer to grant it WFA state-tracking capabilities.

Backup: Why Not Just Use Chain-of-Thought?

CoT can make Transformers Turing Complete via scratchpad.

Why Rational Transductors are preferable for state tracking:

1. **Speed:** CoT forces $O(T)$ autoregressive decoding. Rational Transductors: $O(\log T)$ forward pass.
2. **Context waste:** CoT consumes tokens just to track loop variables.
3. **Precision:** CoT still struggles to learn robust counting over 1,000+ steps. Cayley parameterization is exact by construction.

Backup: Context-Free Languages?

Q: Can Rational Transductors handle CFLs (Dyck-k)?

A: No. By design.

- CFLs require an unbounded pushdown stack.
- A true neural stack forces $O(T)$ sequential execution.
- We restrict to Regular Languages / PNC^1 to guarantee $O(\log T)$ parallel training.

Backup: Scaling to Large Vocabularies?

Q: A matrix per token in a 50k+ vocabulary?

Shared Basis Parameterization:

$$\mathrm{M}_\sigma = \sum_{k=1}^{K} a_{\sigma,k}\, \mathrm{B}_k$$

- K shared basis matrices B_k.
- Each token σ: tiny K-dim coefficient vector.
- Memory: $O(|\Sigma|K + Kd^2)$ instead of $O(|\Sigma|d^2)$.

Backup: Virtual Tensorization

Injecting $\mathbf{h}_t$ into attention queries/keys implicitly computes a kernel over $\mathbf{h}_t \otimes \mathbf{h}_{t'}$.

Virtual Tensorization

Attention score: $S_{\text{rat}}(t, t') = \mathbf{h}_t^\top M_{\text{int}} \mathbf{h}_{t'}$

$\Rightarrow$ linear in $\mathbf{h}_t \otimes \mathbf{h}_{t'}$.

State dimension d simulates d^2 **for free**.