The Expressive Capacity of State Space Models: A Formal Language Perspective

Yash Sarrof, Yana Veitsman, and Michael Hahn

Motivation

Research on Expressivity of RNNs and Transformers

Research on Expressivity of SSMs *

*Merrill, W., Petty, J., & Sabharwal, A. The Illusion of State in State-Space Models. In *Forty-first International Conference on Machine Learning*.
*Jelassi, S., Brandfonbrener, D., & Kakade, S. M. Repeat After Me: Transformers are Better than State Space Models at Copying. In *Forty-first International Conference on Machine Learning*.

Non-negative SSMs

All entries in $A(x_t) >= 0$

- Examples : **Mamba***, GLA, HGRN2

$$h_t = \overline{A}h_{t-1} + \overline{B}x_t \qquad \overline{A} = \exp(\Delta A)$$
$$y_t = Ch_t$$

$$h_t = A(x_t) \circ h_{t-1} + B(x_t)$$

* Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *CoLM 2024*

Star-Free Languages

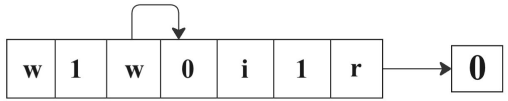
Regular language class that is closed under finite union, product and complement

but not

Kleene-star, aka *

Example :: Flip Flop, Bounded Dyck

Flip-Flop

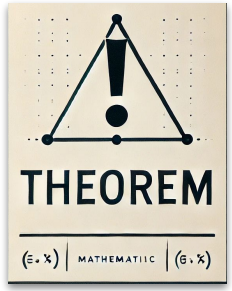


Minimalistic long-range dependency benchmark ~ Proxy for closed domain hallucinations.

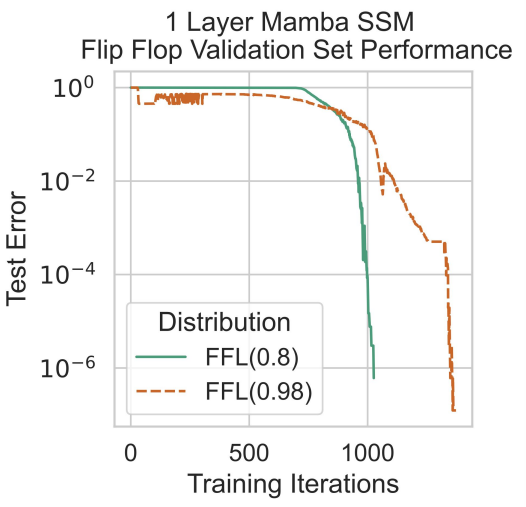
Liu, Bingbin, et al. "Exposing attention glitches with flip-flop language modeling - NeurIPS 2023

A 2 Layer SSM predictively models the Flip flop language

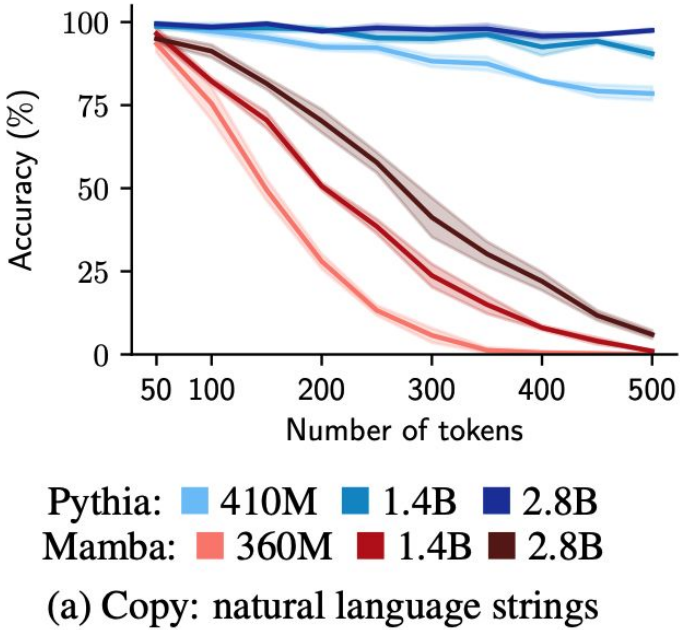
- at **arbitrary input lengths**
- with **finite precision**.

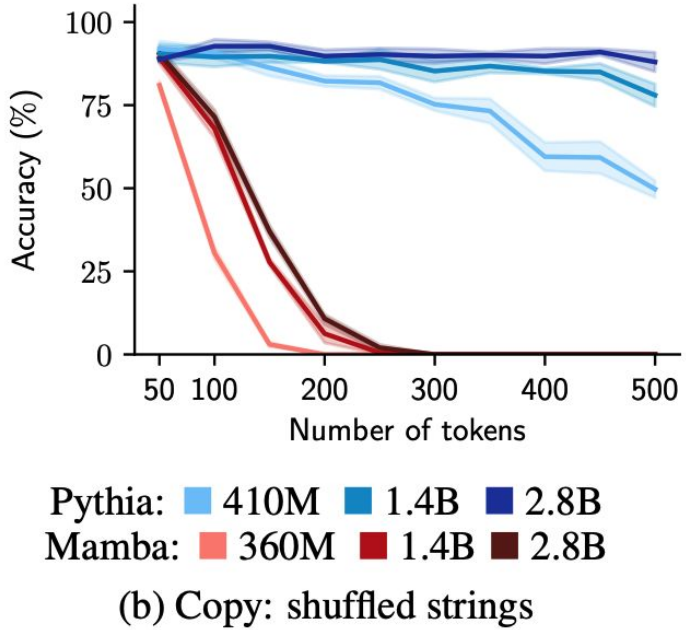


SSMs resolves a critical failure mode of self attention



Transformers can copy, SSMs can’t

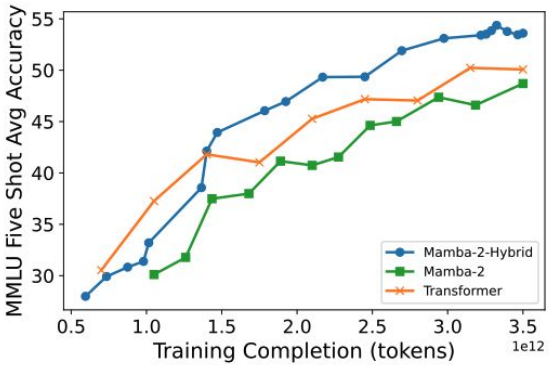




Jelassi, Samy, et al. "Repeat after me: Transformers are better than state space models at copying." *ICML* - 2024

Takeaway #1

- Complementary Abilities b/w SSMs & Transformers
- Future: Hybrid Architecture



- Lieber, Opher, et al. "Jamba: A hybrid transformer-mamba language model." *arXiv preprint arXiv:2403.19887* (2024).
- Waleffe, Roger, et al. "An Empirical Study of Mamba-based Language Models." *arXiv preprint arXiv:2406.07887* (2024).
- Ren, Liliang, et al. "Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling." *arXiv preprint arXiv:2406.07522* (2024).

Non Star-Free Language

All Regular language that are not Star-Free :)

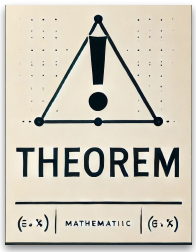
Along with Union, Product and Complement

REQUIRE THE INCLUSION OF
Kleene-star *

Example :: PARITY

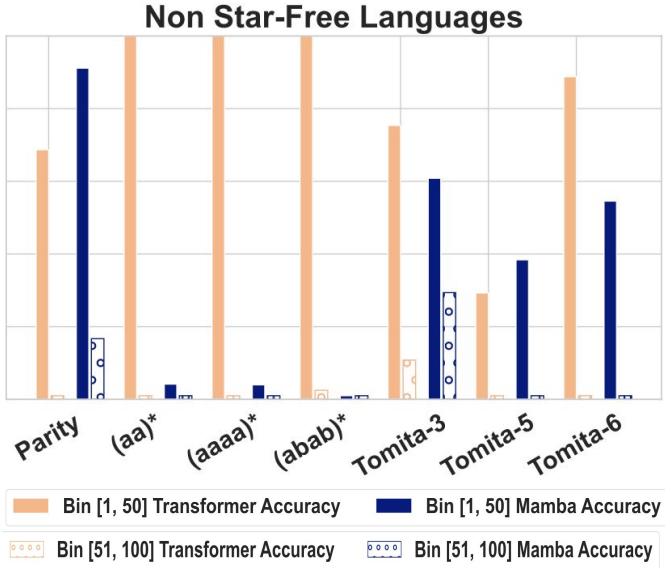
NONNEGATIVE SSMs cannot recognize PARITY

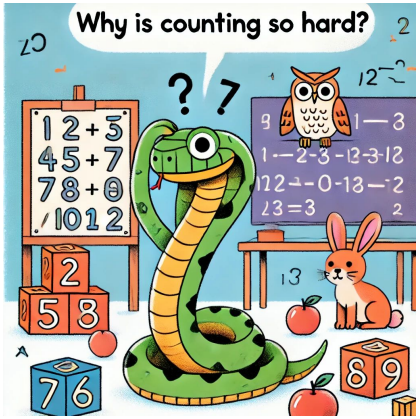
- at **arbitrary input lengths**
- with **finite precision**.



Takeaway #2

- SSMs will struggle with Modular counting whenever required. (Non Star Free languages require it).

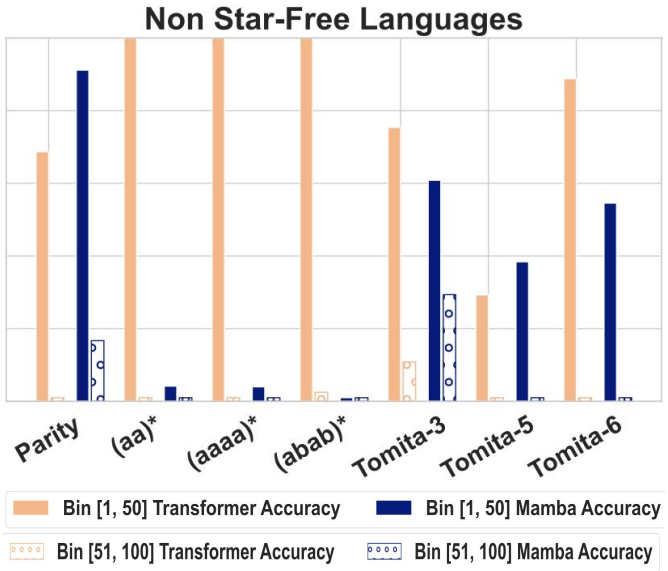


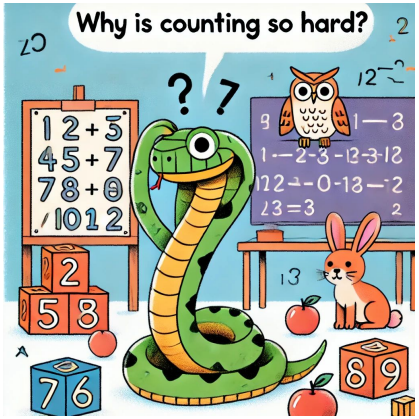


Bhattamishra, Satwik, Kabir Ahuja, and Navin Goyal. "On the Ability and Limitations of Transformers to Recognize Formal Languages."- *EMNLP* 2020.

Takeaway #1

- SSMs will struggle with Modular counting whenever required. (Non Star Free languages require it).
- Certain Design Choices cause this limitation

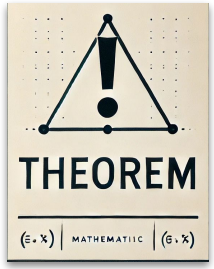




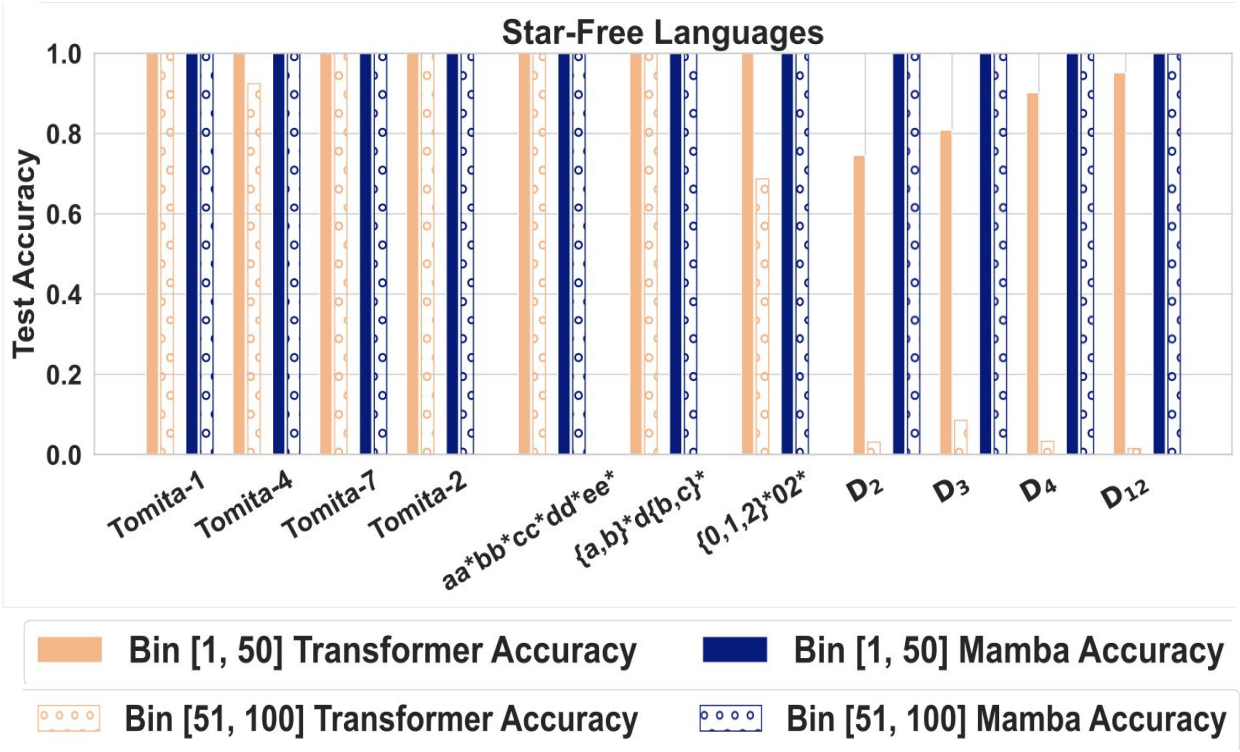
Bhattamishra, Satwik, Kabir Ahuja, and Navin Goyal. "On the Ability and Limitations of Transformers to Recognize Formal Languages."- *EMNLP* 2020.

NON-NEGATIVE SSMs can predictively model Regular Languages

- iff **the language is star free**
- with **finite precision**.



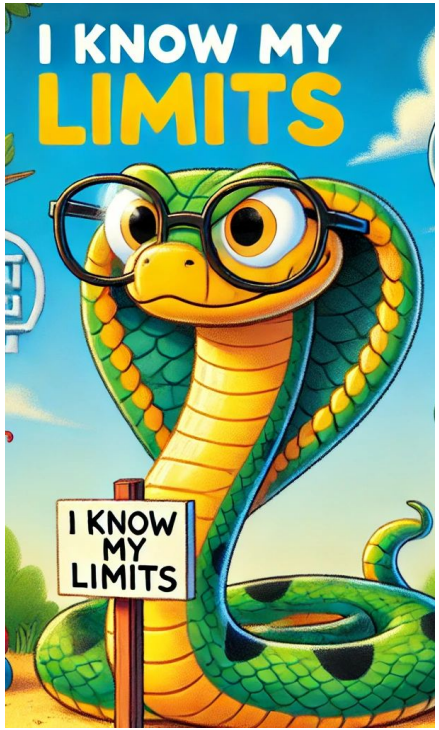
Empirical Results



Bhattamishra, Satwik, Kabir Ahuja, and Navin Goyal. "On the Ability and Limitations of Transformers to Recognize Formal Languages."- *EMNLP* 2020.

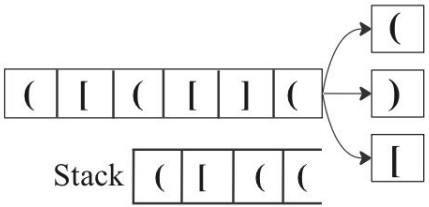
Takeaway #3

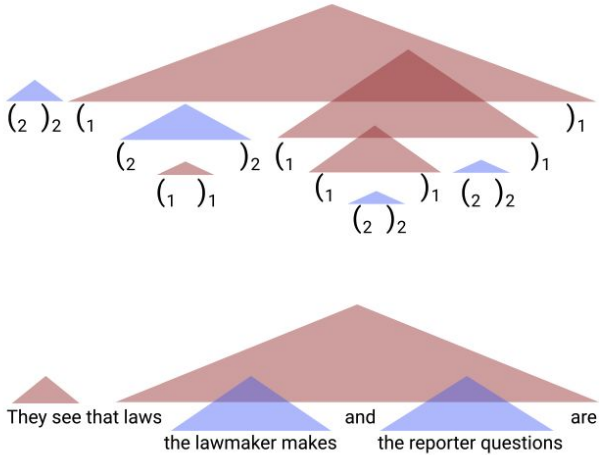
- Exact characterisation of Transformers* in Finite state case : Difficult.
- With SSMs, it’s possible !



*Angluin, Dana, David Chiang, and Andy Yang. "Masked hard-attention transformers and boolean RASP recognize exactly the star-free languages." *NeurIPS 2024*

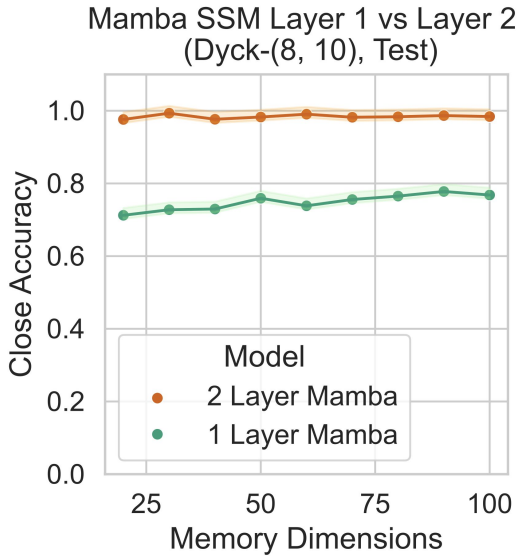
Bounded Dyck : Dyck(K, m)





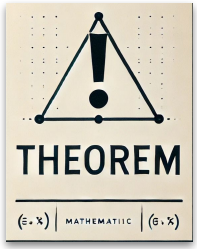
Experimental Results

- Dyck(K, m) : Regular language
  Solution guaranteed (not necessarily efficient)
- Explicit Stack not required
- EFFICIENT (shortcut through counting)



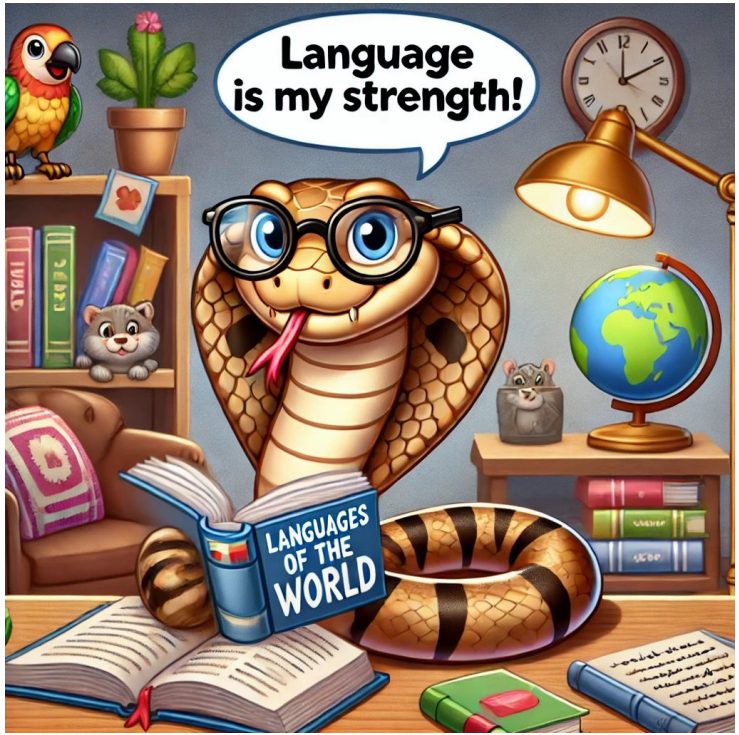
A 2 Layer SSM predictively models Bounded Dyck (K, m)

- with **d = O(mlog K)**
- with **finite precision**.

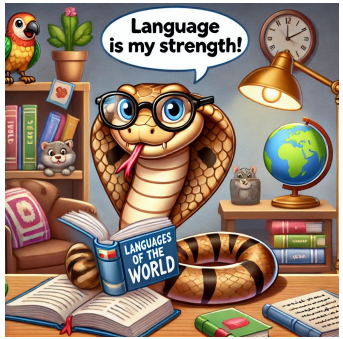


Takeaway #4

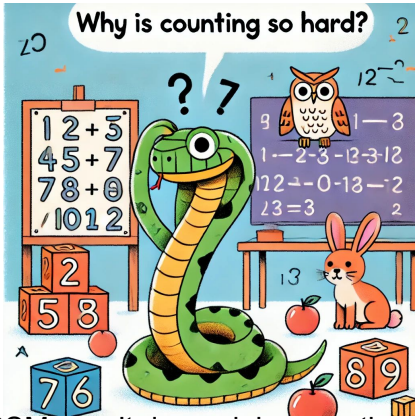
- SSMs can keep track of bounded hierarchical structures EFFICIENTLY !
- SSMs can model hierarchical structure of language



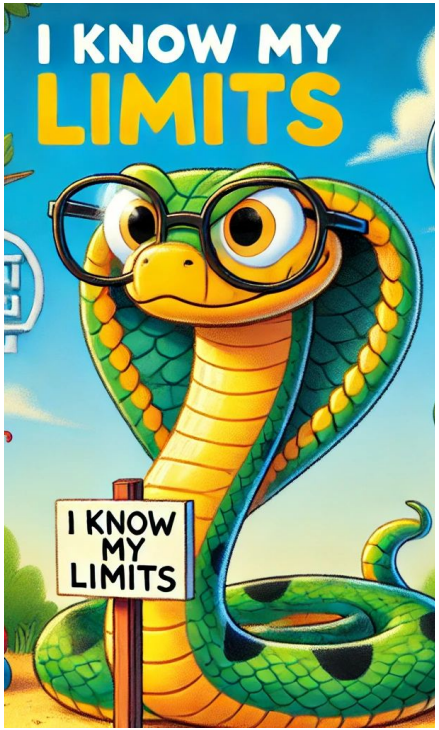
Recapping our takeaways



SSMs can model hierarchical structure of language



SSMs can’t do modular counting



It would be easier to theoretically predict failures & abilities LLMs based on SSMs.

LLMs will have Hybrid Architectures