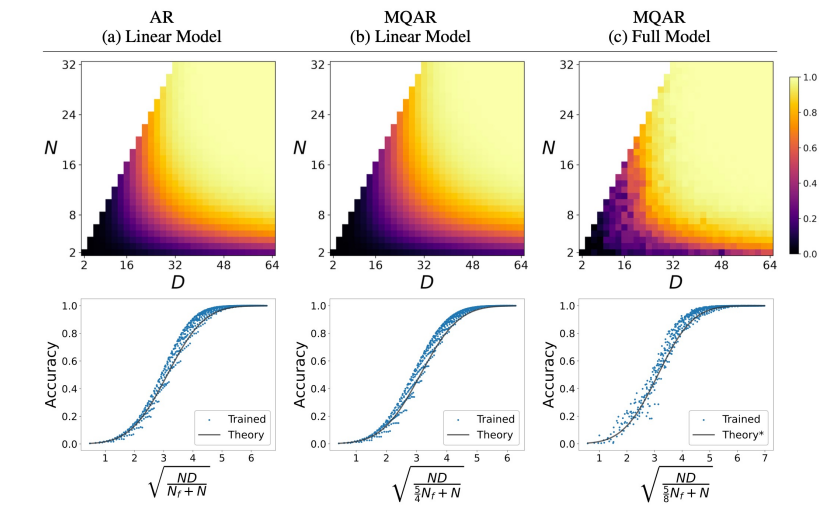
On the Recall Scaling Laws in Mamba: A Theoretical and Mechanistic Study via Hashing

Yuval Koren, Assaf Ben-Kish, Raja Giryes, Lior Wolf, Itamar Zimerman



Paper: https://arxiv.org/abs/2609.07681 GitHub: https://github.com/yuvalko1/mamba-recall-scaling-laws Contact: yuvalkoren4@mail.tau.ac.il

Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

- **Transformers** achieve remarkable success in sequence modeling
- However, their memory complexity grows (linearly) with context length
- **Mamba** and linear RNNs use a **fixed-size state**, enabling constant memory complexity
- Due to limited memory, these models inherently suffer from information loss
- To improve **memory utilization** in Mamba, it is crucial to understand how it solves in-context retrieval tasks, such as **associative recall**

Methodology and Overview

NEURAL INFORMATION PROCESSING SYSTEMS

Methodology

- We use the synthetic ***Associative Recall*** benchmark,
  to study Mamba recall mechanism in a **controlled** setup
- To find the **bottlenecks** for recall, we **remove** model components,
  to get a **minimal** Mamba that can still recall

Our work overview

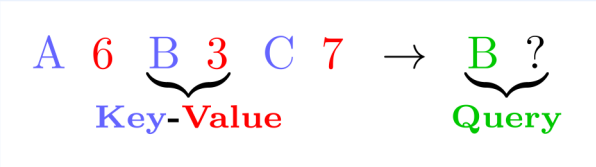
1. **Theoretical circuit:** how *can* Mamba recall?
2. **Mechanistic experiments:** how *does* Mamba recall, in practice?
3. **Theoretical scaling laws:** how should recall *scale* with Mamba dimension?
4. **Scaling Laws Experiments:** verifying the theoretical laws in practice

Associative Recall

NEURAL INFORMATION PROCESSING SYSTEMS

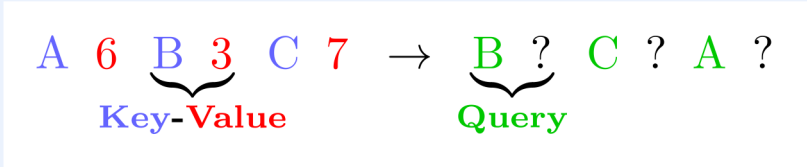
- **AR** (**A**ssociative **R**ecall)

  Facts are followed by a single query



- **MQAR** (**M**ulti-**Q**uery **A**ssociative **R**ecall)

  Facts are followed by multiple queries



Expected input-output pair:
(the * tokens are ignored)

x	A	6	B	3	C	7	B
y	*	*	*	*	*	*	3

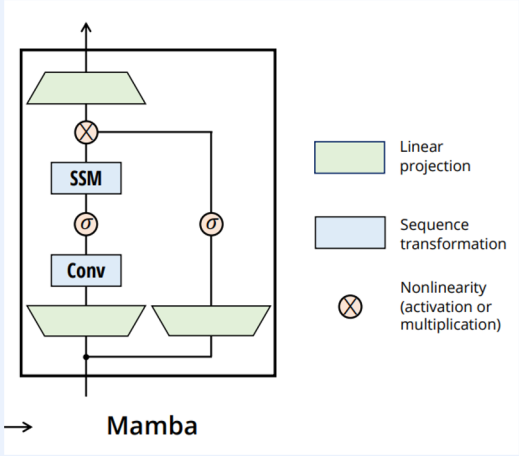
Expected input-output pair:
(the * tokens are ignored, random padding tokens added)

x	A	6	B	3	C	7	B	2	5	0	9	C	4	A	8	1
y	*	*	*	*	*	*	3	*	*	*	*	7	*	6	*	*

Simplified Mamba Model

NEURAL INFORMATION PROCESSING SYSTEMS

- To find the **bottlenecks** for recall, we **remove** model components, to get a **minimal** Mamba that can still recall



Keep:

- Linear layers
- Convolution
- State-Space Model

Remove:

- Gating
- Nonlinearities
- Discretization

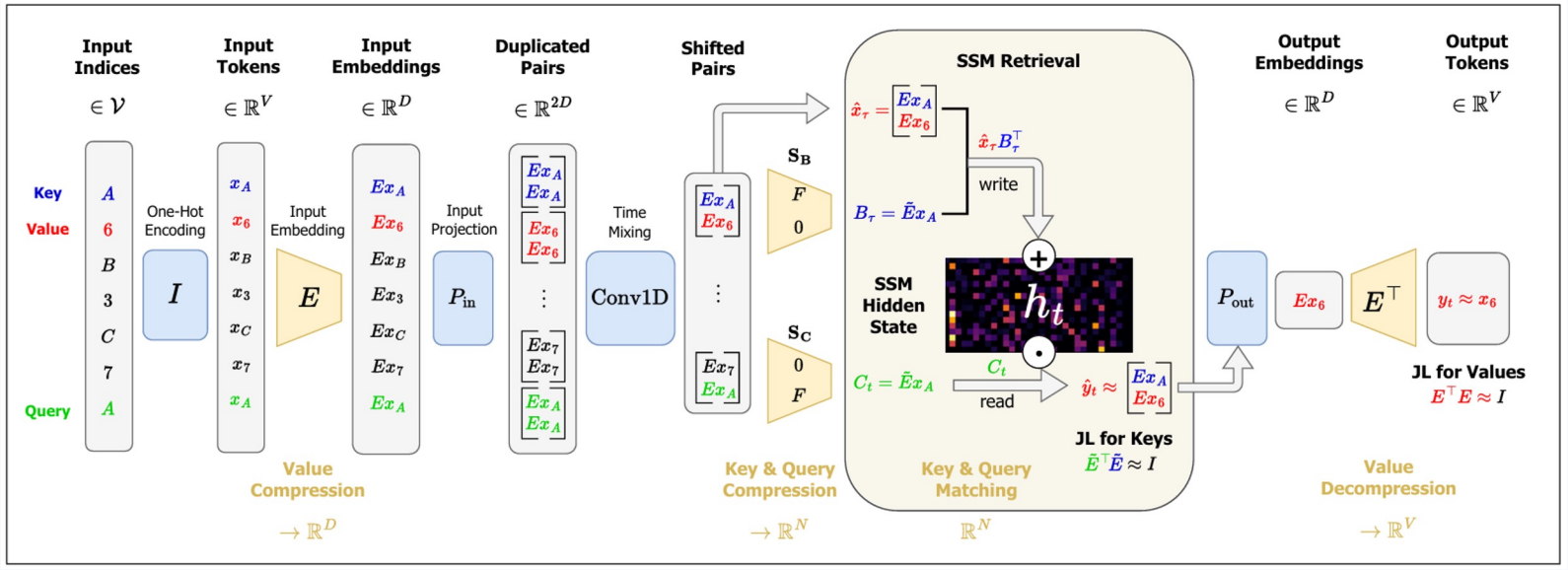
Algorithm 1 Simplified Single-Layer Mamba

Input: $x \in \mathbb{R}^{V \times L}$
Output: $y \in \mathbb{R}^{V \times L}$
Parameters: E, P_{in}, P_{out}, W_{conv}, S_B, S_C
$x_e \leftarrow E\,x$ // input embedding
Mixer:
$x_{\text{in}} \leftarrow P_{\text{in}}\,x_e$
$\hat{x} \leftarrow \text{Conv1D}(x_{\text{in}}, W_{\text{conv}})$
$B \leftarrow S_B\,\hat{x}$
$C \leftarrow S_C\,\hat{x}$
$\hat{y} \leftarrow \text{SSM}(\hat{x}, B, C)$
$y_{\text{out}} \leftarrow P_{\text{out}}\,\hat{y}$
$y \leftarrow E^\top\,y_{\text{out}}$ // output embedding
return y

Theoretical Circuit: How *can* Mamba recall?

NEURAL INFORMATION PROCESSING SYSTEMS

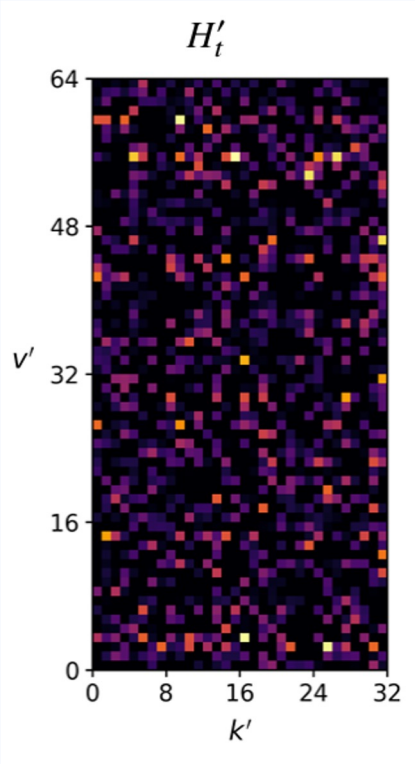
- Projection **duplicates**, Conv1D **copies and shifts**, to form key-value pairs
- SSM **writes** key-value pairs to memory, and **reads** values where query matches key
- Reliable value reconstruction: using ***JL Lemma*** once (on the in/out **embeddings**)
- Reliable query-key matching: using ***JL Lemma*** again (on the **SSM** matrices)



Interpretability: How *does* Mamba recall?

NEURAL INFORMATION PROCESSING SYSTEMS

- **Goal:** find evidence to our **theoretical circuit** in **trained models**
  - Train Mamba on MQAR
  - Extract weights, activations, etc.
  - Inspect the hidden state
  - Interpretation: ???
  - Looks like a random noise

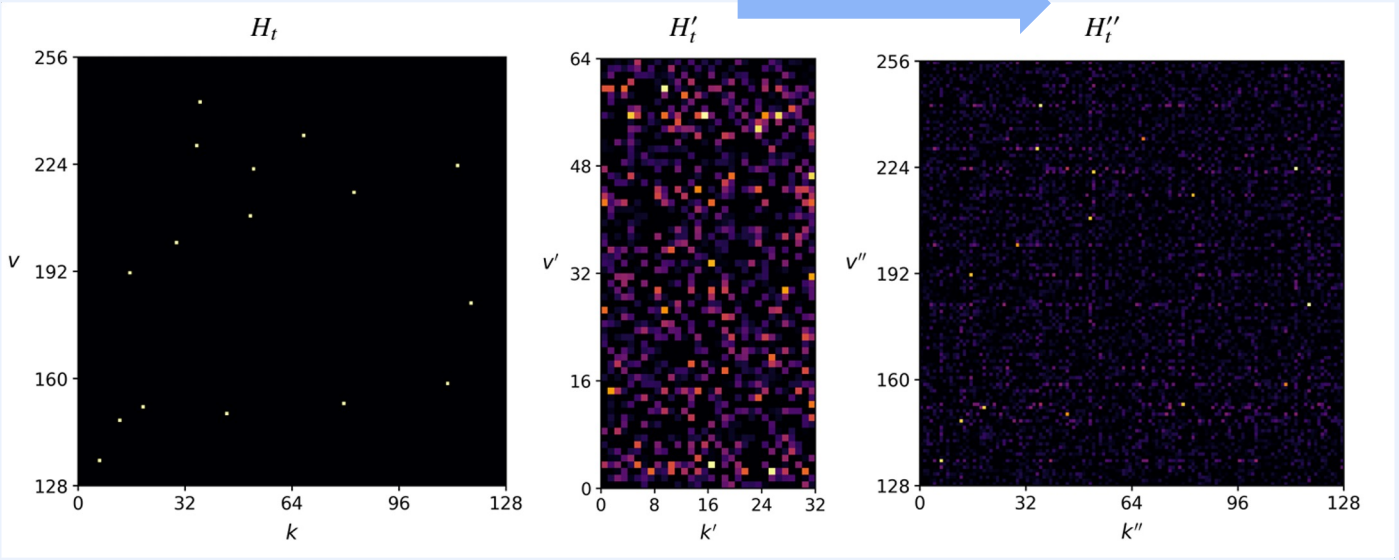


Interpretability: How *does* Mamba recall?

- Ideally, we could simply **project** the hidden state back to vocabulary space
- Interpretation (informal): a linear ***Hash table***,
  where hashed key entries store compressed values

$$H_t = \sum_{n=1}^{N_f} v_n k_n^\top$$

$$H_t'' = E^\top H_t' \tilde{E} \approx H_t.$$



Interpretability: How *does* Mamba recall?

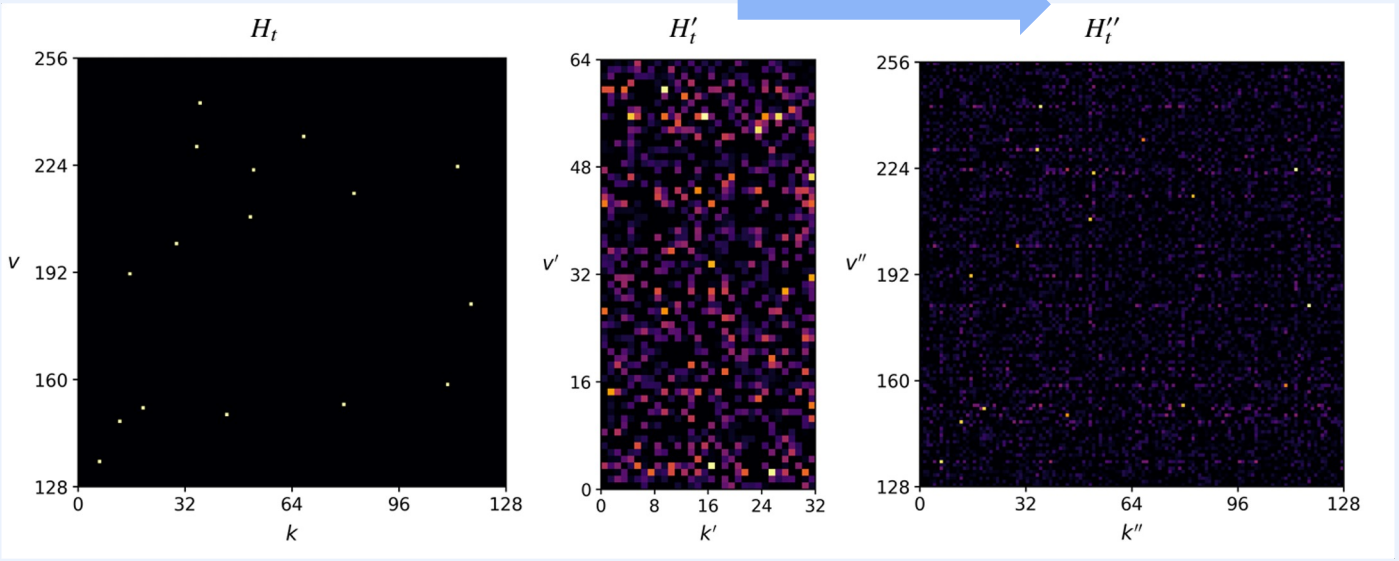
- Ideally, we could simply **project** the hidden state back to vocabulary space
- Actual alg. is more complicated, but similar in principle

$$H_t = \sum_{n=1}^{N_f} v_n k_n^\top$$

Algorithm 2 Hidden State Inversion

Input: $h_t \in \mathbb{R}^{2D \times N}$ // $D_{\text{in}} = 2D$
Output: $h_t^{\text{inv}} \in \mathbb{R}^{2V \times V}$
Parameters:
$E \in \mathbb{R}^{D \times V}$, $P_{\text{in}} \in \mathbb{R}^{2D \times D}$,
$W_{\text{conv}} \in \mathbb{R}^{2D \times 2}$ // $D_{\text{conv}} = 2$
$S_B, S_C \in \mathbb{R}^{N \times 2D}$, $P_{\text{out}} \in \mathbb{R}^{D \times 2D}$
Operators:
$\hat{E}_{\text{in}} \leftarrow \left(\text{diag}(W^0_{\text{conv}})\, P_{\text{in}}\, E \;\middle|\; \text{diag}(W^1_{\text{conv}})\, P_{\text{in}}\, E\right)$
$\Pi_{q,\text{in}} \leftarrow S_C\, \hat{E}_{\text{in}}$ // input projection
$\Pi_{v,\text{out}} \leftarrow E^\top P_{\text{out}}$ // output projection
Inversion:
$h_t^{\text{inv}} \leftarrow \Pi_{v,\text{out}}\, h_t\, \Pi_{q,\text{in}}$

$$\approx H_t.$$



Theoretical Scaling Laws

NEURAL INFORMATION PROCESSING SYSTEMS

- Core question: How does **recall accuracy** depend on **model dimensions**?
- We answer with several theorems and remarks. Importantly:

Theorem 4.2 (Trained model recall scaling laws). *A single-layer simplified Mamba can solve AR and MQAR tasks with probabilities*

$$p_{\mathrm{AR}} \approx \Phi\left(\sqrt{\frac{ND}{N_f}} - \sqrt{2\log V}\right), \quad p_{\mathrm{MQAR}} \approx \frac{1}{L-2N_f}\sum_{t=2N_f}^{L} \Phi\left(\sqrt{\frac{ND}{\frac{1}{2}N_f + \frac{1}{4}t}} - \sqrt{2\log V}\right), \quad (15)$$

in the limits of $N_f \gg N, D$ and $V \gg 1$, where Φ is the normal CDF. (See proof in App. F.4.)

- Compact, unified form:

$$p_{\mathrm{recall}} \approx \Phi\left(\sqrt{\frac{ND}{aN_f+N}} - \sqrt{2\log V}\right)$$

Theoretical Scaling Laws: Derivation (Informal Intuition)

NEURAL INFORMATION PROCESSING SYSTEMS

- **Statistical signal-to-noise analysis** of model output entries

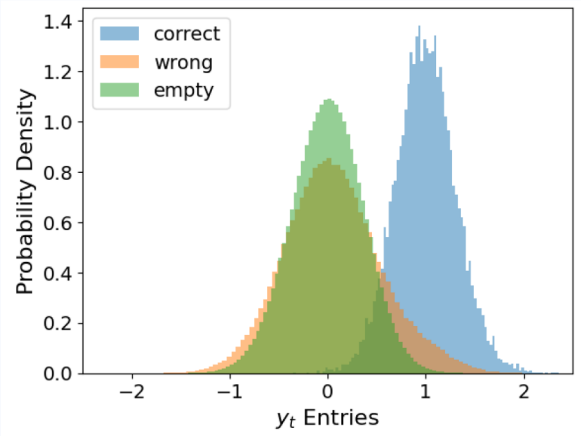
$$y_t = E^\top \sum_{\tau}^{t} (Ex_\tau)(\tilde{E}x_{\tau-1})^\top (\tilde{E}q)$$

$$y_t^i = \sum_{n=1}^{N_f} (v_i^\top G_E\, v_n)(k_n^\top G_{\tilde{E}}\, k_m)$$

Noise ~ 1/D Noise ~ 1/N

Total Noise:

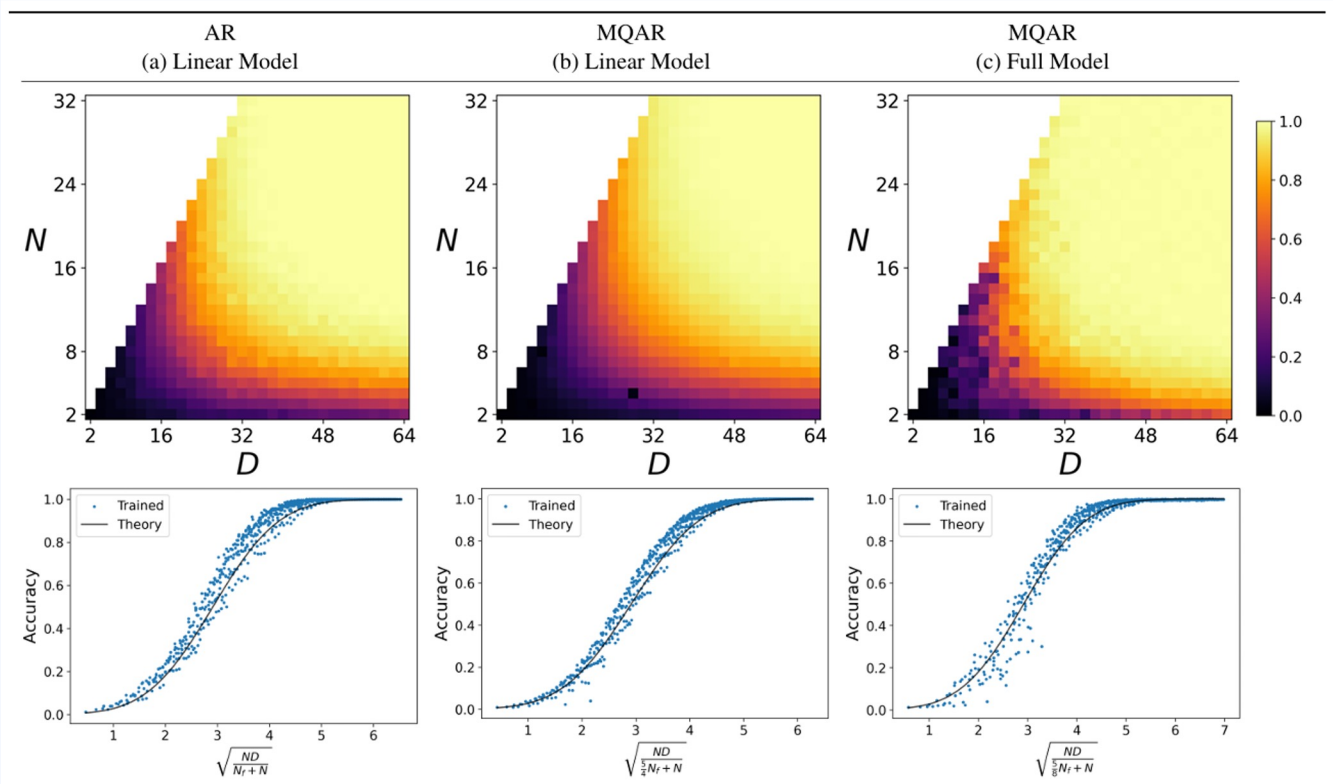
$$\sigma^2 \equiv \frac{N_f}{ND}$$



- 1 correct entry
- V-1 other (wrong fact / empty) entries
- Using the max of V Gaussians:

$$p_{\mathrm{AR}} \approx \Phi\left(\sqrt{\frac{ND}{N_f}} - \sqrt{2\log V}\right),$$

Scaling Laws Experiments



$$p_{\text{recall}} \approx \Phi\left(\sqrt{\frac{ND}{aN_f+N}} - \sqrt{2\log V}\right)$$

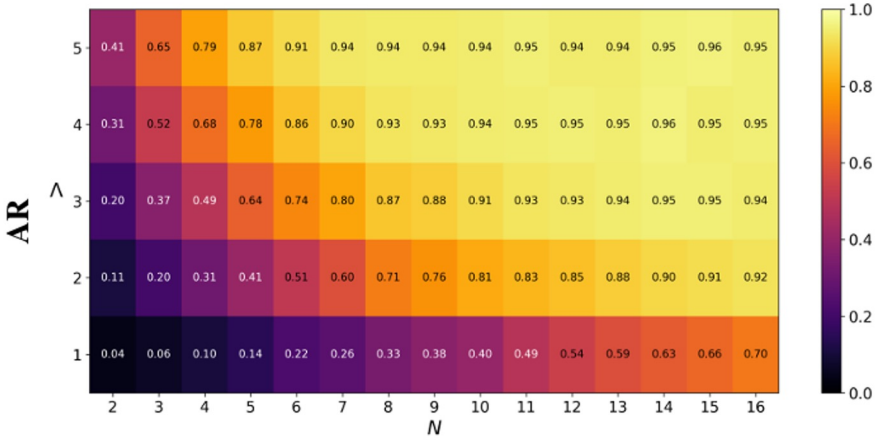
Theory and Experiments: Multi-Layer

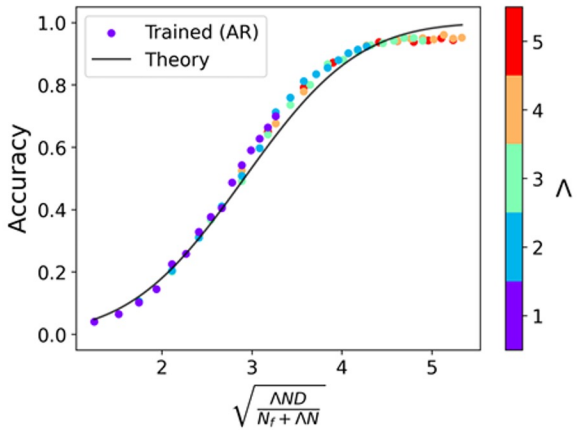
NEURAL INFORMATION PROCESSING SYSTEMS

- Our theory extends to beyond single-layer Mamba models
- Adding layers has the same effect of increasing the hidden state size

Theorem 4.6 (Multi-layer recall scaling laws). *A simplified multi-layer Mamba with Λ layers can solve associative recall tasks with a probability* $p_{\text{recall}} \approx \Phi\left(\sqrt{\frac{\Lambda ND}{aN_f+\Lambda N}} - \sqrt{2\log V}\right)$*, where* $a = 1$ *for AR, and* $a \approx \frac{5}{4}$ *for MQAR. (See proof in App. G.2.)*

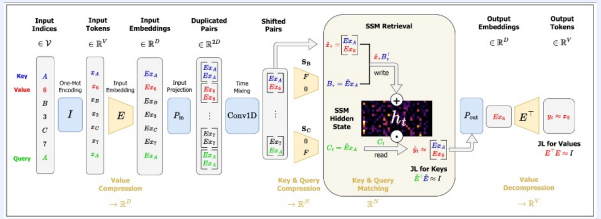
- Again, the theoretical law meets practice

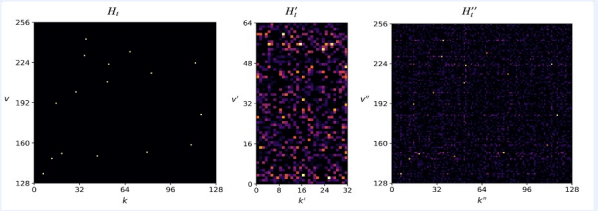


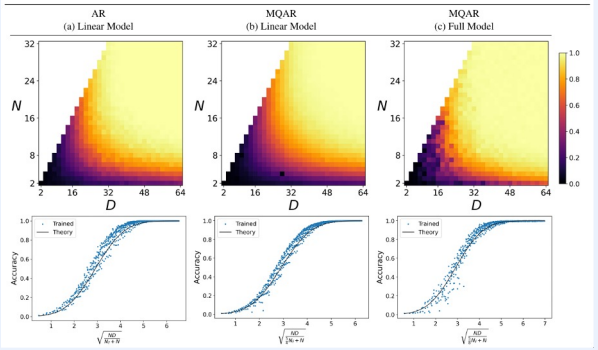


Summary and Conclusions

1. We suggest a **theoretical circuit** for recall in Mamba
2. We verify it with **Mechanistic Interpretability**
3. We derive **theoretical scaling laws**
4. Our theory is **validated** by variety of experiments



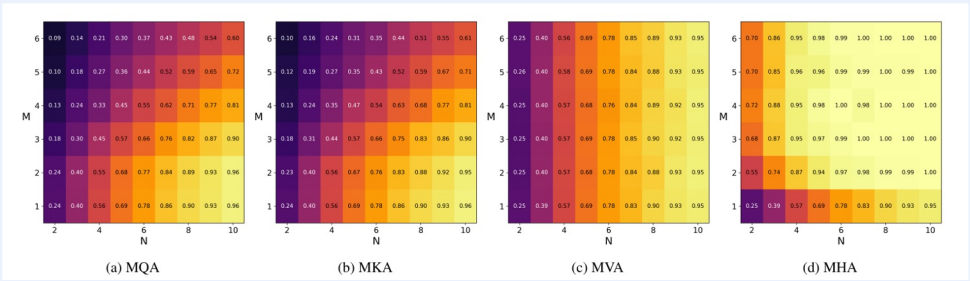


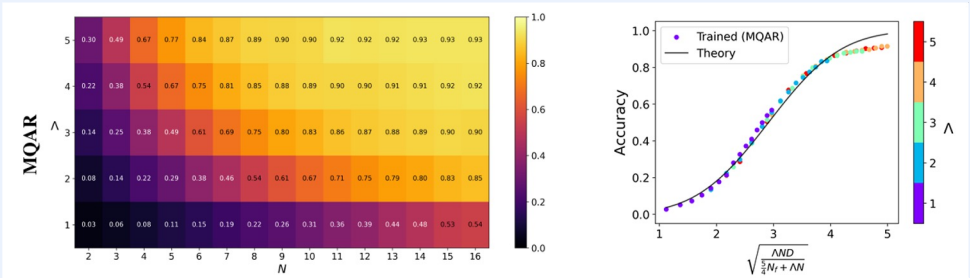


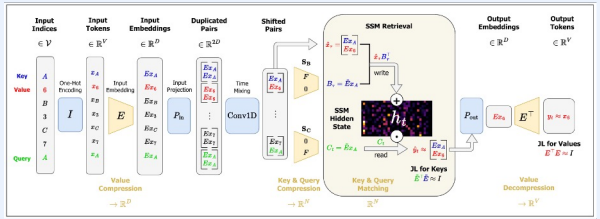
Find the topic interesting?

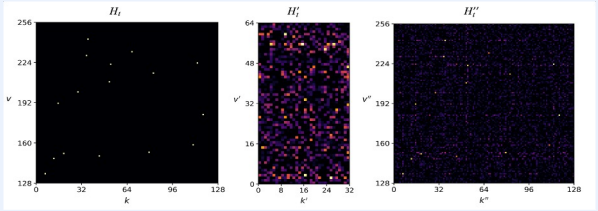
Want to know more?

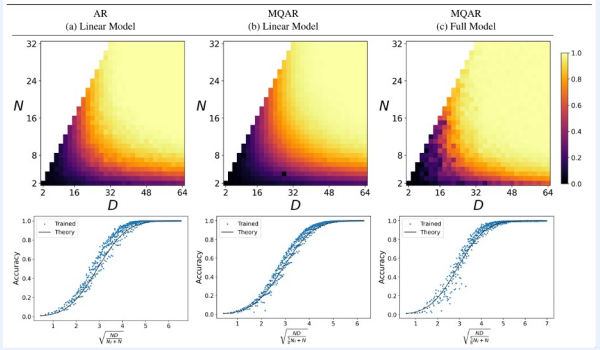
Visit our poster session!











Paper: https://arxiv.org/abs/2609.07681 GitHub: https://github.com/yuvalko1/mamba-recall-scaling-laws Contact: yuvalkoren4@mail.tau.ac.il