GATED DELTA NETWORKS: IMPROVING MAMBA2 WITH DELTA RULE

Songlin Yang *
MIT CSAIL
yangsl66@mit.edu

Jan Kautz
NVIDIA
jkautz@nvidia.com

Ali Hatamizadeh ⋆
NVIDIA
ahatamizadeh@nvidia.com

ABSTRACT

Linear Transformers have gained attention as efficient alternatives to standard Transformers, but their performance in retrieval and long-context tasks has been limited. To address these limitations, recent work has explored two distinct mechanisms: gating for adaptive memory control and the delta update rule for precise memory modifications. We observe that these mechanisms are complementary—gating enables rapid memory erasure while the delta rule facilitates targeted updates. Building on this insight, we introduce the gated delta rule and develop a parallel training algorithm optimized for modern hardware. Our proposed architecture, Gated DeltaNet, consistently surpasses existing models like Mamba2 and DeltaNet across multiple benchmarks, including language modeling, common-sense reasoning, in-context retrieval, length extrapolation, and long-context understanding. We further enhance performance by developing hybrid architectures that combine Gated DeltaNet layers with sliding window attention or Mamba2 layers, achieving both improved training efficiency and superior task performance.
Code: https://github.com/NVlabs/GatedDeltaNet

1 INTRODUCTION

The Transformer architecture has significantly advanced the capabilities of Large Language Models (LLMs), showcasing exceptional performance across a wide range of tasks due to its effective attention mechanism. This mechanism excels in precise sequence modeling and leverages the parallel processing capabilities of modern GPUs during training. However, the self-attention component scales quadratically with sequence length, leading to substantial computational demands that pose challenges for both training and inference.

To mitigate these issues, researchers have explored alternatives such as linear Transformers (Katharopoulos et al., 2020a), which replace traditional softmax-based attention with kernelized dot-product-based linear attention, substantially reducing memory requirements during inference by reframing as a linear RNN with matrix-valued states. While early versions of linear Transformers underperformed in language modeling tasks compared to standard Transformers, recent enhancements—such as incorporating data-dependent gating mechanisms akin to those in LSTMs, exemplified by models like GLA (Yang et al., 2024a) and Mamba2 (Dao & Gu, 2024a)—have shown promising improvements. However, challenges persist in managing information over long sequences, particularly for in-context retrieval tasks where traditional Transformers maintain their advantage (Arora et al., 2023a; 2024a; Jelassi et al., 2024; Wen et al., 2024; Akyürek et al., 2024).

This phenomenon is not surprising: linear Transformers can be interpreted as implementing an outer-product-based key-value association memory, reminiscent of tensor product representation (Smolensky, 1990). However, the number of orthogonal key-value pairs they can store is *bounded* by the model's dimensionality. When the sequence length exceeds this dimension, "memory collisions" become inevitable, hindering exact retrieval (Schlag et al., 2021a).

Mamba2 addresses this limitation by introducing a simple gated update rule, $\mathbf{S}_t = \alpha_t \mathbf{S}_{t-1} + \boldsymbol{v}_t \boldsymbol{k}_t^\intercal$, which uniformly decays all key-value associations at each time step by a dynamic ratio, $\alpha_t \in (0, 1)$. However, this approach does not account for the varying importance of different key-value

*Equation contribution. Work done during SY's internship at NVIDIA.

associations, potentially leading to inefficient memory utilization. If the model needs to forget a specific key-value association, all key-value associations are equally forgotten, making the process less targeted and efficient.

In contrast, the linear Transformer with the delta rule (Widrow et al., 1960), known as DeltaNet (Schlag et al., 2021a; Yang et al., 2024b), selectively updates memory by (softly) replacing an old key-value pair with the incoming one in a sequential manner. This method has demonstrated impressive performance in synthetic benchmarks for in-context retrieval. However, since this process only modifies a single key-value pair at a time, the model lacks the ability to rapidly clear outdated or irrelevant information, especially during context switches where previous data needs to be erased. Consequently, DeltaNet has been found to perform moderately on real-world tasks (Yang et al., 2024b), likely due to the absence of a robust memory-clearing mechanism.

Recognizing the complementary advantages of the gated update rule and the delta rule in memory management, we propose the *gated delta rule*, a simple and intuitive mechanism that combines both approaches. This unified rule enables flexible memory control: it can promptly clear memory by setting $\alpha_t \to 0$, while selectively updating specific content without affecting other information by setting $\alpha_t \to 1$ (effectively switching to the pure delta rule).

The remaining challenge lies in implementing the gated delta rule in a hardware-efficient manner. Building upon Yang et al. (2024b)'s efficient algorithm that parallelizes the delta rule computation using the WY representation (Bischof & Loan, 1985), we carefully extend their approach to incorporate the gating terms. Our extension preserves the benefits of chunkwise parallelism (Hua et al., 2022b; Sun et al., 2023a; Yang et al., 2024a;b), enabling hardware-efficient training.

Our resulting architecture, Gated DeltaNet, consistently outperforms both Mamba2 and DeltaNet across a comprehensive suite of benchmarks, including language modeling, commonsense reasoning, in-context retrieval, length extrapolation, and long-context understanding. Building on these results, we also develop hybrid architectures that strategically combine Gated DeltaNet layers with sliding window attention or Mamba2 layers, further enhancing both training efficiency and model performance.

2 Preliminary

2.1 Mamba2: Linear Attention with decay

It is known that the linear transformer (Katharopoulos et al., 2020b) can be formulated as the following linear recurrence when excluding normalization and query/key activations:

$$\mathbf{S}_t = \mathbf{S}_{t-1} + \boldsymbol{v}_t \boldsymbol{k}_t^\intercal \in \mathbb{R}^{d_v \times d_k}, \qquad \boldsymbol{o}_t = \mathbf{S}_t \boldsymbol{q}_t \in \mathbb{R}^{d_v}$$

where d_k and d_v represent the (head) dimensions for query/key and value, respectively. By expanding the recurrence, we can express it in both vector form (left) and matrix form (right) as follows:

$$\boldsymbol{o}_t = \sum_{i=1}^{t} (\boldsymbol{v}_i \boldsymbol{k}_i^\intercal) \boldsymbol{q}_t = \sum_{i=1}^{t} \boldsymbol{v}_i (\boldsymbol{k}_i^\intercal \boldsymbol{q}_t) \in \mathbb{R}^{d_v}, \qquad \mathbf{O} = (\mathbf{Q}\mathbf{K}^\intercal \odot \mathbf{M})\mathbf{V} \in \mathbb{R}^{L \times d_v}$$

where L is the sequence length, and $\mathbf{M} \in \mathbb{R}^{L \times L}$ is the causal mask defined by $\mathbf{M}_{ij} = 0$ when $i < j$, and 1 otherwise.

However, this vanilla linear attention underperforms Transformers in language modeling by a large margin. To address this, it is common to add a decay term to forget historical information. Here we take Mamba2 (Dao & Gu, 2024a) as an example, which can be represented by the following linear recurrence (up to specific parameterization):

$$\mathbf{S}_t = \alpha_t \mathbf{S}_{t-1} + \boldsymbol{v}_t \boldsymbol{k}_t^\intercal, \qquad \boldsymbol{o}_t = \mathbf{S}_t \boldsymbol{q}_t$$

where $\alpha_t \in (0, 1)$ is a data-dependent scalar-valued decay term that varies with t. Define the cumulative decay product $\gamma_j = \prod_{i=1}^{j} \alpha_i$, and by expanding the recurrence, we can express the result in both a vector form (left) and a matrix parallel form (right):

$$\boldsymbol{o}_t = \sum_{i=1}^{t} \left(\frac{\gamma_t}{\gamma_i} \boldsymbol{v}_i \boldsymbol{k}_i^\intercal \right) \boldsymbol{q}_t = \sum_{i=1}^{t} \boldsymbol{v}_i \left(\frac{\gamma_t}{\gamma_i} \boldsymbol{k}_i^\intercal \boldsymbol{q}_t \right), \qquad \mathbf{O} = ((\mathbf{Q}\mathbf{K}^\intercal) \odot \Gamma)\, \mathbf{V}$$

Here, $\Gamma \in \mathbb{R}^{L\times L}$ is a decay-aware causal mask where $\Gamma_{ij} = \frac{\gamma_i}{\gamma_j}$ if $i \geq j$ and $\Gamma_{ij} = 0$ otherwise. The equivalence between these parallel and recurrent forms is also referred to as the state space duality (SSD) described in Dao & Gu (2024a). This recurrence structure appears in several other architectures including Gated RFA (Peng et al., 2021), xLSTM (Beck et al., 2024), and Gated RetNet (Sun et al., 2024b). When γ_t is data-independent, the formulation reduces to RetNet (Sun et al., 2023a) and Lightning-Attention (Qin et al., 2024a). Furthermore, if γ_t is extended to be matrix-valued rather than scalar-valued, efficient training algorithms remain possible when parameterized with an outer-product structure, as demonstrated by Yang et al. (2024a) and used by Yang et al. (2024a); Peng et al. (2024); Qin et al. (2024b); Zhang et al. (2024); Chou et al. (2024); He et al. (2025); Lu et al. (2025).

Chunkwise training However, both the recurrent and parallel forms are not ideal for efficient training (Hua et al., 2022b; Yang et al., 2024a), which motivates the use of the chunkwise parallel form (Hua et al., 2022b; Sun et al., 2023a) for hardware-efficient, linear-time training, as introduced below. To summarize, the chunkwise parallel form splits inputs and outputs into several chunks of size C, and computes outputs for each chunk based on the final state of the previous chunk and the query/key/value blocks of the current chunk. Following the notation of Sun et al. (2023b); Yang et al. (2024a;b), we take the query block, $\boldsymbol{q}$, as an example. We denote $\mathbf{Q}_{[t]} := \boldsymbol{q}_{tC+1:(t+1)C+1}$ as the query block for chunk t, and $\boldsymbol{q}^r_{[t]} := \boldsymbol{q}_{tC+r}$ as the r-th query within chunk t. The initial state of chunk t is defined as $\mathbf{S}_{[t]} := \mathbf{S}^0_{[t]} = \mathbf{S}^C_{[t-1]}$. By partially expanding the recurrence, we have

$$\mathbf{S}^r_{[t]} = \mathbf{S}_{[t]} + \sum_{i=1}^{r} \boldsymbol{v}^i_{[t]} \boldsymbol{k}^{i\intercal}_{[t]} \in \mathbb{R}^{d_v\times d_k}, \qquad \boldsymbol{o}^r_{[t]} = \mathbf{S}^r_{[t]}\boldsymbol{q}^r_{[t]} = \mathbf{S}_{[t]}\boldsymbol{q}^r_{[t]} + \sum_{i=1}^{r} \boldsymbol{v}^i_{[t]} \left(\boldsymbol{k}^{i\intercal}_{[t]}\boldsymbol{q}^r_{[t]}\right) \in \mathbb{R}^{d_v}$$

Equivalently, in matrix form:

$$\mathbf{S}_{[t+1]} = \mathbf{S}_{[t]} + \mathbf{V}_{[t]}\mathbf{K}^\intercal_{[t]} \in \mathbb{R}^{d_v\times d_k}, \qquad \mathbf{O}_{[t]} = \mathbf{Q}_{[t]}\mathbf{S}^\intercal_{[t]} + \left(\mathbf{Q}_{[t]}\mathbf{K}^\intercal_{[t]} \odot \mathbf{M}\right)\mathbf{V}_{[t]} \in \mathbb{R}^{C\times d_v}$$

where $\mathbf{M} \in \mathbb{R}^{C\times C}$ is the causal mask. The above equations are rich in matrix multiplications (matmuls), allowing for tensor-core-based hardware optimization. This chunkwise algorithm could be easily extended to linear attention with decay:

$$\mathbf{S}_{[t+1]} = \overrightarrow{\mathbf{S}_{[t]}} + \mathbf{V}^\intercal_{[t]}\overrightarrow{\mathbf{K}_{[t]}} \in \mathbb{R}^{d_v\times d_k}, \quad \mathbf{O}_{[t]} = \overleftarrow{\mathbf{Q}_{[t]}}\mathbf{S}^\intercal_{[t]} + \left(\mathbf{Q}_{[t]}\mathbf{K}^\intercal_{[t]} \odot \Gamma_{[t]}\right)\mathbf{V}_{[t]} \in \mathbb{R}^{C\times d_v} \tag{1}$$

where $(\Gamma_{[t]})_{ij} = \frac{\gamma^i_{[t]}}{\gamma^j_{[t]}}, \gamma^j_{[t]} = \prod_{j=tC+1}^{tC+j} \alpha_j$. [1] Here we use the left arrow ($\overleftarrow{\cdot}$) or the right arrow ($\overrightarrow{\cdot}$) to denote a variable decaying to the first position and the last position of each chunk, respectively,

$\overleftarrow{\boldsymbol{q}^r_{[t]}} = \gamma^r_{[t]}\boldsymbol{q}^r_{[t]}$	decaying each vector to the first position of chunk t
$\overrightarrow{\boldsymbol{k}^r_{[t]}} = \frac{\gamma^C_{[t]}}{\gamma^r_{[t]}}\boldsymbol{k}^r_{[t]}$	decaying each vector to the last position of chunk t
$\overrightarrow{\mathbf{S}_{[t]}} = \gamma^C_{[t]}\mathbf{S}_{[t]}$	decaying the state matrix over the entire chunk t

and likewise for other variables (e.g., $\overrightarrow{\boldsymbol{v}}$). The SSD decomposition algorithm introduced in Mamba2 is largely equivalent to this chunkwise algorithm. For a more generalized approach, Yang et al. (2024a) proposed an extended chunkwise algorithm for linear attention that incorporates fine-grained decay mechanisms.

2.2 Delta Networks: Linear Attention with Delta Rule

The delta update rule (Widrow et al., 1960; Schlag et al., 2021b) *dynamically* erases the value ($\boldsymbol{v}^{\text{old}}_t$) associated with the current input key ($\boldsymbol{k}_t$) and writes a new value ($\boldsymbol{v}^{\text{new}}_t$), which is a linear combination of the current input value and the old value based on the "writing strength" $\beta_t \in (0, 1)$.[2]

$$\mathbf{S}_t = \mathbf{S}_{t-1} - \underbrace{(\mathbf{S}_{t-1}\boldsymbol{k}_t)}_{\boldsymbol{v}^{\text{old}}_t}\boldsymbol{k}^\intercal_t + \underbrace{(\beta_t\boldsymbol{v}_t + (1-\beta_t)\mathbf{S}_{t-1}\boldsymbol{k}_t))}_{\boldsymbol{v}^{\text{new}}_t}\boldsymbol{k}^\intercal_t = \mathbf{S}_{t-1}\left(\mathbf{I} - \beta_t\boldsymbol{k}_t\boldsymbol{k}^\intercal_t\right) + \beta_t\boldsymbol{v}_t\boldsymbol{k}^\intercal_t$$

[1] Here we slightly abuse the notation of γ to denote the cumulative product for each chunk (starting with the first position of each chunk separately) instead of the entire sequence.

[2] It is possible to set $\beta_t \in (0, 2)$ to allow negative eigenvalue to unlock the state tracking abilities of DeltaNet (Grazzi et al., 2024; Siems et al., 2025).

As shown above, DeltaNet implements a first-order linear recurrence with generalized Householder transition matrices ($\mathbf{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\intercal$). Despite demonstrating superior associative recall and language modeling performance (Schlag et al., 2021a), DeltaNet received limited attention due to computational inefficiency until Yang et al. (2024b) introduced a hardware-efficient chunkwise training algorithm, as detailed below.

Chunkwise parallel form. By partially expanding the recurrence, we have

$$\mathbf{S}_{[t]}^r = \mathbf{S}_{[t]} \underbrace{\left(\prod_{i=1}^{r} \mathbf{I} - \beta_{[t]}^i \boldsymbol{k}_{[t]}^i \boldsymbol{k}_{[t]}^{i\intercal}\right)}_{:=\mathbf{P}_{[t]}^r} + \underbrace{\sum_{i=1}^{r} \left(\beta_{[t]}^i \boldsymbol{v}_{[t]}^i \boldsymbol{k}_{[t]}^{i\intercal} \prod_{j=i+1}^{r} \left(\mathbf{I} - \beta_{[t]}^j \boldsymbol{k}_{[t]}^j \boldsymbol{k}_{[t]}^{j\intercal}\right)\right)}_{:=\mathbf{H}_t^r} \tag{2}$$

where $\mathbf{P}_{[t]}^j$ involves cumulative products of generalized Householder matrices, which could be optimized by the classical WY representation (Bischof & Loan, 1985):

$$\mathbf{P}_{[t]}^r = \mathbf{I} - \sum_{i=1}^{r} \mathbf{w}_{[t]}^i \boldsymbol{k}_{[t]}^{i\intercal} \in \mathbb{R}^{d_k \times d_k} \qquad \mathbf{w}_{[t]}^r = \beta_{[t]}^r \left(\boldsymbol{k}_{[t]}^r - \sum_{i=1}^{r-1} \left(\mathbf{w}_{[t]}^i (\boldsymbol{k}_{[t]}^{i\intercal} \boldsymbol{k}_{[t]}^r)\right)\right) \in \mathbb{R}^{d_k} \tag{3}$$

Likewise, $\mathbf{H}_{[t]}^r$ could be represented as:

$$\mathbf{H}_{[t]}^r = \sum_{i=1}^{r} \mathbf{u}_{[t]}^i \boldsymbol{k}_{[t]}^{i\intercal} \in \mathbb{R}^{d_v \times d_k} \qquad \mathbf{u}_{[t]}^r = \beta_{[t]}^r \left(\boldsymbol{v}_{[t]}^r - \sum_{i=1}^{r-1} \left(\mathbf{u}_{[t]}^i (\boldsymbol{k}_{[t]}^{i\intercal} \boldsymbol{k}_{[t]}^r)\right)\right) \in \mathbb{R}^{d_v} \tag{4}$$

and in matrix form: $\mathbf{P}_{[t]} = \mathbf{I} - \mathbf{W}_{[t]}^\top \mathbf{K}_{[t]} \in \mathbb{R}^{d_k \times d_k}$, $\mathbf{H}_{[t]} = \mathbf{U}_{[t]}^\top \mathbf{K}_{[t]} \in \mathbb{R}^{d_v \times d_k}$. By using the UT transform (Joffrain et al., 2006), we can further write $\mathbf{W}$ and $\mathbf{U}$ in matrix form:

$$\mathbf{T}_{[t]} = \left[\mathbf{I} + \text{strictLower}\left(\text{diag}(\beta_{[t]}) \mathbf{K}_{[t]} \mathbf{K}_{[t]}^\intercal\right)\right]^{-1} \text{diag}\left(\beta_{[t]}\right) \in \mathbb{R}^{C \times C} \tag{5}$$

$$\mathbf{W}_{[t]} = \mathbf{T}_{[t]} \mathbf{K}_{[t]} \in \mathbb{R}^{C \times d_k}, \qquad \mathbf{U}_{[t]} = \mathbf{T}_{[t]} \mathbf{V}_{[t]} \in \mathbb{R}^{C \times d_v} \tag{6}$$

Substituting these back into Eq. 2 yields a hardware-efficient chunkwise algorithm for DeltaNet that leverages matmuls, enabling tensor core based GPU optimization:

$$\mathbf{S}_{[t+1]} = \mathbf{S}_{[t]} \mathbf{P}_{[t]} + \mathbf{H}_{[t]} = \mathbf{S}_{[t]} + \left(\mathbf{U}_{[t]} - \mathbf{W}_{[t]} \mathbf{S}_{[t]}^\intercal\right)^\intercal \mathbf{K}_{[t]} \qquad \in \mathbb{R}^{d_v \times d_k} \tag{7}$$

$$\mathbf{O}_{[t]} = \mathbf{Q}_{[t]} \mathbf{S}_{[t]}^\intercal + (\mathbf{Q}_{[t]} \mathbf{K}_{[t]}^\intercal \odot \mathbf{M}) \left(\mathbf{U}_{[t]} - \mathbf{W}_{[t]} \mathbf{S}_{[t]}^\intercal\right) \qquad \in \mathbb{R}^{C \times d_v} \tag{8}$$

3 Gated Delta Networks

3.1 Formulation: Gated Delta Rule

The proposed gated delta rule is simple yet effective:

$$\mathbf{S}_t = \mathbf{S}_{t-1} \left(\alpha_t (\mathbf{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\intercal)\right) + \beta_t \boldsymbol{v}_t \boldsymbol{k}_t^\intercal \tag{9}$$

where the data-dependent gating term $\alpha_t \in (0, 1)$ controls state decay. This formulation unifies the advantages of both gating mechanisms and the delta rule: the gating term enables adaptive memory management, while the delta update structure facilitates effective key-value association learning.

We present a formal analysis of the gated delta rule through the lens of the online learning framework introduced by Liu et al. (2024). In this framework, recurrent state updates emerge as *closed-form* solutions to an online learning problem, as shown in Table 1. Recent linear RNN architectures typically incorporate a regularization term in their online learning objective to prevent state divergence from previous values, thereby enabling memory retention. However, this retention mechanism becomes problematic when the state becomes saturated with information. In such cases, each state would encode a superposition of multiple information pieces, making precise retrieval challenging. To address this limitation, Mamba2 and Gated DeltaNet introduce an adaptive scaling factor α_t that relaxes the regularization term, allowing controlled deviations between $\mathbf{S}_t$ and $\mathbf{S}_{t-1}$. This modification enables dynamic memory management through selective forgetting, which could be useful in filtering out irrelevant information (see §3.2).

Table 1: Comparison of different linear RNN models and their corresponding online learning objectives using the framework from Liu et al. (2024). For convenience, we simplify Longhorn's vector-valued $\boldsymbol{\beta}$ to scalar β.

Method	Online Learning Objective	Online Update
LA	$\|\mathbf{S}_t - \mathbf{S}_{t-1}\|_F^2 - 2\langle \mathbf{S}_t \boldsymbol{k}_t, \boldsymbol{v}_t \rangle$	$\mathbf{S}_t = \mathbf{S}_{t-1} + \boldsymbol{v}_t \boldsymbol{k}_t^T$
Mamba2	$\|\mathbf{S}_t - \alpha_t \mathbf{S}_{t-1}\|_F^2 - 2\langle \mathbf{S}_t \boldsymbol{k}_t, \boldsymbol{v}_t \rangle$	$\mathbf{S}_t = \alpha_t \mathbf{S}_{t-1} + \boldsymbol{v}_t \boldsymbol{k}_t^T$
Longhorn	$\|\mathbf{S}_t - \mathbf{S}_{t-1}\|_F^2 - \beta_t \|\mathbf{S}_t \boldsymbol{k}_t - \boldsymbol{v}_t\|^2$	$\mathbf{S}_t = \mathbf{S}_{t-1}(\mathbf{I} - \epsilon \boldsymbol{k}_t \boldsymbol{k}_t^T) + \epsilon_t \boldsymbol{v}_t \boldsymbol{k}_t^T, \epsilon_t = \dfrac{\beta_t}{1 + \beta_t \boldsymbol{k}_t^\top \boldsymbol{k}_t}$
DeltaNet	$\|\mathbf{S}_t - \mathbf{S}_{t-1}\|_F^2 - 2\langle \mathbf{S}_t \boldsymbol{k}_t, \beta_t (\boldsymbol{v}_t - \mathbf{S}_{t-1} \boldsymbol{k}_t) \rangle$	$\mathbf{S}_t = \mathbf{S}_{t-1}(\mathbf{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^T) + \beta_t \boldsymbol{v}_t \boldsymbol{k}_t^T$
Gated DeltaNet	$\|\mathbf{S}_t - \alpha_t \mathbf{S}_{t-1}\|_F^2 - 2\langle \mathbf{S}_t \boldsymbol{k}_t, \beta_t (\boldsymbol{v}_t - \alpha_t \mathbf{S}_{t-1} \boldsymbol{k}_t) \rangle$	$\mathbf{S}_t = \mathbf{S}_{t-1} \left(\alpha_t (\mathbf{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^T) \right) + \beta_t \boldsymbol{v}_t \boldsymbol{k}_t^T$

Table 2: Zero-shot performance comparison on S-NIAH benchmark suite for 1.3B models (see §4 for setups)

	S-NIAH-1 (pass-key retrieval)				S-NIAH-2 (number in haystack)				S-NIAH-3 (uuid in haystack)		
Model	1K	2K	4K	8K	1K	2K	4K	8K	1K	2K	4K
DeltaNet	97.4	96.8	**99.0**	**98.8**	98.4	45.6	18.6	14.4	85.2	47.0	22.4
Mamba2	**99.2**	**98.8**	65.4	30.4	99.4	98.8	56.2	17.0	64.4	47.6	4.6
Gated DeltaNet	98.4	88.4	91.4	91.8	**100.0**	**99.8**	**92.2**	**29.6**	**86.6**	**84.2**	**27.6**

On the other hand, Linear Attention (LA) and Mamba2 use a simple negative inner-product loss -$\langle \mathbf{S}_t \boldsymbol{k}_t, \boldsymbol{v}_t \rangle$, while Longhorn (Liu et al., 2024) uses a more expressive online regression objective $\|\mathbf{S}_t \boldsymbol{k}_t - \boldsymbol{v}_t\|^2$ for better modeling of key-value associations. The resulting Longhorn's update rule closely resembles the delta update rule, [3] suggesting the superiority of the (gated) delta rule over Mamba2 in in-context associative recall.

From the perspective of fast weight programming (Irie et al., 2022a) and test-time training (Sun et al., 2024a) and regression (Wang et al., 2025), the hidden state $\mathbf{S}$ can be interpreted as a (fast) weight matrix, with the delta rule optimizing the online regression objective $\mathcal{L}(\mathbf{S}_t) = \frac{1}{2}\|\mathbf{S}_t \boldsymbol{k}_t - \boldsymbol{v}_t\|^2$ via *test-time* stochastic gradient descent (SGD):

$$\mathbf{S}_{t+1} = \mathbf{S}_t - \beta_t \nabla \mathcal{L}(\mathbf{S}_t) = \mathbf{S}_t - \beta_t (\mathbf{S}_t \boldsymbol{k}_t - \boldsymbol{v}_t) \boldsymbol{k}_t^\intercal = \mathbf{S}_t \left(\mathbf{I} - \beta_t \boldsymbol{k}_t \boldsymbol{k}_t^\intercal \right) + \beta_t \boldsymbol{v}_t \boldsymbol{k}_t^\intercal$$

where β_t represents the (adaptive) learning rate. From this perspective, the gated delta rule can be viewed as incorporating an adaptive weight decay term α_t into the SGD update, a technique widely used in deep learning (Krogh & Hertz, 1991; Andriushchenko et al., 2023). Concurrently, Titans (Behrouz et al., 2024) demonstrated the effectiveness of incorporating weight decay mechanisms in RNN test-time SGD updates.

3.2 Case Study: Single Needle in a Haystack (S-NIAH)

To better understand the complementary strength between the delta rule and the gated rule, we present a case study on the Single Needle-In-A-Haystack (S-NIAH) benchmark suite from RULER (Hsieh et al., 2024), where a key-value pair acts as a needle in the haystack (context) and the model must recall the value when given the key. Table 2 presents the results and we draw three main observations:

Decay hurts memory retention. In the simplest S-NIAH-1 setting with repeated synthetic context, models memorize minimal information, testing long-term retention. DeltaNet achieves near-perfect performance across all sequence lengths. Mamba2 degrades significantly beyond 2K sequences since its decay discards information radically, while Gated DeltaNet's degradation is less severe thanks to the use of delta rule.

Gating facilitates filtering. In S-NIAH-2/3 with real-world-essay context, models store all potentially relevant information, testing efficient memory management. With fixed state size, lack of clearance causes memory collision—information becomes superimposed and indistinguishable. DeltaNet's performance drops significantly at longer sequences due to poor memory clearance. Mamba2 and Gated DeltaNet maintain better performance through gating mechanisms that filter irrelevant information.

[3] The theoretical distinction lies in the optimization approach: Longhorn uses implicit online learning (Kulis & Bartlett, 2010) to derive closed-form globally optimal updates, while DeltaNet optimizes the same objective through one-step explicit gradient descent, as noted by Liu et al. (2024).

Delta rule helps memorization. In S-NIAH-3, values change from numbers to UUIDs, testing complex pattern memorization. Mamba2's performance drops quickly, while Gated DeltaNet performs better, verifying that the delta rule indeed has better memorization ability.

3.3 Algorithm: Hardware-efficient Chunkwise Training

In this subsection, we derive a hardware-efficient chunkwise algorithm for training Gated DeltaNet. By partially expanding the recurrence in Eq. 9, we have

$$\mathbf{S}_{[t]}^r = \mathbf{S}_{[t]} \underbrace{\left(\prod_{i=1}^{r} \alpha_{[t]}^i \left(\mathbf{I} - \beta_{[t]}^i \boldsymbol{k}_{[t]}^i \boldsymbol{k}_{[t]}^{i\intercal}\right)\right)}_{:=\mathbf{F}_{[t]}^r} + \underbrace{\sum_{i=1}^{r} \left(\beta_{[t]}^i \boldsymbol{v}_{[t]}^i \boldsymbol{k}_{[t]}^{i\intercal} \prod_{j=i+1}^{r} \alpha_{[t]}^j \left(\mathbf{I} - \beta_{[t]}^j \boldsymbol{k}_{[t]}^j \boldsymbol{k}_{[t]}^{j\intercal}\right)\right)}_{:=\mathbf{G}_{[t]}^r}$$

It is easy to see that $\mathbf{F}_{[t]}^r = \gamma_{[t]}^r \mathrm{P}_{[t]}^r = \overleftarrow{\mathrm{P}_{[t]}^r}$. As for $\mathbf{G}_{[t]}^r$, we adapt Eq. 4 as follows,

$$\mathbf{G}_{[t]}^r = \sum_{i=1}^{r} \frac{\gamma_{[t]}^r}{\gamma_{[t]}^i} \tilde{\mathbf{u}}_{[t]}^i \boldsymbol{k}_{[t]}^{i\intercal} \in \mathbb{R}^{d_v \times d_k} \qquad \tilde{\mathbf{u}}_{[t]}^r = \beta_{[t]}^r \left(\boldsymbol{v}_{[t]}^r - \sum_{i=1}^{r-1} \left(\tilde{\mathbf{u}}_{[t]}^i \left(\frac{\gamma_{[t]}^r}{\gamma_{[t]}^i} \boldsymbol{k}_{[t]}^{i\intercal} \boldsymbol{k}_{[t]}^r\right)\right)\right) \in \mathbb{R}^{d_v} \tag{10}$$

(see §B for a proof). By UT transform, we have the matrix form:

$$\widetilde{\mathbf{U}_{[t]}} = \left[\mathbf{I} + \mathrm{strictLower}\left(\mathrm{diag}\left(\beta_{[t]}\right)\left(\Gamma_{[t]} \odot \mathbf{K}_{[t]}\mathbf{K}_{[t]}^{\intercal}\right)\right)\right]^{-1} \mathrm{diag}\left(\beta_{[t]}\right) \mathbf{V}_{[t]} \qquad \in \mathbb{R}^{C \times d_v}$$

Similar to how Mamba2 extends linear attention (Eq. 1), we can adapt DeltaNet's chunkwise algorithm (Eq. 7-8) for Gated DeltaNet to enable hardware-efficient training as follows:

$$\mathbf{S}_{[t+1]} = \overrightarrow{\mathbf{S}_{[t]}} + \left(\widetilde{\mathbf{U}_{[t]}} - \overleftarrow{\mathbf{W}_{[t]}} \mathbf{S}_{[t]}^{\intercal}\right)^{\intercal} \overrightarrow{\mathbf{K}_{[t]}} \qquad \in \mathbb{R}^{d_v \times d_k} \tag{11}$$

$$\mathbf{O}_{[t]} = \overleftarrow{\mathbf{Q}_{[t]}} \mathbf{S}_{[t]}^{\intercal} + \left(\mathbf{Q}_{[t]}\mathbf{K}_{[t]}^{\intercal} \odot \mathbf{M}\right)\left(\widetilde{\mathbf{U}_{[t]}} - \overleftarrow{\mathbf{W}_{[t]}} \mathbf{S}_{[t]}^{\intercal}\right) \qquad \in \mathbb{R}^{C \times d_v} \tag{12}$$

3.4 Gated Delta Networks and Hybrid Models

Token mixer block. The basic Gated DeltaNet follows Llama's macro architecture, stacking token mixer layers with SwiGLU MLP layers, but replaces self-attention with gated delta rule token mixing. Fig. 1 (right) shows its block design. For the gated delta rule (Eq. 9), queries, keys and values $\{\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}\}$ are generated through linear projection, short convolution and SiLU, with L2 normalization applied to $\boldsymbol{q}, \boldsymbol{k}$ for training stability. α, β use linear projection only.[4] Following Sun et al. (2023a), the output is processed through normalization and gating before applying output projection.

Hybrid models. Linear transformers have limitations in modeling local shifts and comparisons, and their fixed state size makes it hard for retrieval tasks (Arora et al., 2024a). Following recent hybrid architectures like Griffin (De et al., 2024) and Samba (Ren et al., 2024), we combine linear recurrent layers with sliding window attention (SWA), resulting in GatedDeltaNet-H1. We also stack Mamba2, GatedDeltaNet and SWA, resulting in GatedDeltaNet-H2.

4 Experiments

Setup Our experiments encompass a comprehensive comparison of recent state-of-the-art architectures, including pure Transformer models, RNN-based approaches, and hybrid architectures. We evaluate against the following baselines: RetNet (Sun et al., 2023a), HGRN2 (Qin et al., 2024b), Mamba (Gu & Dao, 2023), Mamba2 (Dao & Gu, 2024b), Samba (Ren et al., 2024), and DeltaNet (Yang et al., 2024b). For fair comparison, all models are trained under identical conditions with 1.3B parameters on 100B tokens sampled from the FineWeb-Edu dataset (Penedo et al., 2024). We use the AdamW optimizer with a peak learning rate of 4e-4, weight decay of 0.1, and gradient clipping of 1.0. The learning rate follows a cosine annealing schedule with a 1B token warm-up period and batch size of 0.5M tokens. All models employ the Llama2 tokenizer with a vocabulary size of 32,000. For sequence modeling, we set the training length to 4K tokens, with Samba and our hybrid models using a sliding window size of 2K. See appendix for evaluation settings.

[4] We use Mamba2's parameterization for α but omit it for brevity.

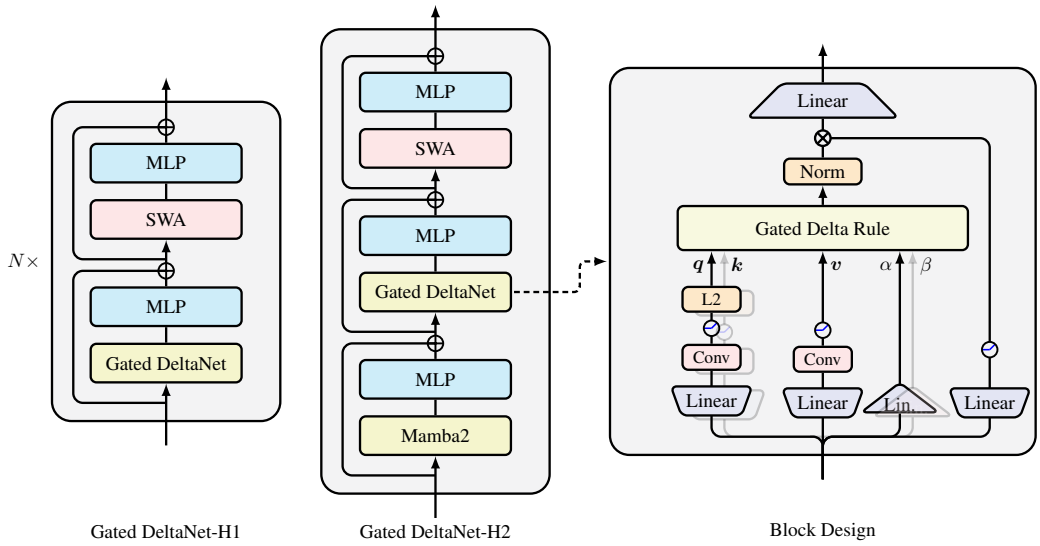

Figure 1: Visualization of the (hybrid) architecture and block design of Gated DeltaNet models. Gated DeltaNet-H1 and H2 use Gated DeltaNet + SWA and Mamba2 + Gated DeltaNet + SWA patterns, respectively. In the block design, query/key paths consist of linear proj., shortconv., SiLU and L2 norm; value path includes linear proj., shortconv. and SiLU; alpha/beta use linear proj.; and output gate applies linear proj. with SiLU.

Model	Wiki. ppl ↓	LMB. ppl ↓	LMB. acc ↑	PIQA acc ↑	Hella. acc_n ↑	Wino. acc ↑	ARC-e acc ↑	ARC-c acc_n ↑	SIQA acc ↑	BoolQ acc ↑	Avg.
Recurrent models											
RetNet	19.08	17.27	40.52	70.07	49.16	54.14	67.34	33.78	40.78	60.39	52.02
HGRN2	19.10	17.69	39.54	70.45	49.53	52.80	69.40	35.32	40.63	56.66	51.79
Mamba	17.92	15.06	43.98	71.32	52.91	52.95	69.52	35.40	37.76	61.13	53.12
Mamba2	16.56	12.56	45.66	71.87	55.67	55.24	**72.47**	37.88	40.20	60.13	54.89
DeltaNet	17.71	16.88	42.46	70.72	50.93	53.35	68.47	35.66	40.22	55.29	52.14
Gated DeltaNet	**16.42**	**12.17**	**46.65**	**72.25**	**55.76**	**57.45**	**71.21**	**38.39**	**40.63**	60.24	**55.32**
Attention or hybrid models											
Transformer++	18.53	18.32	42.60	70.02	50.23	53.51	68.83	35.10	40.66	57.09	52.25
Samba	16.13	13.29	44.94	70.94	53.42	55.56	68.81	36.17	39.96	62.11	54.00
Gated DeltaNet-H1	16.07	**12.12**	47.73	**72.57**	56.53	**58.40**	71.75	**40.10**	41.40	**63.21**	**56.40**
Gated DeltaNet-H2	**15.91**	12.55	**48.76**	72.19	**56.88**	57.77	71.33	39.07	**41.91**	61.55	56.18

Table 3: Performance comparison on language modeling and zero-shot common-sense reasoning.

Common-sense reasoning In Table 3, we present the language modeling perplexity and **zero-shot** accuracy on common-sense reasoning benchmarks for models with 400M and 1.3B parameters. Gated DeltaNet consistently outperforms other linear models, including RetNet, HGRN2, Mamba, Mamba2, and DeltaNet, at both scales. As expected, the hybrid variant further enhances performance.

Models	SWDE	SQD	FDA	TQA	NQ	Drop	Avg
Recurrent models							
RetNet	14.0	28.5	7.0	54.4	16.2	17.3	22.9
HGRN2	8.3	25.3	4.8	51.2	14.2	16.9	20.1
Mamba	9.8	25.8	3.7	54.3	14.9	17.4	21.0
Mamba2	19.1	33.6	**25.3**	**61.0**	**20.8**	19.2	29.8
DeltaNet	17.9	30.9	18.4	53.9	17.3	18.6	26.2
Gated DeltaNet	**25.4**	**34.8**	23.7	60.0	20.0	**19.8**	**30.6**
Attention or hybrid models							
Transformer++	29.5	38.0	**52.2**	58.3	22.5	21.6	37.0
Samba	33.0	39.2	50.5	57.7	23.5	20.2	37.3
Gated DeltaNet-H1	35.6	39.7	52.0	60.1	24.6	22.2	39.0
Gated DeltaNet-H2	**38.2**	**40.4**	50.7	**63.3**	**24.8**	**23.3**	**40.1**

Table 4: Accuracy on recall-world retrieval tasks with input truncated to 2K tokens. SQD: SQUADE. TQA: Trivial QA.

In-context retrieval on real-world data Table 4 presents results on real-world recall-intensive tasks used by Arora et al. (2024b). As expected, linear recurrent models show a significant performance gap compared to Transformers, while hybrid models combining linear recurrence and attention outperform pure attention models in retrieval tasks.

For pure recurrent models, despite DeltaNet's superior performance on synthetic in-context retrieval tasks (Yang et al., 2024b), its real-world retrieval performance lags behind Mamba2, consistent with our observations in S-NIAH-2 and S-NIAH-3 (Table 2). Gated DeltaNet outperforms both DeltaNet and Mamba2 thanks to its gated delta rule, though the improvement margin is smaller than in Table 2. We attribute this reduced performance gap to instruction-unaligned small language models being prone to repetition errors, which are the primary source of errors in these tasks (cf. Arora et al. (2024b, Appendix E)). Since this issue is largely independent of the update rule choice, the performance differences between models are less pronounced compared to Table 2.

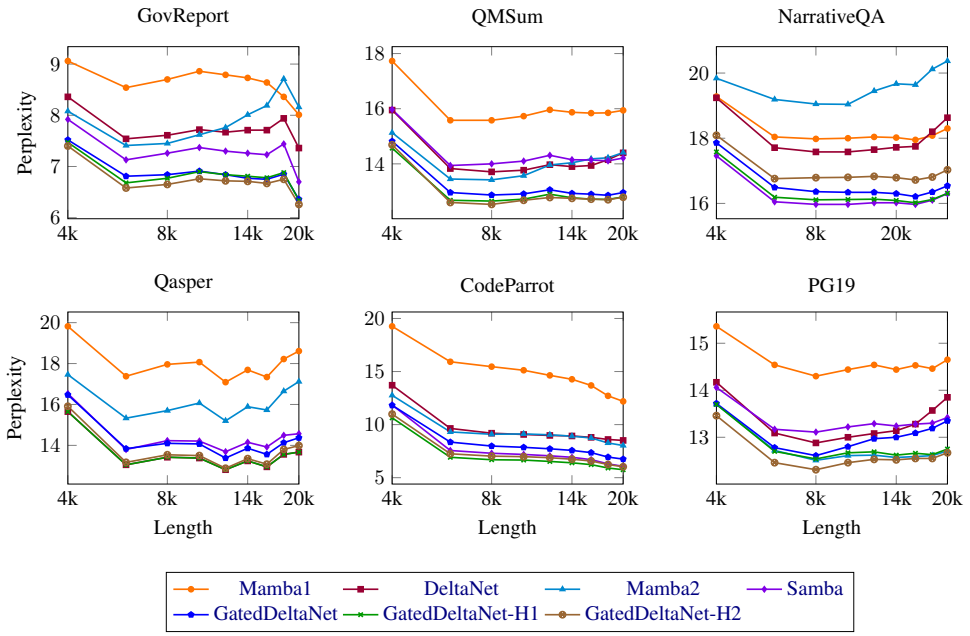

Figure 2: Length extrapolation on six long benchmarks.

Length extrapolation on long sequences. As shown in Fig.2, we evaluate the models' capacity to extrapolate to sequences of up to 20K tokens across six long-context benchmarks. Gated DeltaNet achieves the lowest overall perplexity across tasks among RNN models. While we observe mixed results in length extrapolation, Gated DeltaNet exhibits relatively more robust performance, suggesting better memory management. The hybrid models further improve upon this by leveraging attention for local context modeling, which reduces the memory management burden on their recurrent components. Future work will explore these models' capabilities on even longer sequences.

Long context understanding As demonstrated in Table 5, we evaluated the models' performance on LongBench (Bai et al., 2023). In recurrent models, Gated DeltaNet shows consistent advantages, especially in single-doc QA, few-shot in-context learning, and Code tasks, demonstrating its superior capabilities in retrieval, in-context learning, and state tracking, respectively.

Throughput Comparison. The training throughput comparison across different models is presented in Fig. 3. As our analysis shows, the proposed gated delta rule introduces only marginal overhead compared to the original delta rule, with Gated DeltaNet achieving essentially the same throughput as DeltaNet. Both are slightly slower than Mamba2 (2-3K tokens/sec) due to their more expressive transition matrices.

The Transformer++ achieves the best performance in the 2K context window domain, thanks to the highly optimized Flash-Attention-2 kernel (Dao, 2023). Consequently, hybrid approaches combining 2K window-size SWA attention with other token mixers demonstrate higher throughput than standalone mixers: Samba outperforms Mamba, while Gated DeltaNet-H1 and -H2 outperform Gated DeltaNet. Notably, Gated DeltaNet-H1 maintains compelling training throughput across all sequence lengths, even on short sequences.

Model	Single-Doc QA			Multi-Doc QA			Summarization			Few-shot			Code		Avg
	NQA	QQA	MFQ	HQA	2WM	Mus	GvR	QMS	MNs	TRC	TQA	SSM	LCC	RBP	
Recurrent models															
RetNet	12.1	10.7	19.1	10.7	**18.0**	5.8	4.8	15.8	7.9	19.0	18.0	12.8	14.1	17.9	13.2
HGRN2	10.7	12.1	19.1	11.3	15.7	6.0	5.2	15.1	**9.2**	16.0	15.8	10.3	18.6	20.8	13.5
Mamba	13.0	10.1	20.4	10.1	16.7	6.0	7.2	15.9	8.4	23.1	21.9	11.2	17.9	19.0	14.6
DeltaNet	12.9	10.8	21.5	10.9	13.2	5.1	6.5	13.5	7.2	15.5	23.3	11.6	17.6	20.3	13.6
Mamba2	11.1	11.3	18.6	11.8	15.1	**6.7**	6.7	14.5	7.4	13.0	**23.6**	8.4	17.9	20.6	13.5
Gated DeltaNet	**14.1**	**14.0**	**23.3**	**13.7**	14.4	5.8	**7.5**	**16.4**	7.9	**30.0**	22.4	**23.0**	**18.7**	**22.1**	**16.6**
Attention or hyrbid models															
Transformer++	11.8	9.3	10.0	10.9	4.2	6.1	7.4	15.8	6.6	16.9	13.5	3.9	17.2	18.7	11.0
Samba	12.5	12.9	25.4	11.2	19.7	6.8	9.1	15.7	11.0	20.0	22.7	22.8	18.1	21.1	15.9
Gated DeltaNet-H1	**14.5**	12.3	26.6	12.6	**23.6**	6.1	9.1	16.1	12.8	33.5	**23.9**	26.8	15.5	19.2	17.8
Gated DeltaNet-H2	12.7	**13.0**	**27.1**	**12.7**	20.6	**7.5**	**10.4**	**16.2**	**13.0**	**40.5**	22.7	**27.9**	**19.9**	**22.1**	**18.4**

Table 5: Accuracy on 14 tasks from LongBench (Bai et al., 2023): Narrative QA, QasperQA, MultiField QA, HotpotQA, 2WikiMulti QA, Musique, GovReport, QMSum, MultiNews, TRec, Trivia QA, SamSum, LCC, and RepoBench-P by order.

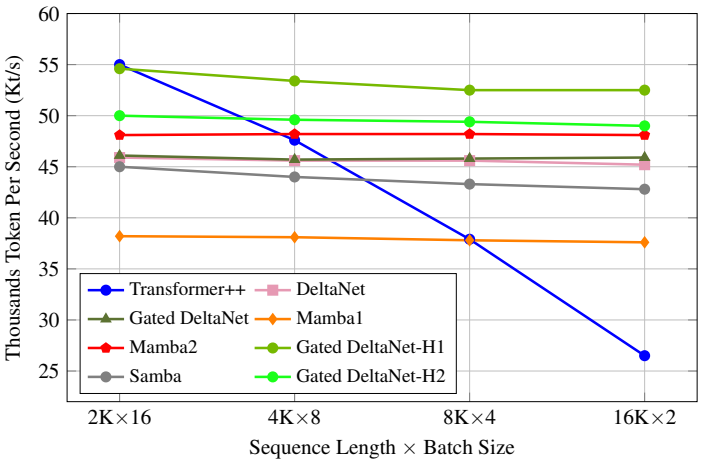

Figure 3: Training throughput comparison of 1.3B models on a single H100 GPU.

5 Related Work

Gated linear RNN. Large linear recurrent language models have attracted significant attention due to their training and inference efficiency. The field of linear RNNs has rapidly evolved from using data-independent decay mechanisms, as exemplified by models like S4 (Gu et al., 2022), S5 (Smith et al., 2023), LRU (Orvieto et al., 2023), RWKV4/5 (Peng et al., 2023), and RetNet (Sun et al., 2023a), to incorporating data-dependent decay mechanisms in more recent architectures such as HGRN1/2 (Qin et al., 2024b; 2023b), Mamba1/2 Gu & Dao (2023); Dao & Gu (2024a), RWKV6 (Peng et al., 2024), GSA (Zhang et al., 2024). This transition stems from the proven advantages of gating/forgetting mechanisms (termed selective mechanisms in Mamba)—a classical concept originating in the gated RNN literature (Gers et al., 2000) whose significance has been consistently reaffirmed (Greff et al., 2015; van der Westhuizen & Lasenby, 2018; Qin et al., 2024b; 2023b; Gu & Dao, 2023).

Modern forget gates differ from traditional designs like those in LSTM by removing the dependency on the previous hidden state, relying solely on input data. This modification enables efficient parallelism across sequence lengths (Martin & Cundy, 2018; Qin et al., 2023b). The absence of a forget gate has been a notable limitation in DeltaNet, and our gated extension addresses this gap in a natural, effective, and hardware-efficient way. We also note a recent concurrent work RWKV-7 [5] using a similar idea, but with a more relaxable formalism using diagonal-plus-low-rank transitions:

[5] https://github.com/BlinkDL/RWKV-LM/tree/main/RWKV-v7

$\mathbf{S}_t = \mathbf{S}_{t-1}(\mathrm{diag}(\mathbf{d}_t) - \mathbf{a}_t\mathbf{b}_t^\top) + \boldsymbol{v}_t\boldsymbol{k}_t^\top$ where $\mathbf{d}_t, \mathbf{a}_t, \mathbf{b}_t \in \mathbb{R}^{d_k}$. The chunkwise algorithm could be similarly adapted to this case, as implemented in Flash Linear Attention (Yang & Zhang, 2024). [6]

Delta rule. The delta learning rule demonstrates superior memory capacity compared to Hebbian learning (Gardner, 1988; Prados & Kak, 1989), an advantage DeltaNet leverages while linear transformers rely on Hebbian-like rules. This memory capacity advantage is evident in synthetic in-context learning tasks and extends to language modeling (Irie et al., 2021; Yang et al., 2024b), reinforcement learning (Irie et al., 2022b), and image generation (Irie & Schmidhuber, 2023). Yang et al. (2024b) parallelized delta rule computation and demonstrated how DeltaNet's data-dependent identity-plus-low-rank structure ($\mathbf{I} - \beta_t\boldsymbol{k}_t\boldsymbol{k}_t^\intercal$) offers greater flexibility than Mamba2's data-dependent diagonal matrices ($\alpha_t\mathbf{I}$). This structural advantage could enable complex reasoning, including regular language recognition (Fan et al., 2024; Grazzi et al., 2024) and state-tracking beyond TC^0 complexity (Merrill et al., 2024)—crucial for coding and reasoning applications.

Despite these significant advantages, the delta rule faces theoretical limitations (Irie et al., 2023) and shows only moderate performance on real-world datasets (Yang et al., 2024b), suggesting room for improvement. Previous attempts to enhance expressiveness through nonlinear recurrence (Irie et al., 2021; 2022b) addressed some limitations but sacrificed training parallelism, creating a performance-efficiency tradeoff. Recent work proposes some enhancements without compromising parallelism for better state tracking performance, including using negative eigenvalues (Grazzi et al., 2024) and multiple products of householder transition matrices (Siems et al., 2025) which enable high-rank transformations. These methods could be applied to Gated DeltaNet seamlessly.

From a (online) learning objective perspective, alternative formulations could further extend expressiveness: nonlinear regression ($\mathcal{L}(\mathbf{S}_t) = \frac{1}{2}||f_{\mathbf{S}_t}(\boldsymbol{k}_t) - \boldsymbol{v}_t||^2$) as in TTT (Sun et al., 2024a) and Titans (Behrouz et al., 2024), where $f_{\mathbf{S}}$ is a nonlinear function parameterized by $\mathbf{S}$; or regression considering the entire history ($\mathcal{L}(\mathbf{S}_t) = \frac{1}{2}\sum_{i=1}^{t}||\mathbf{S}_t\boldsymbol{k}_i - \boldsymbol{v}_i||^2$) as in Mesa layer (von Oswald et al., 2024)—analogous to the difference between Least Mean Square and Recursive Least Square algorithms. However, these more expressive variants introduce nonlinear recurrence and require workarounds, such as performing nonlinear updates only after processing entire chunks (as in TTT and Titans); or approximating nonlinear recurrence methods like Lim et al. (2024); Gonzalez et al. (2024); Schöne et al. (2025).

Hybrid models. In this work, we explore interleaving hybrid attention layers across layers, which is commonly used such as in MiniMax-01 (MiniMax et al., 2025) and Hybrid Mamba2-Attention (Waleffe et al., 2024). It is also interesting to investigate hybrid linear/softmax attention within a single layer (Hua et al., 2022a; Zancato et al., 2024; Munkhdalai et al., 2024; Nunez et al., 2024; Dong et al., 2025; Zhang et al., 2025).

6 Conclusion

In this work, we introduced Gated DeltaNet, which enables better key-value association learning compared to Mamba2 and more adaptive memory clearance than DeltaNet, leading to consistently better empirical results across various tasks. We extended the parallel algorithm from Yang et al. (2024b) to enable hardware-efficient training of Gated DeltaNet. Our hybrid Gated DeltaNet model achieves even higher training throughput and overall performance, making it well-suited for practical deployment.

Acknowledgment

We thank Yu Zhang for assistance with figure creation and model evaluation; Kazuki Irie for providing valuable feedback on the draft; Simeng Sun and Zhixuan Lin for insightful discussions on long-sequence task evaluation settings; and Eric Alcaide and Volodymyr Kyrylov for their helpful discussions on the online learning perspective of DeltaNet.

[6] https://github.com/fla-org/flash-linear-attention/tree/main/fla/ops/generalized_delta_rule.

REFERENCES

Ekin Akyürek, Bailin Wang, Yoon Kim, and Jacob Andreas. In-Context Language Learning: Architectures and Algorithms, 2024. URL https://arxiv.org/abs/2401.12973.

Maksym Andriushchenko, Francesco D'Angelo, Aditya Vardhan Varre, and Nicolas Flammarion. Why do we need weight decay in modern deep learning? *ArXiv*, abs/2310.04415, 2023. URL https://api.semanticscholar.org/CorpusID:263829417.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. *ArXiv preprint*, abs/2312.04927, 2023a. URL https://arxiv.org/abs/2312.04927.

Simran Arora, Brandon Yang, Sabri Eyuboglu, Avanika Narayan, Andrew Hojel, Immanuel Trummer, and Christopher Ré. Language Models Enable Simple Systems for Generating Structured Views of Heterogeneous Data Lakes, 2023b. URL https://arxiv.org/abs/2304.09433.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, Dylan Zinsley, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff. *ArXiv preprint*, abs/2402.18668, 2024a. URL https://arxiv.org/abs/2402.18668.

Simran Arora, Aman Timalsina, Aaryan Singhal, Benjamin Spector, Sabri Eyuboglu, Xinyi Zhao, Ashish Rao, Atri Rudra, and Christopher Ré. Just read twice: closing the recall gap for recurrent language models, 2024b. URL https://arxiv.org/abs/2407.05483.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *ArXiv preprint*, abs/2308.14508, 2023. URL https://arxiv.org/abs/2308.14508.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *ArXiv preprint*, abs/2405.04517, 2024. URL https://arxiv.org/abs/2405.04517.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time, 2024. URL https://arxiv.org/abs/2501.00663.

Christian H. Bischof and Charles Van Loan. The WY representation for products of householder matrices. In *SIAM Conference on Parallel Processing for Scientific Computing*, 1985. URL https://api.semanticscholar.org/CorpusID:36094006.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 7432–7439. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6239.

Yuhong Chou, Man Yao, Kexin Wang, Yuqi Pan, Rui-Jie Zhu, Jibin Wu, Yiran Zhong, Yu Qiao, Bo XU, and Guoqi Li. MetaLA: Unified optimal linear approximation to softmax attention map. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Y8YVCOMEpz.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv preprint*, abs/1803.05457, 2018. URL https://arxiv.org/abs/1803.05457.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *ArXiv preprint*, abs/2307.08691, 2023. URL https://arxiv.org/abs/2307.08691.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv: 2405.21060*, 2024a.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071. PMLR, 2024b. URL https://proceedings.mlr.press/v235/dao24a.html.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models, 2024. URL https://arxiv.org/abs/2402.19427.

Xin Dong, Yonggan Fu, Shizhe Diao, Wonmin Byeon, ZIJIA CHEN, Ameya Sunil Mahabaleshwarkar, Shih-Yang Liu, Matthijs Van keirsbilck, Min-Hung Chen, Yoshi Suhara, Yingyan Celine Lin, Jan Kautz, and Pavlo Molchanov. Hymba: A hybrid-head architecture for small language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=A1ztozypga.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1074–1084, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1102. URL https://aclanthology.org/P19-1102.

Ting-Han Fan, Ta-Chung Chi, and Alexander Rudnicky. Advancing regular language reasoning in linear recurrent neural networks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 45–53, Mexico City, Mexico, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-short.4.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 2021.

E. Gardner. The space of interactions in neural network models. *Journal of Physics A*, 21:257–270, 1988. URL https://api.semanticscholar.org/CorpusID:15378089.

Felix A. Gers, Jürgen Schmidhuber, and Fred A. Cummins. Learning to forget: Continual prediction with LSTM. *Neural Comput.*, 12(10):2451–2471, 2000.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu (eds.), *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL https://aclanthology.org/D19-5409.

Xavier Gonzalez, Andrew Warrington, Jimmy T.H. Smith, and Scott Linderman. Towards scalable and stable parallelization of nonlinear RNNs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=hBCxxVQDBw.

Riccardo Grazzi, Julien N. Siems, Jorg K. H. Franke, Arber Zela, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear rnns through negative eigenvalues. 2024. URL https://api.semanticscholar.org/CorpusID:274141450.

Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28: 2222–2232, 2015. URL https://api.semanticscholar.org/CorpusID:3356463.

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

Daya Guo, Canwen Xu, Nan Duan, Jian Yin, and Julian J. McAuley. Longcoder: A long-range pre-trained language model for code completion. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12098–12107. PMLR, 2023. URL https://proceedings.mlr.press/v202/guo23j.html.

Zhihao He, Hang Yu, Zi Gong, Shizhan Liu, Jianguo Li, and Weiyao Lin. Rodimus*: Breaking the accuracy-efficiency trade-off with efficient attentions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IIVYiJ1ggK.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *ArXiv preprint*, abs/2404.06654, 2024. URL https://arxiv.org/abs/2404.06654.

Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc V. Le. Transformer quality in linear time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9099–9117. PMLR, 2022a. URL https://proceedings.mlr.press/v162/hua22a.html.

Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc V. Le. Transformer quality in linear time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9099–9117. PMLR, 2022b. URL https://proceedings.mlr.press/v162/hua22a.html.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. Efficient attentions for long document summarization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1419–1436, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.112. URL https://aclanthology.org/2021.naacl-main.112.

Kazuki Irie and Jürgen Schmidhuber. Images as weight matrices: Sequential image generation through synaptic learning rules. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=ddad0PNUvV.

Kazuki Irie, Imanol Schlag, Róbert Csordás, and Jürgen Schmidhuber. Going beyond linear transformers with recurrent fast weight programmers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 7703–7717, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/3f9e3767ef3b10a0de4c256d7ef9805d-Abstract.html.

Kazuki Irie, Róbert Csordás, and Jürgen Schmidhuber. The dual form of neural networks revisited: Connecting test time predictions to training patterns via spotlights of attention. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9639–9659. PMLR, 2022a. URL https://proceedings.mlr.press/v162/irie22a.html.

Kazuki Irie, Imanol Schlag, Róbert Csordás, and Jürgen Schmidhuber. A modern self-referential weight matrix that learns to modify itself. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9660–9677. PMLR, 2022b. URL https://proceedings.mlr.press/v162/irie22b.html.

Kazuki Irie, Róbert Csordás, and Jürgen Schmidhuber. Practical computational power of linear transformers and their recurrent and self-referential extensions. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9455–9465, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.588. URL https://aclanthology.org/2023.emnlp-main.588.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. Repeat After Me: Transformers are Better than State Space Models at Copying. *ArXiv preprint*, abs/2402.01032, 2024. URL https://arxiv.org/abs/2402.01032.

Thierry Joffrain, Tze Meng Low, Enrique S. Quintana-Ortí, Robert A. van de Geijn, and Field G. Van Zee. Accumulating householder transformations, revisited. *ACM Trans. Math. Softw.*, 32:169–179, 2006. URL https://api.semanticscholar.org/CorpusID:15723171.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, 2017a. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, 2017b. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165. PMLR, 2020a. URL http://proceedings.mlr.press/v119/katharopoulos20a.html.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165. PMLR, 2020b. URL http://proceedings.mlr.press/v119/katharopoulos20a.html.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023.

Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *Neural Information Processing Systems*, 1991. URL https://api.semanticscholar.org/CorpusID:10137788.

Brian Kulis and Peter L. Bartlett. Implicit online learning. In Johannes Fürnkranz and Thorsten Joachims (eds.), *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pp. 575–582. Omnipress, 2010. URL https://icml.cc/Conferences/2010/papers/429.pdf.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.

Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://aclanthology.org/C02-1150.

Yi Heng Lim, Qi Zhu, Joshua Selfridge, and Muhammad Firmansyah Kasim. Parallelizing non-linear sequential models over the sequence length, 2024. URL https://arxiv.org/abs/2309.12252.

Bo Liu, Rui Wang, Lemeng Wu, Yihao Feng, Peter Stone, and Qiang Liu. @articleDBLP:journals/corr/abs-2407-14207, author = Bo Liu and Rui Wang and Lemeng Wu and Yihao Feng and Peter Stone and Qiang Liu, title = Longhorn: State Space Models are Amortized Online Learners, journal = CoRR, volume = abs/2407.14207, year = 2024, url = https://doi.org/10.48550/arXiv.2407.14207, doi = 10.48550/ARXIV.2407.14207, eprinttype = arXiv, eprint = 2407.14207, timestamp = Fri, 23 Aug 2024 08:12:16 +0200, biburl = https://dblp.org/rec/journals/corr/abs-2407-14207.bib, bibsource = dblp computer science bibliography, https://dblp.org : State space models are amortized online learners. *ArXiv preprint*, abs/2407.14207, 2024. URL https://arxiv.org/abs/2407.14207.

Tianyang Liu, Canwen Xu, and Julian McAuley. RepoBench: Benchmarking repository-level code auto-completion systems. *ArXiv preprint*, abs/2306.03091, 2023. URL https://arxiv.org/abs/2306.03091.

Colin Lockard, Prashant Shiralkar, and Xin Luna Dong. OpenCeres: When open information extraction meets the semi-structured web. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3047–3056, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1309. URL https://aclanthology.org/N19-1309.

Peng Lu, Ivan Kobyzev, Mehdi Rezagholizadeh, Boxing Chen, and Philippe Langlais. Regla: Refining gated linear attention, 2025. URL https://arxiv.org/abs/2502.01578.

Eric Martin and Chris Cundy. Parallelizing linear recurrent neural nets over sequence length. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=HyUNwulC-.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Byj72udxe.

William Merrill, Jackson Petty, and Ashish Sabharwal. The Illusion of State in State-Space Models, 2024. URL https://arxiv.org/abs/2404.08819.

MiniMax, Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, Enwei Jiao, Gengxin Li, Guojun Zhang, Haohai Sun, Houze Dong, Jiadai Zhu, Jiaqi Zhuang, Jiayuan Song, Jin Zhu, Jingtao Han, Jingyang Li, Junbin Xie, Junhao Xu, Junjie Yan, Kaishun Zhang, Kecheng Xiao, Kexi Kang, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Zheng, Linbo Chai, Long Xing, Meizhi Ju, Mingyuan Chi, Mozhi Zhang, Peikai Huang, Pengcheng Niu, Pengfei Li, Pengyu Zhao, Qi Yang, Qidi Xu, Qiexiang Wang, Qin Wang, Qiuhui Li, Ruitao Leng, Shengmin Shi, Shuqi Yu, Sichen Li, Songquan Zhu, Tao Huang, Tianrun Liang, Weigao Sun, Weixuan Sun, Weiyu Cheng, Wenkai Li, Xiangjun Song, Xiao Su, Xiaodong Han, Xinjie Zhang, Xinzhu Hou, Xu Min, Xun Zou, Xuyang Shen, Yan Gong, Yingjie Zhu, Yipeng Zhou, Yiran Zhong, Yongyi Hu, Yuanxiang Fan, Yue Yu, Yufeng Yang, Yuhao Li, Yunan Huang, Yunji Li, Yunpeng Huang, Yunzhi Xu, Yuxin Mao, Zehan Li, Zekang Li, Zewei Tao, Zewen Ying, Zhaoyang Cong, Zhen Qin, Zhenhua Fan, Zhihang Yu, Zhuo Jiang, and Zijia Wu. Minimax-01: Scaling foundation models with lightning attention, 2025. URL https://arxiv.org/abs/2501.08313.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention. *ArXiv preprint*, abs/2404.07143, 2024. URL https://arxiv.org/abs/2404.07143.

Elvis Nunez, Luca Zancato, Benjamin Bowman, Aditya Golatkar, Wei Xia, and Stefano Soatto. Expansion span: Combining fading memory and retrieval in hybrid state space models, 2024. URL https://arxiv.org/abs/2412.13328.

Antonio Orvieto, Samuel L. Smith, Albert Gu, Anushan Fernando, Çaglar Gülçehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26670–26698. PMLR, 2023. URL https://proceedings.mlr.press/v202/orvieto23a.html.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1144. URL https://aclanthology.org/P16-1144.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *ArXiv preprint*, abs/2406.17557, 2024. URL https://arxiv.org/abs/2406.17557.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. RWKV: Reinventing RNNs for the transformer era. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*,

pp. 14048–14077, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.936. URL https://aclanthology.org/2023.findings-emnlp.936.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Xingjian Du, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, Kranthi Kiran GV, Jan Kocoń, Bartłomiej Koptyra, Satyapriya Krishna, Ronald McClelland Jr., Niklas Muennighoff, Fares Obeid, Atsushi Saito, Guangyu Song, Haoqin Tu, Stanisław Woźniak, Ruichong Zhang, Bingchen Zhao, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui-Jie Zhu. Eagle and Finch: RWKV with Matrix-Valued States and Dynamic Recurrence, 2024. URL https://arxiv.org/abs/2404.05892.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah A. Smith, and Lingpeng Kong. Random feature attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=QtTKTdVrFBB.

DL Prados and SC Kak. Neural network capacity using delta rule. *Electronics Letters*, 3(25):197–199, 1989.

Zhen Qin, Dong Li, Weigao Sun, Weixuan Sun, Xuyang Shen, Xiaodong Han, Yunshen Wei, Baohong Lv, Fei Yuan, Xiao Luo, Y. Qiao, and Yiran Zhong. Transnormerllm: A faster and better large language model with improved transnormer. 2023a. URL https://api.semanticscholar.org/CorpusID:260203124.

Zhen Qin, Songlin Yang, and Yiran Zhong. Hierarchically gated recurrent neural network for sequence modeling. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b. URL http://papers.nips.cc/paper_files/paper/2023/hash/694be3548697e9cc8999d45e8d16fe1e-Abstract-Conference.html.

Zhen Qin, Weigao Sun, Dong Li, Xuyang Shen, Weixuan Sun, and Yiran Zhong. Lightning attention-2: A free lunch for handling unlimited sequence lengths in large language models. 2024a.

Zhen Qin, Songlin Yang, Weixuan Sun, Xuyang Shen, Dong Li, Weigao Sun, and Yiran Zhong. Hgrn2: Gated linear rnns with state expansion. *ArXiv preprint*, abs/2404.07904, 2024b. URL https://arxiv.org/abs/2404.07904.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124.

Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. Samba: Simple hybrid state space models for efficient unlimited context language modeling. *ArXiv preprint*, abs/2406.07522, 2024. URL https://arxiv.org/abs/2406.07522.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 8732–8740. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6399.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9355–9366. PMLR, 2021a. URL http://proceedings.mlr.press/v139/schlag21a.html.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9355–9366. PMLR, 2021b. URL http://proceedings.mlr.press/v139/schlag21a.html.

Mark Schöne, Babak Rahmani, Heiner Kremer, Fabian Falck, Hitesh Ballani, and Jannes Gladrow. Implicit language models are rnns: Balancing parallelization and expressivity, 2025. URL https://arxiv.org/abs/2502.07827.

Julien Siems, Timur Carstensen, Arber Zela, Frank Hutter, Massimiliano Pontil, and Riccardo Grazzi. Deltaproduct: Increasing the expressivity of deltanet through products of householders, 2025. URL https://arxiv.org/abs/2502.10297.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=Ai8Hw3AXqks.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artif. Intell.*, 46(1-2):159–216, 1990. doi: 10.1016/0004-3702(90)90007-M. URL https://doi.org/10.1016/0004-3702(90)90007-M.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): Rnns with expressive hidden states. *ArXiv preprint*, abs/2407.04620, 2024a. URL https://arxiv.org/abs/2407.04620.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *ArXiv preprint*, abs/2307.08621, 2023a. URL https://arxiv.org/abs/2307.08621.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *ArXiv preprint*, abs/2307.08621, 2023b. URL https://arxiv.org/abs/2307.08621.

Yutao Sun, Li Dong, Yi Zhu, Shaohan Huang, Wenhui Wang, Shuming Ma, Quanlu Zhang, Jianyong Wang, and Furu Wei. You only cache once: Decoder-decoder architectures for language models. *ArXiv preprint*, abs/2405.05254, 2024b. URL https://arxiv.org/abs/2405.05254.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31.

Jos van der Westhuizen and Joan Lasenby. The unreasonable effectiveness of the forget gate. *ArXiv preprint*, abs/1804.04849, 2018. URL https://arxiv.org/abs/1804.04849.

Johannes von Oswald, Maximilian Schlegel, Alexander Meulemans, Seijin Kobayashi, Eyvind Niklasson, Nicolas Zucchet, Nino Scherrer, Nolan Miller, Mark Sandler, Blaise Agüera y Arcas, Max Vladymyrov, Razvan Pascanu, and João Sacramento. Uncovering mesa-optimization algorithms in transformers, 2024. URL https://arxiv.org/abs/2309.05858.

Roger Waleffe, Wonmin Byeon, Duncan Riach, Brandon Norick, Vijay Korthikanti, Tri Dao, Albert Gu, Ali Hatamizadeh, Sudhakar Singh, Deepak Narayanan, Garvit Kulshreshtha, Vartika Singh, Jared Casper, Jan Kautz, Mohammad Shoeybi, and Bryan Catanzaro. An empirical study of mamba-based language models, 2024. URL https://arxiv.org/abs/2406.07887.

Ke Alexander Wang, Jiaxin Shi, and Emily B. Fox. Test-time regression: a unifying framework for designing sequence models with associative memory, 2025. URL https://arxiv.org/abs/2501.12352.

Kaiyue Wen, Xingyu Dang, and Kaifeng Lyu. RNNs are not Transformers (Yet): The Key Bottleneck on In-context Retrieval. *ArXiv preprint*, abs/2402.18510, 2024. URL https://arxiv.org/abs/2402.18510.

Bernard Widrow, Marcian E Hoff, et al. Adaptive switching circuits. In *IRE WESCON convention record*, volume 4, pp. 96–104. New York, 1960.

Songlin Yang and Yu Zhang. FLA: A Triton-Based Library for Hardware-Efficient Implementations of Linear Attention Mechanism, 2024. URL https://github.com/sustcsonglin/flash-linear-attention. original-date: 2023-12-20T06:50:18Z.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56501–56523. PMLR, 2024a. URL https://proceedings.mlr.press/v235/yang24ab.html.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *NeurIPS*, 2024b.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259.

Luca Zancato, Arjun Seshadri, Yonatan Dukler, Aditya Golatkar, Yantao Shen, Benjamin Bowman, Matthew Trager, Alessandro Achille, and Stefano Soatto. B'MOJO: Hybrid state space realizations of foundation models with eidetic and fading memory. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=RnQdRY1h5v.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Michael Zhang, Simran Arora, Rahul Chalamala, Benjamin Frederick Spector, Alan Wu, Krithik Ramesh, Aaryan Singhal, and Christopher Re. LoLCATs: On low-rank linearizing of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8VtGeyJyx9.

Yu Zhang, Songlin Yang, Ruijie Zhu, Yue Zhang, Leyang Cui, Yiqiao Wang, Bolun Wang, Freda Shi, Bailin Wang, Wei Bi, Peng Zhou, and Guohong Fu. Gated slot attention for efficient linear-time sequence modeling. 2024. URL https://api.semanticscholar.org/CorpusID:272593079.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. QMSum: A new benchmark for query-based multi-domain meeting summarization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5905–5921, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.472. URL https://aclanthology.org/2021.naacl-main.472.

Table 6: Ablation study on the Gated DeltaNet block. Avg-PPL and Avg-Acc denote average perplexity and zero-shot commonsense reasoning accuracy (as in Table 3), respectively. All models have 400M parameters and are trained for 15B tokens on the same subset of FineWeb-Edu dataset (Penedo et al., 2024).

Gated DeltaNet Ablations (400M)	Avg-PPL (↓)	Avg-Acc (↑)
Gated DeltaNet *w* Head Dim 128,	27.35	47.26
Macro Design		
w. naive Delta Rule	30.87	45.12
w/o. Short Conv	28.95	46.16
w/o. Output Gate	29.12	45.46
w/o. Output Norm	27.55	47.07
Normalization & Feature Map		
w. L_1-norm & ReLU	30.79	45.92
w. L_1-norm & 1+ELU	30.34	46.05
w. L_1-norm & SiLU	30.18	46.09
w. L_2-norm & ReLU	27.67	46.94
w. L_2-norm & 1+ELU	27.58	47.17
Model Dimensions		
w. Head Dim 64	28.31	46.35
w. Head Dim 256	27.13	47.38

A Ablation Study

Model	**Wiki.** ppl ↓	**LMB.** ppl ↓	**LMB.** acc ↑	**PIQA** acc ↑	**Hella.** acc_n ↑	**Wino.** acc ↑	**ARC-e** acc ↑	**ARC-c** acc_n ↑	**SIQA** acc ↑	**BoolQ** acc ↑	**Avg.**
Hybrid Ablations (500M/15B)											
Gated DeltaNet + SWA + Mamba2	24.02	28.20	34.77	67.08	40.84	50.74	60.35	28.83	38.94	61.49	47.88
Gated DeltaNet + Mamba2 + SWA	23.69	26.83	36.17	67.51	41.51	51.85	61.19	29.77	38.58	53.73	47.54
Mamba2 + SWA + Gated DeltaNet	24.14	25.21	36.79	64.96	41.18	52.01	60.90	30.03	38.07	59.44	47.92
Mamba2 + Gated DeltaNet + SWA	**23.54**	**24.11**	36.92	66.48	41.70	52.72	61.06	30.54	39.91	60.51	**48.73**

Table 7: Ablation studies of Gated DeltaNet models. All evaluations are performed by using `lm-evaluation-harness` (Gao et al., 2021). All models use the Llama tokenizer and are trained on the same subset of the FineWeb-Edu dataset (Penedo et al., 2024).

Table 6 presents ablation studies on the Gated DeltaNet block's components. Our experiments demonstrate that both the short convolution and output gate are crucial for model performance, while output normalization yields marginal improvements. Consistent with Yang et al. (2024b), we found L2 normalization to be essential for optimal performance, though the choice of feature map was less influential. Nevertheless, SiLU consistently outperformed other activation functions, aligning with observations from Qin et al. (2023a). Through empirical analysis, we determined that a head dimension of 128 provides an optimal trade-off between performance and computational efficiency. Additionally, Table 7 demonstrates that among various hybrid architectures, the combination of Mamba2, Gated DeltaNet, and SWA in this specific order produces superior results.

B Extended WY Representation for Gated Delta Rule

To reduce notation clutter, we only consider the first chunk here.

For $\mathbf{S}_t$, the extended WY representation is

$$\mathbf{S}_t = \sum_{i=1}^{t} \frac{\gamma_t}{\gamma_i} \mathbf{u}_i \boldsymbol{k}_i^\intercal, \qquad \mathbf{u}_t = \beta_t \left(\boldsymbol{v}_t - \sum_{i=1}^{t-1} \frac{\gamma_t}{\gamma_i} \mathbf{u}_i \boldsymbol{k}_i^T \boldsymbol{k}_t \right)$$

We proof this by mathmetical induction.

Proof.

$$
\begin{aligned}
\mathbf{S}_{t+1} &= \mathbf{S}_t \left(\alpha_{t+1}(\mathbf{I} - \beta_{t+1}\boldsymbol{k}_{t+1}\boldsymbol{k}_{t+1}^{\intercal})\right) + \beta_{t+1}\boldsymbol{v}_{t+1}\boldsymbol{k}_{t+1}^{\intercal} \\
&= \alpha_{t+1}(\sum_{i=1}^{t} \frac{\gamma_t}{\gamma_i}\mathbf{u}_i\boldsymbol{k}_i^{\intercal}) - \alpha_{t+1}\beta_{t+1}(\sum_{i=1}^{t} \frac{\gamma_t}{\gamma_i}\mathbf{u}_i\boldsymbol{k}_i^{\intercal}\boldsymbol{k}_i\boldsymbol{k}_{t+1}^{\intercal}) + \beta_{t+1}\boldsymbol{v}_{t+1}\boldsymbol{k}_{t+1}^{\intercal} \\
&= \sum_{i=1}^{t} \frac{\gamma_{t+1}}{\gamma_i}\mathbf{u}_i\boldsymbol{k}_i^{\intercal} + \underbrace{\beta_{t+1}\left(\boldsymbol{v}_{t+1} - \sum_{i=1}^{t} \frac{\gamma_{t+1}}{\gamma_i}\mathbf{u}_i\boldsymbol{k}_i^{T}\boldsymbol{k}_{t+1}\right)}_{\mathbf{u}_{t+1}}\boldsymbol{k}_{t+1}^{\intercal} \\
&= \sum_{i=1}^{t} \frac{\gamma_{t+1}}{\gamma_i}\mathbf{u}_i\boldsymbol{k}_i^{\intercal} + \underbrace{\frac{\gamma_{t+1}}{\gamma_{t+1}}}_{1}\mathbf{u}_{t+1}\boldsymbol{k}_{t+1}^{\intercal} \\
&= \sum_{i=1}^{t+1} \frac{\gamma_{t+1}}{\gamma_i}\mathbf{u}_i\boldsymbol{k}_i^{\intercal}
\end{aligned}
$$

□

C EXPERIMENTAL SETTINGS

C.1 EVALUATION

Commonsense reasoning Following Gu & Dao (2023), we evaluate our model on multiple commonsense reasoning benchmarks: PIQA (Bisk et al., 2020), HellaSwag (Hella.; Zellers et al., 2019), WinoGrande (Wino.; Sakaguchi et al., 2020), ARC-easy (ARC-e) and ARC-challenge (ARC-c) (Clark et al., 2018), SIQA (Sap et al., 2019), BoolQ (Clark et al., 2019), Wikitext (Wiki.; Merity et al., 2017), and LAMBADA (LMB.; Paperno et al., 2016).

In-context retrieval Our evaluation comprises both synthetic and real-world tasks. For synthetic tasks, we utilize the Needle-In-A-Haystack Single (NIAH-S) benchmark suite from RULER (Hsieh et al., 2024), which includes three increasingly complex tasks: S-NIAH-1 (passkey retrieval), S-NIAH-2 (numerical needle in haystack), and S-NIAH-3 (word-based needle in haystack). For real-world tasks, following Arora et al. (2024b), we evaluate on diverse datasets: SWDE (Lockard et al., 2019) for structured HTML relation extraction, FDA (Arora et al., 2023b) for PDF key-value retrieval, and several question-answering datasets including SQuAD (Rajpurkar et al., 2018), TriviaQA (Joshi et al., 2017a), Drop (Dua et al., 2019), and NQ (Kwiatkowski et al., 2019). Since our pretrained models lack instruction tuning, we employ the Cloze Completion Formatting prompts provided by Arora et al. (2024b), which better align with our models' next-word-prediction training objective.

Long context understanding We evaluate on 14 tasks from Longbench (Bai et al., 2023), encompassing: narrative comprehension (Narrative QA (Kočiský et al., 2018)), scientific understanding (QasperQA (Dasigi et al., 2021)), multi-hop reasoning (MultiField QA, HotpotQA (Yang et al., 2018), 2WikiMulti QA (Ho et al., 2020), Musique (Trivedi et al., 2022)), document summarization (GovReport (Huang et al., 2021), QMSum (Zhong et al., 2021), MultiNews (Fabbri et al., 2019)), and various specialized tasks (TRec (Li & Roth, 2002), Trivia QA (Joshi et al., 2017b), SamSum (Gliwa et al., 2019), LCC (Guo et al., 2023), and RepoBench-P (Liu et al., 2023)).