Mamba Modulation
On the Length Generalization of Mamba

Peng Lu[*]
Université de Montréal
peng.lu@umontreal.ca

Jerry Huang[*]
Université de Montréal & Mila - Quebec AI Institute
jerry.huang@mila.quebec

Qiuhao Zeng
Western University
qzeng53@uwo.ca

Xinyu Wang
McGill University
xinyu.wang5@mail.mcgill.ca

Boxing Chen
Noah's Ark Lab
boxing.chen@huawei.com

Philippe Langlais
Université de Montréal
felipe@iro.umontreal.ca

Yufei Cui[†]
Noah's Ark Lab
yufei.cui@huawei.com

Abstract

The quadratic complexity of the attention mechanism in Transformer models has motivated the development of alternative architectures with sub-quadratic scaling, such as state-space models. Among these, Mamba has emerged as a leading architecture, achieving state-of-the-art results across a range of language modeling tasks. However, Mambas performance significantly deteriorates when applied to contexts longer than those seen during pre-training, revealing a sharp sensitivity to context length extension. Through detailed analysis, we attribute this limitation to the out-of-distribution behavior of its state-space dynamics, particularly within the parameterization of the state transition matrix $\boldsymbol{A}$. Unlike recent works which attribute this sensitivity to the vanished accumulation of discretization time steps, $\exp(-\sum_{t=1}^{N}\Delta_t)$, we establish a connection between state convergence behavior as the input length approaches infinity and the spectrum of the transition matrix $\boldsymbol{A}$, offering a well-founded explanation of its role in length extension. Next, to overcome this challenge, we propose an approach that applies spectrum scaling to pre-trained Mamba models to enable robust long-context generalization by selectively modulating the spectrum of $\boldsymbol{A}$ matrices in each layer. We show that this can significantly improve performance in settings where simply modulating Δ_t fails, validating our insights and providing avenues for better length generalization of state-space models with structured transition matrices. Our code is available at https://github.com/gnepul-ace/mamba_modulation.

1 Introduction

In the new age of deep learning, the Transformers [129] architecture has spurred a new age of research into large language models (LLMs) [31, 141, 25, 92, 148] that has largely dominated the space of natural language processing (NLP) research since their introduction. Their surprising capabilities and rapid development have led to their wide application across various domains, including chatbots,

[*]Equal contribution.
[†]Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

https://doi.org/10.52202/085713-0653

intelligent agents, code assistants, etc. However, the Transformer comes with various deficiencies, which has led to research into alternative paradigms that seek to resolve these outstanding concerns. One of these competitors, `Mamba` [45, 22], is based off the state-space model (SSM) paradigm from control theory [47, 49] that have enabled the training of recurrent models that have overcome the sequential bottleneck of traditional models [109, 55, 15].

Among the various motivation for `Mamba` and its successors is the goal of length extrapolation, whereby a model that is initially trained on a limited context length (e.g. 2048 tokens in each training sequence) is capable of generalizing to longer sequences at test time (i.e. without further training) due to a more efficient inference-time token processing methodology. However, various works [56, 29, 57] have brought about challenges to this claim. Meanwhile, a key component in the Transformer is the position embedding, for which Rotary Position Embeddings (RoPE) [121] has been a popular choice and applied in many LLMs. Various works have studied RoPE [139, 83] and shown it to be intuitive to manipulate to extend the context window within Transformers [14, 11, 35, 96], whereas no equivalent method yet exists for `Mamba`-style models. A common explanation for the ability to extend this context length is through as avoiding out-of-distribution (OOD) rotation angles [83, 52] in RoPE, meaning the extended context length (OOD) can be mapped to the in-distribution (ID) context length on which models have been properly trained. However, `Mamba` does not utilize knowledge of token positions during training, thus making such methods broadly inapplicable.

Recent works [146, 9, 3] have meanwhile made attempts at exploring how to conduct length generalization with `Mamba` models. A shared feature among these is the focus on a specific input-dependent parameter, Δ, which is used to discretize the underlying state-space model and control for both the state decay over time as well as the incoming input contribution to the state. Their observations rely on the implicit notion that since the duration of the time-step influences the state, it can act as a proxy to filter out (or ignore) parts of the input, or be scaled to influence the long-term information decay within the model state. However, despite the compelling intuition behind such a notion, there remains no fundamental theoretical justification for this.

In this work, we attempt to build a better understanding of how to better scale `Mamba` models for improved length generalization. We begin with an analysis of the model and the implicit effects this will have on the convergence behavior of the hidden state as the input length goes to infinity. From this, we identify two ways in which this process can be controlled: either through the state-transition matrix $\boldsymbol{A}$, or through the discretization time-steps Δ. We then motivate why controlling for or adjusting $\boldsymbol{A}$ is more well-founded, through arguments relating the eigenvalue spectrum with long-term information decay. Through experiments on standard long-context extension settings, such as long-context language modeling and passkey retrieval, we demonstrate empirically how scaling $\boldsymbol{A}$ is more effective compared to scaling Δ, in the case of both `Mamba` and `Mamba2` models. Broadly, we summarize our contributions as follows:

I) We first provide a broader understanding of the length generalization ability of `Mamba`-based models via spectrum analysis of their transition matrix. We demonstrate and justify that the convergence behavior of the hidden states hinders their length generalization in `Mamba` models.

II) Based on our analysis, we identify how the scaling of $\boldsymbol{A}$ as opposed to the more common practice of scaling Δ is a more effective proposition.

III) Results on a series of long-context generalization tasks show such an intuition holds empirically on `Mamba` models, highlighting the potential benefits of using $\boldsymbol{A}$ for length generalization.

2 Related Works

2.1 Language Models and Long Contexts

Being capable of modeling long sequences is an important desiderate in various LLM applications. However, due to the quadratic complexity (relative to the sequence length) of the self-attention mechanism in Transformers, long sequence modeling requires a large computational overhead [125]. Early work in efficient Transformers [74, 147, 8, 131, 16] attempted to reduce the computational complexity of attention by inducing greater sparsity. Additional work has explored the use of linear attention [72, 86, 143, 144, 150, 87] to remove the softmax activation that induces this

quadratic complexity, however, such methods may impair performance on tasks that demand precise contextual recall [1, 133]. Furthermore, hardware optimizations for more efficient computation [22, 21, 114, 80] as well as inference-time acceleration methods [138] to reduce the computational and memory complexity of Transformers. However, a broader class of linear recurrent models [49, 45, 93, 7, 101, 102], which resemble traditional recurrent neural networks but provide an additional benefit of parallel training over the sequence elements, have emerged as an alternative for long sequences through a sub-quadratic complexity relative to sequence length as well as constant-time inference complexity.

2.2 Length Generalization and Extrapolation

Various restrictions on the data available for training make it difficult to directly collect data of extreme lengths (e.g., 100K+ tokens), hence there have been a great deal of efforts devoted to enabling models to generalize beyond the training length. However, various works have demonstrated the collapse of the performance [122, 83, 139], thus leaving this an open area of research. Based on the wide dominance of RoPE as the positional embedding of choice, many recent works have focused on extending the context window by scaling the rotary angles [11, 35, 96, 14] with potentially some additional tuning, enabling extension to sometimes up to $10\times$ the original training context length. Alternatively, linear recurrent models present promise through their lack of direct positional encoding; rather, a fixed-size hidden state is often utilized to maintain information from the past while the sequence is being processed. While some promise has been shown on synthetic tasks [1, 99], where these models have been shown to be able to filter out noise from the sequence while maintaining useful information within the state, these observations have not extended to tasks such as real-world long-context language tasks [56].

Yet because many existing methods relevant to Transformer length extrapolation rely explicitly on positional information, it remains an open work to find ways to enable such linear recurrent models to generalize beyond their training lengths. Alternatively, recent works [9, 146, 3] have investigated the post-hoc length extension in `Mamba` models, with a particular focus on using the discretization time-steps Δ_t for context extension. Ben-Kish et al. [9] use the value of these time-steps to 'decimate' or remove tokens from the processing of the sequence at specific layers, resulting in a shortened sequence length. Similarly, Ye et al. [146] use the value of the time-steps to filter out tokens. Azizi et al. [3] meanwhile calibrate scaling factors for these time-steps to adjust the long-term decay within the model, extending the context length. Unlike these works, we do not analyze the effect of discretization (Δ) on extrapolation ability. Instead, we focus on establishing a connection between the spectral characteristics of the state transition matrix and the asymptotic convergence behavior of the hidden state as the input length approaches infinity.

2.3 Spectrum Analysis of Linear Recurrent Models

Previous works have provided specific analysis of the eigenvalue spectrum of linear recurrent models as a way of understanding their state dynamics and the downstream influence this can have on performance. Gu et al. [46] initially provided an understanding of the specific parameterization of the state transition matrix in SSMs, determining the necessity of a Hurwitz matrix for effective sequence modeling. Orvieto et al. [93] further demonstrated how the eigenvalues have a specific influence on state decay as well as long-term dynamics during training. Beck et al. [7] further bound the state of the recurrence, implicitly bounding the spectrum as well. Finally, Grazzi et al. [44] also recently demonstrate the importance of negative eigenvalues for state-tracking tasks.

3 Background

3.1 State-Space Models (SSMs) and `Mamba`

The SSM-based models, i.e., structured state space sequence models (S4) [49] and `Mamba` [45] are inspired by the continuous system, which maps a 1-D function or sequence $\boldsymbol{x}(t) \in \mathbb{R}^{d_m}$ to an output $\boldsymbol{y}(t) \in \mathbb{R}^{d_m}$ through a hidden state $\boldsymbol{h}(t) \in \mathbb{R}^{d_h}$. The system uses evolution parameters $\boldsymbol{A} \in \mathbb{R}^{d_h \times d_h}$, $\boldsymbol{B} \in \mathbb{R}^{d_h \times d_m}$, and $\boldsymbol{C} \in \mathbb{R}^{d_m \times d_h}$, creating a continuous system whose dynamics are governed by

$$\boldsymbol{h}'(t) = \boldsymbol{A}\boldsymbol{h}(t) + \boldsymbol{B}x(t), \quad \boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{h}(t) \tag{1}$$

https://doi.org/10.52202/085713-0653

The Mamba model uses the selective SSM blocks, which leverage the input-dependent discretization into the recurrence computation. This is done by including an input-dependent timescale parameter $\boldsymbol{\Delta}(\boldsymbol{x}_t)$ to transform the continuous parameters $\boldsymbol{A}$, $\boldsymbol{B}$ to discrete parameters $\overline{\boldsymbol{A}}_t$ and $\overline{\boldsymbol{B}}_t$. We follow the official implementation of Mamba [45]:

$$\boldsymbol{h}_t = \overline{\boldsymbol{A}}_t \boldsymbol{h}_{t-1} + \overline{\boldsymbol{B}}_t \boldsymbol{x}_t, \quad \boldsymbol{y}_t = \boldsymbol{C}_t \boldsymbol{h}_t, \tag{2}$$

This method uses a Zero-Order Hold (ZOH) for the matrix $\overline{\boldsymbol{A}}_t$ and a simplified Euler discretization for the matrix $\overline{\boldsymbol{B}}_t$, omiting the computation of matrix inversion for $\overline{\boldsymbol{B}}$ as required by the ZOH:

$$\overline{\boldsymbol{A}}_t = \exp\left(-\Delta_t \odot \boldsymbol{A}\right), \quad \overline{\boldsymbol{B}}_t = \Delta_t \otimes \left(\left(\Delta_t \boldsymbol{A}\right)^{-1} \left(\exp\left(\Delta_t \boldsymbol{A}\right) - \boldsymbol{I}\right) \boldsymbol{B}\right), \tag{3}$$

The key improvement of Mamba is making the parameterization ($\overline{\boldsymbol{A}}_t$, $\overline{\boldsymbol{B}}_t$ and $\boldsymbol{C}_t$) input-dependent. Specifically, each part of them can be computed as follows:

$$\boldsymbol{\Delta}_t = \mathsf{softplus}\left(\mathsf{Linear}_{\Delta}\left(\boldsymbol{x}_t\right)\right), \quad \boldsymbol{B}_t = \mathsf{Linear}_{B}\left(\boldsymbol{x}_t\right), \quad \boldsymbol{C}_t = \mathsf{Linear}_{C}\left(\boldsymbol{x}_t\right) \tag{4}$$

where Linear_{Δ}, Linear_{B} and Linear_{C} are regular linear projections, $\odot$ is the Hadamard product, $\otimes$ is the outer product, $\boldsymbol{\Delta}_t \in \mathbb{R}_+^d$ and $\boldsymbol{A} = \mathsf{diag}\left(\alpha_1, \ldots, \alpha_d\right)$ s.t. $\alpha_i > 0 \, \forall i \in \{1, \ldots, d\}$. In Mamba, Δ, $\boldsymbol{B}$ and $\boldsymbol{C}$ are input-dependent, such that at each time-step unique transition matrices can be used to update the system ($\boldsymbol{A}$ is left as a fixed parameter in as the dynamics of the state should be consistent across steps). This is based on the observation that some elements in a discrete sequence may not be as important as others, therefore there is an incentive to possibly update the system differently based on this factor. This results in unique update matrices at each time-step $\left(\Delta_t, \overline{\boldsymbol{A}}_t, \overline{\boldsymbol{B}}_t, \boldsymbol{C}_t\right)$, enabling the ability to solve problems that require selective processing of the sequence. In order to maintain computational efficiency $\boldsymbol{A}$ is restricted to having a diagonal structure such that only the diagonal elements of these matrices need to be stored. Mamba2 [22] further restricts the diagonal matrix to have the form of a scalar-times-identity matrix, enabling further computational improvements.

3.2 Limitations of Mamba in Long Context

The output can be reformulated as a matrix product form as follows:

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_L \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}_1 \bar{\boldsymbol{B}}_1 & 0 & \cdots & 0 \\ \boldsymbol{C}_2 \bar{\boldsymbol{A}}_2 \bar{\boldsymbol{B}}_1 & \boldsymbol{C}_2 \bar{\boldsymbol{B}}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \boldsymbol{C}_L \prod_{t=2}^{L} \bar{\boldsymbol{A}}_t \bar{\boldsymbol{B}}_1 & \boldsymbol{C}_L \prod_{t=3}^{L} \bar{\boldsymbol{A}}_t \bar{\boldsymbol{B}}_2 & \cdots & \boldsymbol{C}_L \bar{\boldsymbol{B}}_L \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_L \end{bmatrix} = \boldsymbol{M}\boldsymbol{X} \tag{5}$$

In this formulation, each output $\boldsymbol{y}_t$ is computed as a weighted sum of all inputs, with each weight involving a product of transition matrices $\prod_{t=j+1}^{L} \bar{\boldsymbol{A}}_t$. This product term plays a crucial role in determining the influence of past states and can be further disentangled to enable fine-grained analysis, facilitating a deeper understanding of state evolution and transition behavior.

$$\prod_{t=j+1}^{L} \bar{\boldsymbol{A}}_t = \prod_{t=j+1}^{L} \exp\left(-\boldsymbol{A}\Delta_t\right) = \exp\left(-\boldsymbol{A} \sum_{t=j+1}^{L} \Delta_t\right) \tag{6}$$

Previous work [9, 146, 3] has primarily focused on analyzing the discretization step Δ_t for long context, particularly the vanishing effect of the accumulated term $\exp\left(-\sum_{t=1}^{N} \Delta_t\right)$ when N is large and propose different solutions to overcome this out-of-distribution (OOD) issue. For instance, Azizi et al. [3] propose applying scalar values $s \leq 1$ across different model layers to mitigate OOD discretization steps, ensuring smaller Δ_t to prevent the vanishing issue of distant inputs. They introduce two calibration methods and demonstrate superior length generalization performance in calibrated Mamba models with unconstrained scaling factors. However, their work does not explain why some of resulted scaling factors $s > 1$ could still enhance generalization performance.

4 Spectrum-Based Analysis of Mambas Length Generalization

In this section, we examine the length generalization ability of Mamba-based language models from the perspective of spectrum analysis of their transition matrix. Specifically, we analyze the state convergence behavior of the hidden state in Mamba. Based on our findings, we propose a spectrum scaling method to enhance the length generalization capability of pre-trained Mamba models.

4.1 Spectrum of `Mamba` Transition Matrix

We first visualize the spectrum of the continuous transition matrix $\mathbf{\Lambda} = \text{diag}\left(\exp\left(-\boldsymbol{A}\right)\right)$ of `Mamba` models. The $\exp\left(-\boldsymbol{A}\right)$ parameterization guarantees all values $\lambda \in (0, 1)$. We stack all 48 layers and rank eigenvalues in descending order for each row. The magnitudes (appearing to range from 0 to 1) imply that all eigenvalues λ lie well inside the unit circle, which is critical for the stability of the dynamics governed by the transition matrix for `Mamba` training. High eigenvalue zones can be viewed as dominant temporal modes, useful for modeling long-term dependencies, especially in language or time series tasks. We also observe that low-eigenvalue regions in the transition matrix spectrum correspond to rapidly decaying modes, which specialize in modeling local dependencies and high-frequency dynamics.

The heatmap in Figure 1 reveals a consistent spectral pattern across layers: the coexistence of both large (near 1) and small (near 0) eigenvalues. Next, we establish the connection between the failure of length generalization and the spectrum of the transition matrix by showing a divergent tendency in the convergence of the state norm.

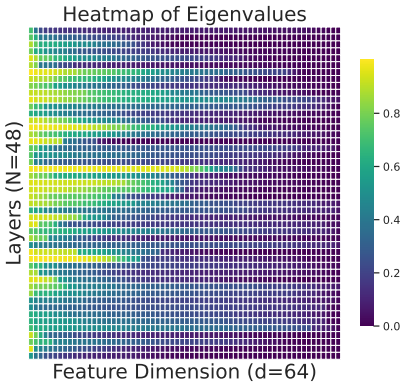

Figure 1: Heatmap of eigenvalues of $\text{diag}(\exp(-\boldsymbol{A}))$ of `mamba2-1.3b`.

4.2 State Convergence in SSMs for Long Contexts

The previous section presented the numerical spectrum of the transition matrix $\exp(-\boldsymbol{A})$. Next, we theoretically investigate its influence on the convergence behavior of `Mamba` states. We begin by introducing the following lemma, which establishes an expected bound on the norm of inputs.

Lemma 4.1. *Let $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ be a matrix with $\|\boldsymbol{B}\|_2 = \sigma_B$, and let $\boldsymbol{x} \in \mathbb{R}^d$ be a vector such that each entry of $\boldsymbol{x}$ satisfies $|x_i| \le \sigma_x$. The upper bound for $\|\boldsymbol{B}\boldsymbol{x}\|_2$ is:*

$$\|\boldsymbol{B}\boldsymbol{x}\|_2 \le \sigma_B \cdot \sigma_x \cdot \sqrt{d}. \tag{7}$$

Theorem 4.2. *(Convergence of State Norm with Real-Valued Diagonal Transition Matrix). Let the real-valued transition matrix $\mathbf{\Lambda} \in \mathbb{R}^{d\times d}$ be diagonal with eigenvalues $\lambda_i \sim \text{Uniform}[\lambda_{\min}, \lambda_{\max}]$, where $0 < \lambda_{\min} < \lambda_{\max} < 1$. Consider the system dynamics: $\boldsymbol{h}_t = \mathbf{\Lambda}\boldsymbol{h}_{t-1} + \boldsymbol{B}\boldsymbol{x}_t$ where $\boldsymbol{x}_t$ is the input vector at time step t, and $\boldsymbol{B}$ is a weight matrix whose rows are independently sampled as $\boldsymbol{b} \sim \mathcal{N}\left(0, \frac{1}{2d}\mathbf{I}\right)$. Then, as $t \to \infty$, the expected squared norm of the state $\boldsymbol{h}_t$ converges to:*

$$\mathbb{E}[\|\boldsymbol{h}_\infty\|^2] = \frac{1}{2(\lambda_{\max} - \lambda_{\min})} \log\left(\frac{1-\lambda_{\min}^2}{1-\lambda_{\max}^2}\right) \cdot \mathbb{E}[\|\boldsymbol{B}\boldsymbol{x}\|^2]. \tag{8}$$

Under the setting of Theorem 4.2 (proofs in Appendix A), we consider two cases (`Mamba` and `Mamba2`) corresponding to the architectural variants of structured state-space models, each characterized by a different form of the structured transition matrix.

Corollary 4.3 (Norm of `Mamba` State)**.** *Suppose the diagonal entries of $\mathbf{\Lambda}$ are independently drawn from a uniform distribution on $[0, \lambda]$, a moderate discretized step value Δ and the system evolves as $\boldsymbol{h}_t = \mathbf{\Lambda}\boldsymbol{h}_{t-1} + \overline{\boldsymbol{B}}\boldsymbol{x}_t = \text{diag}\left(\exp(-\Delta\alpha)\right)\boldsymbol{h}_{t-1} + \Delta\boldsymbol{B}\boldsymbol{x}_t$. Then the growth rate ρ of the expected squared norm of the limiting state satisfies $\mathcal{O}\left(\frac{\Delta}{2\lambda}\log\left(\frac{1}{1-\lambda^2}\right)\right)$.*

Corollary 4.4 (Norm of `Mamba2` State)**.** *Suppose $\mathbf{\Lambda} = \lambda\boldsymbol{I} = \exp(-\Delta\alpha)\boldsymbol{I}$ is a scalar multiple of the identity matrix, where $\lambda \in (0, 1)$, a moderate discretized step value Δ and the system evolves as $\boldsymbol{h}_t = \mathbf{\Lambda}\boldsymbol{h}_{t-1} + \overline{\boldsymbol{B}}\boldsymbol{x}_t = \exp(-\Delta\alpha) \odot \boldsymbol{h}_{t-1} + \Delta\boldsymbol{B}\boldsymbol{x}_t$. Then the convergence rate ρ of the expected squared norm of the limiting state can be estimated as $\mathcal{O}(\frac{\Delta\cdot\lambda}{1-\lambda})$.*

These provide insight into the asymptotic convergence behavior of `Mamba` states as the input sequence length grows with different eigenvalues. If $\lambda \to 1$, or $\lambda \to 0$, then

$$\lim_{\lambda\to 1^-} \rho = \infty, \qquad \lim_{\lambda\to 0^+} \rho = 0 \tag{9}$$

https://doi.org/10.52202/085713-0653

These rates shed light on challenges in length generalization for structured state-space models (SSMs) with constrained diagonal transition matrices. In particular, both extremely large eigenvalues (approaching 1) and extremely small eigenvalues (approaching 0) can induce instability in the `Mamba` state norm as input length increasesleading to state explosion or vanishing, respectively. While tuning the discretization step $\boldsymbol{\Delta}$ can help modulate the convergence rate (as suggested by Corollary 4.3 and 4.4), it does not address the root cause: the distribution of the transition matrix eigenvalues. To directly tackle this issue, we propose a *spectrum scaling* method that adjusts the spectral distribution of a pre-trained `Mamba` model by compressing large eigenvalues and inflating small ones. This rescaling aims to stabilize the state norm across longer sequences, thereby improving the models ability to generalize over input length.

4.3 State Norm Analysis across different input lengths

ProofPile	1K	2K	4K	8K	16K	32K	64K
State norm (max)	131.4489	174.6176	892.7114	1330.5321	1141.5253	1130.6537	1157.5122
State norm (min)	0.0046	0.0030	0.0007	0.0007	0.0004	0.0003	0.0005
PG19	1K	2K	4K	8K	16K	32K	64K
State norm (max)	158.7656	160.8465	166.3788	876.0833	1241.3979	1201.5530	1242.7589
State norm (min)	0.0033	0.0014	0.0050	0.0004	0.0005	0.0003	0.0002

Table 1: State norm statistics for ProofPile and PG19 datasets across different sequence lengths.

We conducted a series of experiments to examine how the hidden state of the SSM evolves with increasing input sequence length, as shown in Table 1. Using randomly sampled data from ProofPile and PG19, we measured the maximum difference between the largest and smallest SSM state norms across all layers of the `mamba2-1.3b` model. The observed divergence in state norm magnitude as sequence length grows provides empirical validation for the theoretical predictions outlined in the previous section.

5 Mamba Modulation for Length Extrapolation

In the following sections, we describe a series of experiments that we conduct to validate our previous intuitions. Appendix C provides more specific implementation details and design choices.

5.1 A Simple Case Analysis on Constant Scaling

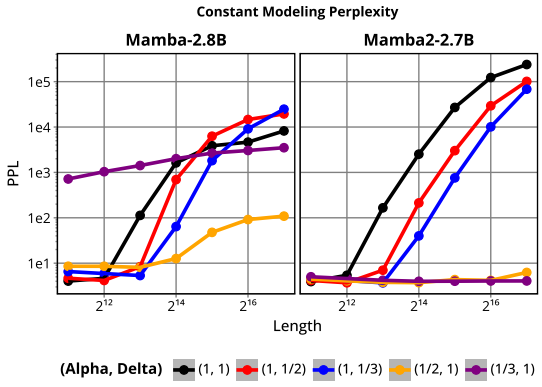

Figure 2: Language modeling perplexity on ProofPile after applying a constant scaling factor to either $\boldsymbol{A}$ or Δ_t. Lines are distinguished by their colors. $(1, 1)$ means the baseline where nothing is scaled.

To confirm our intuition, we first attempt a simple comparison between the effects of scaling Δ_t and $\boldsymbol{A}$. Here, across all layers, we use a fixed, constant-valued scaling factor. We evaluate language modeling perplexity on the ProofPile dataset [40], following Peng et al. [96], across a varying number of context lengths. This method uses no tuning or training; the scaling is applied explicitly during the forward pass. Figure 2 shows these results after applying a scaling on perplexity on context lengths from 2K to 128K tokens, with scaling factors of 2 or 3 applied.

We see that scaling $\boldsymbol{A}$ by a constant scaling factor is significantly better at incurring a lower perplexity, however, it remains the case that such a constant scaling factor needs to be properly tuned for, particularly in the case of `mamba-2.8b`. Given the simple setting/scenario on which we experiment, this is unsurprising; as we investigated in Section 4.2, different layers have different underlying behavior in terms of their eigenvalues, making it likely difficult

to find a constant scaling factor that can work across all layers. In the case of `mamba2-2.7b`, we can see that applying these scaling factors can significantly bound the long-context perplexity from exploding.

5.2 Adapting MambaExtend to Scale A

Given our observations and analysis regarding the relationship between $\boldsymbol{A}$ and Δ_t, a natural method against which we can compare is `MambaExtend` is a training-free method that scales the discretizations steps at each layer. For a model with L layers, the objective is to learn a set of constant scaling factors for each layer $\{\boldsymbol{s}_i\}_{i=1}^{L}$ which can be used to adjust the discretizations steps Δ_t. In general, $\boldsymbol{s}_i$ can be set to either a scalar or a vector. These scaling factors serve as learnable parameters in within the model but are consequentially tuned in a manner that does not require training any other parameters within the model. The idea of the algorithm is to take a pre-trained model along with a small set of samples for calibrating the scaling factors; depending on the setting, the calibration function can vary, with the only restriction being that the original model parameters are not modified during calibration. Appendix C.3 describes the implementation in further detail.

Algorithm 1 `MambaExtend` methodology.

Input: Model $\mathcal{M}$, calibration set $\mathcal{C}$ and function CF
Output: Scaling factors $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L] \in \mathbb{R}_{+}^{d_s \times L}$
for $i \leq L$ **do**
    $\boldsymbol{s}_i \leftarrow U(0, 1)$
end for
$\boldsymbol{S} \leftarrow \mathsf{CF}(\boldsymbol{S}, \mathcal{C}, \mathcal{M})$
return $\boldsymbol{S}$

Although the original `MambaExtend` learns scaling factors only for Δ_t, their methodology is adaptable to usage with $\boldsymbol{A}$ instead; given the shared dimensionality for both $\boldsymbol{A}$ and Δ_t, the scaling factors can be directly used for calibrating $\boldsymbol{A}$. Furthermore, this means that tuning scaling factors for $\boldsymbol{A}$ does not require any additional computation, time or memory requirements as compared to tuning them directly Δ_t, leading to a simple yet effective algorithm that can directly be applied to the adaptation of $\boldsymbol{A}$ for long-context generalization. The following sections evaluates the performance and efficiency of tuning these scaling factors for $\boldsymbol{A}$ on a number of standard settings for evaluating long-context generalization of models. As a baseline, we compare directly with `MambaExtend`.

6 Experiments and Results

6.1 Language Modeling Perplexity

We first experiment by measuring language modeling perplexity after calibrating scaling factors for either Δ_t or $\boldsymbol{A}$. In this task, we use the black-box zeroth-order calibration method suggested by Azizi et al. [3]; we train a single scaling factor $s_i \in \mathbb{R}_{+}$ for every layer i in the model. For a L-layer model, this means L individual scaling factors are used. To calibrate, 20 samples of the corresponding context length are used. For example, for a length of 16K, 20 samples of this length are used for the calibration of the set of s_i. Figure 3 shows these perplexity results on a number of validation datasets, namely ProofPile [40], PG19 [103] and GovReport [59].

In particular, scaling $\boldsymbol{A}$ leads to better perplexity on nearly all validation datasets, for both `Mamba` and `Mamba2` models. In many cases, this gap can be significant, particularly in the case of `mamba2-2.7b`, where the perplexity at long sequences when calibrating Δ_t explodes for all three datasets whereas calibrating $\boldsymbol{A}$ can lead the model to maintain a consistent perplexity up to $1000\times$ lower.

6.2 Passkey Retrieval

Next, we conduct experiments on the Passkey Retrieval task, also known as the Needle-in-A-Haystack. Similar to before, we conduct this to compare the effectiveness of tuning scaling factors for $\boldsymbol{A}$ as opposed to Δ_t; we again conduct this experiment across different `Mamba` models. Unlike the language modeling perplexity task however, we train the model on a training set. This training set contains samples of length 4096 corresponding to the task, where the objective is standard instruction-tuning [32]. However, we freeze all parameters except the scaling parameters for each layer. For `Mamba`, it is equivalent to the number of inner state dimensions, i.e. each inner state utilizes the same scaling factor for each dimension of the SSM state. For `Mamba2`, this is the number of heads, meaning that each head shares the same scaling factor for each component of its state.

https://doi.org/10.52202/085713-0653

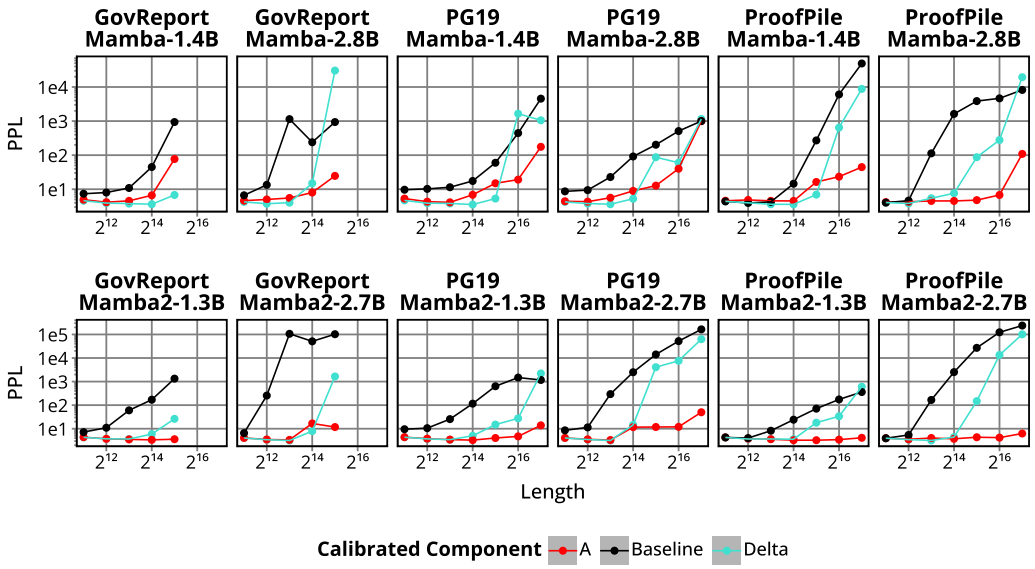

Figure 3: Model perplexity by calibrating scaling factors for either $\log(\boldsymbol{A})$ (red) or Δ_t (turquoise), across different datasets and sizes. `Baseline` means the base model with no calibration, i.e. the model is used directly without modification.

Evaluation is conducted on a set of fixed lengths and depths to evaluate for both generalization ability as well as potential biases to relative location within the sequence. The exact setup follows from Ben-Kish et al. [9], in particular, the task comprises of a 5-digit code embedded at a random sequence depth within samples from the WikiText-103 dataset [89]. Models are deemed to have solved length/depth pair if they can correctly solve all evaluation examples, i.e. retrieve the code within the example.

Figure 4 and Appendix C.3.2 visualize these results. In particular, we see very consistent results similar to our language modeling perplexity results; for `Mamba`, smaller models appear to fare slightly better when trained to scale Δ_t, but as the models get larger, learning to scale $\boldsymbol{A}$ closes the gap and eventually exceeds the performance of scaling Δ_t. Similarly, for `Mamba2` models, scaling $\boldsymbol{A}$ appears to nearly always be a more appropriate choice in comparison to scaling Δ_t, as seen by a nearly constant improvement in performance on the task. Further comparing with a full-fine-tuning of the model, we observe that scaling $\boldsymbol{A}$ is as effective despite fewer parameters being trained, whereas scaling Δ_t does observe a drop-off in performance.

6.3 LongBench

LongBench [6] is a popular benchmark for testing the long-context abilities of LLMs, serving as a more suitable real-world benchmark on which we can explore how the scaling of $\boldsymbol{A}$ as opposed to Δ_t can influence performance. Here, we again use the zeroth-order optimization method as we used for our initial perplexity experiments. More specifically, a constant scaling factor is used for each individual layer. We compare against both the initial base model, as well as `MambaExtend`. Table 2 shows results on `mamba2-2.7b`. In particular, we show that we can increase performance by over 6% through the calibrated scaling of $\boldsymbol{A}$, with a relative improvement of nearly 10% compared to if the scaling was instead calibrated for Δ_t.

Table 2: Results on LongBench [6].

Model	Strategy	Qasper	HotpotQA	2WMHQA	TREC	TriviaQA	LCC	RB-P	Average
`mamba2-2.7b`	Base Model	1.17	1.54	2.18	8.33	10.60	23.46	14.97	8.75
	`MambaExtend`	12.53	1.63	5.99	24.63	10.33	23.00	17.09	13.60
	Calibrated Scaling $\boldsymbol{A}$	12.90	5.69	11.18	24.32	10.49	23.36	16.91	14.98

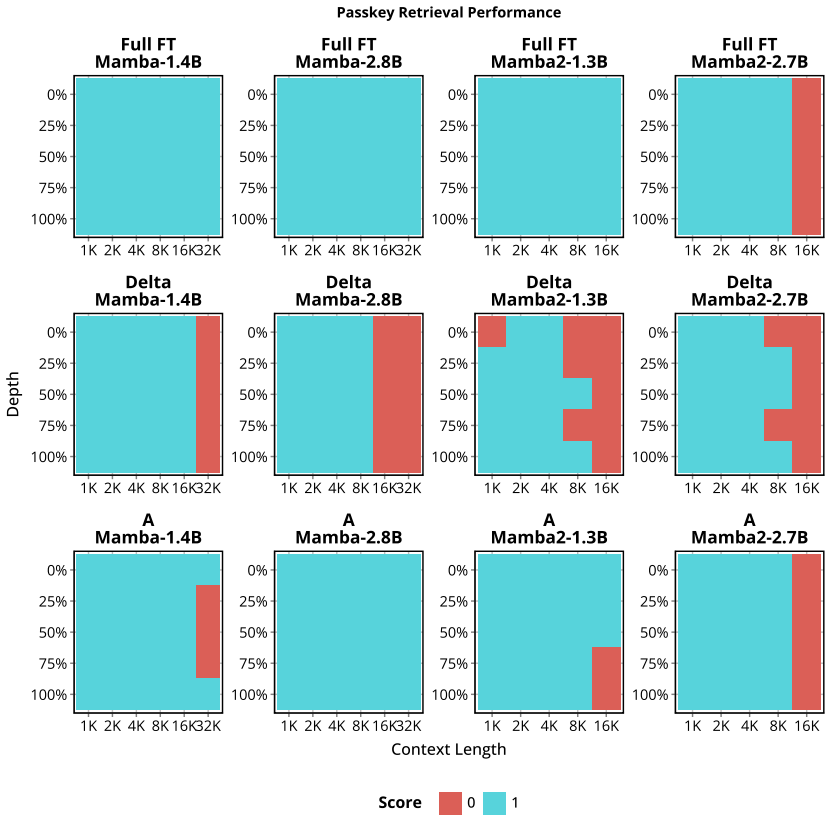

Figure 4: Passkey Retrieval performance of `Mamba` models by calibrating scaling factors for either $\log(\boldsymbol{A})$ or Δ_t. Turquoise squares mean that the model was able to solve all examples of the given evaluation length/depth pair after tuning scaling factors, while red squares means otherwise.

Furthermore, if looking more specifically at individual tasks, there are no settings where calibrating Δ_t results in a meaningful performance increase compared to $\boldsymbol{A}$, whereas calibrating $\boldsymbol{A}$ instead appears to significantly increase performance on HotpotQA and 2WikiMultihopQA.

6.4 Comparison with Alternative Methods

As a final point of comparison of our proposed methodology, we compare against other proposals that have aimed towards extending the context of `Mamba`. Unlike `MambaExtend`, both of these methods use a filtering mechanism rather than directly scale Δ_t; in `LongMamba` [146], channels are prevented from exponential decaying by filtering out tokens from the training sequence if the update of a specific token within the sequence Δ_t is smaller than a preset threshold. `DeciMamba` [9] instead defines *decimating layers* that directly filter out tokens that are then not passed to the following layer, significantly shortening the sequence that the last layers within the model observe. Both models require additional tuning; `LongMamba` calibrates multiple hyper-parameters to tune their filtering mechanism, while `DeciMamba` requires training the decimation layers on longer sequences.

For reasons of public code availability[3] and methodology[4], we compare `DeciMamba` against `mamba-1.4b/2.8b` and `LongMamba` against `mamba2-1.3b`. We also provide results using the initial base model, `MambaExtend`, as well as the previous two ways we tested for scaling $\boldsymbol{A}$, namely constant scaling as well as the calibrated scaling based on `MambaExtend`. Table 3 compares the effectiveness of these different methods on perplexity on the PG19 dataset. We note that in all cases, the calibrated scaling of $\boldsymbol{A}$ performs either the best or second best on all context lengths across the different tested models with marginal gaps when not the best performing method, while other

[3] `LongMamba` did not release their tuning code: `https://github.com/GATECH-EIC/LongMamba`

[4] `DeciMamba` only modified `Mamba` CUDA kernels: `https://github.com/assafbk/DeciMamba`

https://doi.org/10.52202/085713-0653

Table 3: Comparison of PG19 perplexity at varying lengths. Cases where scaling $\boldsymbol{A}$ leads to the lowest perplexity are **bolded** and underlined when second best. If the best method does not involve scaling $\boldsymbol{A}$, it is highlighted in violet.

Model	Context Length 2K	4K	8K	16K	32K	64K
`mamba-1.4b`						
Base Model	9.67	10.23	11.43	17.46	59.77	444.09
`DeciMamba`	11.45	12.34	14.65	19.83	24.85	28.48
`MambaExtend`	4.69	3.89	3.83	3.55	5.31	1648.0
Constant Scaling $\boldsymbol{A}$	44.68	53.46	59.56	63.51	75.86	114.44
Calibrated Scaling $\boldsymbol{A}$	5.31	4.31	4.13	6.88	14.94	**19.13**
`mamba-2.8b`						
Base Model	8.66	9.42	22.78	91.43	202.20	508.88
`DeciMamba`	11.34	13.45	15.63	18.34	21.53	26.54
`MambaExtend`	4.25	3.80	3.63	5.25	87.00	60.00
Constant Scaling $\boldsymbol{A}$	28.80	33.93	39.80	69.37	162.77	355.77
Calibrated Scaling $\boldsymbol{A}$	4.44	4.31	5.63	8.94	**12.75**	40.00
`mamba2-1.3b`						
Base Model	9.52	10.54	25.49	115.65	634.32	1479.45
`LongMamba`	10.12	10.31	11.36	11.61	12.81	13.55
`MambaExtend`	4.34	3.69	3.44	5.00	14.94	27.50
Constant Scaling $\boldsymbol{A}$	11.12	11.83	12.47	12.71	12.85	13.22
Calibrated Scaling $\boldsymbol{A}$	4.38	3.78	**3.44**	**3.28**	**4.03**	**4.72**

methods are fairly inconsistent on this front. Meanwhile, a constant scaling is generally ineffective, confirming previous doubts from Section 5.1 regarding the usefulness of a single constant factor based on the previous eigenvalue analysis. These results further support our analysis regarding the use of scaling factors for $\boldsymbol{A}$ for length generalization compared to a wide variety of methods.

7 Conclusion

In this work, we conduct an in-depth exploration regarding the state transition matrix of `Mamba` models. We first provide a broader understanding of the SSM parameterization and how it can affect length generalization in `Mamba` models. In particular, we analyze the eigenvalue spectrum of both `Mamba` and `Mamba2` models, identifying the specific role this can have on the convergence of SSMs given long inputs. Then we identify how the scaling of $\boldsymbol{A}$ as opposed to the more common practice of scaling Δ can be more effective at tuning this spectrum, enabling models to better generalize to long-contexts that far exceed the training context. We experiment on multiple long-context generalization tasks to validate that this newly built intuition holds empirically, on both `Mamba` and `Mamba2` models, highlighting the potential benefits of using $\boldsymbol{A}$ for length generalization.

8 Acknowledgements

Jerry Huang was supported by the NSERC Canada Graduate Scholarships Doctoral (CGS-D) program (funding reference number 589326) as well as the Bourse d'Excellence Hydro-Québec program.

References

[1] S. Arora, S. Eyuboglu, A. Timalsina, I. Johnson, M. Poli, J. Zou, A. Rudra, and C. Ré. Zoology: Measuring and Improving Recall in Efficient Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=LY3ukUANko.

[2] K. J. Åström. *Introduction to stochastic control theory*. Courier Corporation, 2012.

[3] S. Azizi, S. Kundu, M. E. Sadeghi, and M. Pedram. MambaExtend: A Training-Free Approach to Improve Long Context Extension of Mamba. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=LgzRo1RpLS.

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

[5] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu. Qwen Technical Report, 2023. URL https://doi.org/10.48550/arXiv.2309.16609. arXiv: 2309.16609.

[6] Y. Bai, X. Lv, J. Zhang, H. Lyu, J. Tang, Z. Huang, Z. Du, X. Liu, A. Zeng, L. Hou, Y. Dong, J. Tang, and J. Li. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3119–3137. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.172. URL https://doi.org/10.18653/v1/2024.acl-long.172.

[7] M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter. xLSTM: Extended Long Short-Term Memory. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/c2ce2f2701c10a2b2f2ea0bfa43cfaa3-Abstract-Conference.html.

[8] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The Long-Document Transformer, 2020. URL https://arxiv.org/abs/2004.05150. arXiv: 2004.05150.

[9] A. Ben-Kish, I. Zimerman, S. Abu-Hussein, N. Cohen, A. Globerson, L. Wolf, and R. Giryes. DeciMamba: Exploring the Length Extrapolation Potential of Mamba. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=iWSl5Zyjjw.

[10] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, H. Gao, K. Gao, W. Gao, R. Ge, K. Guan, D. Guo, J. Guo, G. Hao, Z. Hao, Y. He, W. Hu, P. Huang, E. Li, G. Li, J. Li, Y. Li, Y. K. Li, W. Liang, F. Lin, A. X. Liu, B. Liu, W. Liu, X. Liu, X. Liu, Y. Liu, H. Lu, S. Lu, F. Luo, S. Ma, X. Nie, T. Pei, Y. Piao, J. Qiu, H. Qu, T. Ren, Z. Ren, C. Ruan, Z. Sha, Z. Shao, J. Song, X. Su, J. Sun, Y. Sun, M. Tang, B. Wang, P. Wang, S. Wang, Y. Wang, Y. Wang, T. Wu, Y. Wu, X. Xie, Z. Xie, Z. Xie, Y. Xiong, H. Xu, R. X. Xu, Y. Xu, D. Yang, Y. You, S. Yu, X. Yu, B. Zhang, H. Zhang, L. Zhang, L. Zhang, M. Zhang, M. Zhang, W. Zhang, Y. Zhang, C. Zhao, Y. Zhao, S. Zhou, S. Zhou, Q. Zhu, and Y. Zou. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism, 2024. URL https://doi.org/10.48550/arXiv.2401.02954. arXiv: 2401.02954.

 https://doi.org/10.52202/085713-0653

[11] bloc97. NTK-Aware Scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation., June 2023. URL https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/.

[12] A. Botev, S. De, S. L. Smith, A. Fernando, G.-C. Muraru, R. Haroun, L. Berrada, R. Pascanu, P. G. Sessa, R. Dadashi, L. Hussenot, J. Ferret, S. Girgin, O. Bachem, A. Andreev, K. Kenealy, T. Mesnard, C. Hardin, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, A. Joulin, N. Fiedel, E. Senter, Y. Chen, S. Srinivasan, G. Desjardins, D. Budden, A. Doucet, S. Vikram, A. Paszke, T. Gale, S. Borgeaud, C. Chen, A. Brock, A. Paterson, J. Brennan, M. Risdal, R. Gundluru, N. Devanathan, P. Mooney, N. Chauhan, P. Culliton, L. G. Martins, E. Bandy, D. Huntsperger, G. Cameron, A. Zucker, T. Warkentin, L. Peran, M. Giang, Z. Ghahramani, C. Farabet, K. Kavukcuoglu, D. Hassabis, R. Hadsell, Y. W. Teh, and N. d. Frietas. RecurrentGemma: Moving Past Transformers for Efficient Open Language Models. *CoRR*, abs/2404.07839, 2024. doi: 10.48550/ARXIV.2404.07839. URL https://doi.org/10.48550/arXiv.2404.07839. arXiv: 2404.07839.

[13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H.-T. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[14] S. Chen, S. Wong, L. Chen, and Y. Tian. Extending Context Window of Large Language Models via Positional Interpolation, 2023. URL https://doi.org/10.48550/arXiv.2306.15595. arXiv: 2306.15595.

[15] K. Cho, B. v. Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL, 2014. doi: 10.3115/V1/D14-1179. URL https://doi.org/10.3115/v1/d14-1179.

[16] K. M. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlós, P. Hawkins, J. Q. Davis, A. Mohiuddin, L. Kaiser, D. B. Belanger, L. J. Colwell, and A. Weller. Rethinking Attention with Performers. 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=Ua6zuk0WRH.

[17] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. PaLM: Scaling Language Modeling with Pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2023. URL https://jmlr.org/papers/v24/22-1144.html.

[18] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Y. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling Instruction-Finetuned Language Models. *J. Mach. Learn. Res.*, 25:70:1–70:53, 2024. URL https://jmlr.org/papers/v25/23-0870.html.

[19] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms, 3rd Edition*. MIT Press, 2009. ISBN 978-0-262-03384-8. URL http://mitpress.mit.edu/books/introduction-algorithms.

[20] D. Dai, C. Deng, C. Zhao, R. X. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, Z. Xie, Y. K. Li, P. Huang, F. Luo, C. Ruan, Z. Sui, and W. Liang. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models, 2024. URL https://doi.org/10.48550/arXiv.2401.06066. arXiv: 2401.06066.

[21] T. Dao. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=mZn2Xyh9Ec.

[22] T. Dao and A. Gu. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=ztn8FCR1td.

[23] P. Dasigi, K. Lo, I. Beltagy, A. Cohan, N. A. Smith, and M. Gardner. A dataset of information-seeking questions and answers anchored in research papers. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4599–4610. Association for Computational Linguistics, 2021. URL https://doi.org/10.18653/v1/2021.naacl-main.365.

[24] DeepSeek-AI, A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Deng, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Xu, H. Yang, H. Zhang, H. Ding, H. Xin, H. Gao, H. Li, H. Qu, J. L. Cai, J. Liang, J. Guo, J. Ni, J. Li, J. Chen, J. Yuan, J. Qiu, J. Song, K. Dong, K. Gao, K. Guan, L. Wang, L. Zhang, L. Xu, L. Xia, L. Zhao, L. Zhang, M. Li, M. Wang, M. Zhang, M. Zhang, M. Tang, M. Li, N. Tian, P. Huang, P. Wang, P. Zhang, Q. Zhu, Q. Chen, Q. Du, R. J. Chen, R. L. Jin, R. Ge, R. Pan, R. Xu, R. Chen, S. S. Li, S. Lu, S. Zhou, S. Chen, S. Wu, S. Ye, S. Ma, S. Wang, S. Zhou, S. Yu, S. Zhou, S. Zheng, T. Wang, T. Pei, T. Yuan, T. Sun, W. L. Xiao, W. Zeng, W. An, W. Liu, W. Liang, W. Gao, W. Zhang, X. Q. Li, X. Jin, X. Wang, X. Bi, X. Liu, X. Wang, X. Shen, X. Chen, X. Chen, X. Nie, and X. Sun. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model, 2024. URL https://doi.org/10.48550/arXiv.2405.04434. arXiv: 2405.04434.

[25] DeepSeek-AI, A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Zhang, H. Ding, H. Xin, H. Gao, H. Li, H. Qu, J. L. Cai, J. Liang, J. Guo, J. Ni, J. Li, J. Wang, J. Chen, J. Chen, J. Yuan, J. Qiu, J. Li, J. Song, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Xu, L. Xia, L. Zhao, L. Wang, L. Zhang, M. Li, M. Wang, M. Zhang, M. Zhang, M. Tang, M. Li, N. Tian, P. Huang, P. Wang, P. Zhang, Q. Wang, Q. Zhu, Q. Chen, Q. Du, R. J. Chen, R. L. Jin, R. Ge, R. Zhang, R. Pan, R. Wang, R. Xu, R. Zhang, R. Chen, S. S. Li, S. Lu, S. Zhou, S. Chen, S. Wu, S. Ye, S. Ye, S. Ma, S. Wang, S. Zhou, S. Yu, S. Zhou, S. Pan, T. Wang, T. Yun, T. Pei, T. Sun, W. L. Xiao, and W. Zeng. DeepSeek-V3 Technical Report, 2024. URL https://doi.org/10.48550/arXiv.2412.19437. arXiv: 2412.19437.

[26] DeepSeek-AI, Q. Zhu, D. Guo, Z. Shao, D. Yang, P. Wang, R. Xu, Y. Wu, Y. Li, H. Gao, S. Ma, W. Zeng, X. Bi, Z. Gu, H. Xu, D. Dai, K. Dong, L. Zhang, Y. Piao, Z. Gou, Z. Xie, Z. Hao, B. Wang, J. Song, D. Chen, X. Xie, K. Guan, Y. You, A. Liu, Q. Du, W. Gao, X. Lu, Q. Chen, Y. Wang, C. Deng, J. Li, C. Zhao, C. Ruan, F. Luo, and W. Liang. DeepSeek-Coder-V2: Breaking the Barrier of Closed-Source Models in Code Intelligence, 2024. URL https://doi.org/10.48550/arXiv.2406.11931. arXiv: 2406.11931.

[27] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan,

 https://doi.org/10.52202/085713-0653

D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, and S. S. Li. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025. URL https://doi.org/10.48550/arXiv.2501.12948. arXiv: 2501.12948.

[28] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and L. Kaiser. Universal Transformers. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

[29] Z. Dong, T. Tang, J. Li, W. X. Zhao, and J.-R. Wen. BAMBOO: A Comprehensive Benchmark for Evaluating Long Text Modeling Capacities of Large Language Models. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 2086–2099. ELRA and ICCL, 2024. URL https://aclanthology.org/2024.lrec-main.188.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 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=YicbFdNTTy.

[31] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Rozière, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. M. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. v. d. Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, and e. al. The Llama 3 Herd of Models, 2024. URL https://doi.org/10.48550/arXiv.2407.21783. arXiv: 2407.21783.

[32] Y. Dubois, C. X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, and T. B. Hashimoto. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/5fc47800ee5b30b8777fdd30abcaaf3b-Abstract-Conference.html.

[33] J. C. Duchi, E. Hazan, and Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. In A. T. Kalai and M. Mohri, editors, *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 257–269. Omnipress, 2010. URL http://colt2010.haifa.il.ibm.com/papers/COLT2010proceedings.pdf#page=265.

[34] J. C. Duchi, E. Hazan, and Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011. doi: 10.5555/1953048.2021068. URL https://dl.acm.org/doi/10.5555/1953048.2021068.

[35] emozilla. Dynamically Scaled RoPE further increases performance of long context LLaMA with zero fine-tuning, June 2023. URL https://www.reddit.com/r/LocalLLaMA/comments/14mrgpr/dynamically_scaled_rope_further_increases/.

[36] W. Fedus, B. Zoph, and N. Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39, 2022. URL https://jmlr.org/papers/v23/21-0998.html.

[37] D. Y. Fu, S. Arora, J. Grogan, I. Johnson, E. S. Eyuboglu, A. W. Thomas, B. Spector, M. Poli, A. Rudra, and C. Ré. Monarch Mixer: A Simple Sub-Quadratic GEMM-Based Architecture. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/f498c1ce6bff52eb04febf87438dd84b-Abstract-Conference.html.

[38] D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, and C. Ré. Hungry Hungry Hippos: Towards Language Modeling with State Space Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=COZDy0WYGg.

[39] D. Y. Fu, E. L. Epstein, E. Nguyen, A. W. Thomas, M. Zhang, T. Dao, A. Rudra, and C. Ré. Simple Hardware-Efficient Long Convolutions for Sequence Modeling. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10373–10391. PMLR, 2023. URL https://proceedings.mlr.press/v202/fu23a.html.

[40] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy. The Pile: An 800GB Dataset of Diverse Text for Language Modeling, 2021. URL https://arxiv.org/abs/2101.00027. arXiv: 2101.00027.

[41] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. C. Courville, and Y. Bengio. Maxout Networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1319–1327. JMLR.org, 2013. URL http://proceedings.mlr.press/v28/goodfellow13.html.

[42] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

[43] A. Graves, G. Wayne, and I. Danihelka. Neural Turing Machines. *CoRR*, abs/1410.5401, 2014. URL http://arxiv.org/abs/1410.5401. arXiv: 1410.5401.

[44] R. Grazzi, J. Siems, A. Zela, J. K. H. Franke, F. Hutter, and M. Pontil. Unlocking State-Tracking in Linear RNNs Through Negative Eigenvalues. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=UvTo3tVBk2.

[45] A. Gu and T. Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In *The First Conference on Language Modeling, COLM 2024, Philadelphia, Pennsylvania, USA, October 7-9, 2024*, Oct. 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

[46] A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré. HiPPO: Recurrent Memory with Optimal Polynomial Projections. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H.-T. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/102f0bb6efb3a6128a3c750dd16729be-Abstract.html.

 https://doi.org/10.52202/085713-0653

[47] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré. Combining Recurrent, Convolutional, and Continuous-time Models with Linear State Space Layers. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 572–585, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/05546b0e38ab9175cd905eebcc6ebb76-Abstract.html.

[48] A. Gu, K. Goel, A. Gupta, and C. Ré. On the Parameterization and Initialization of Diagonal State Space Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/e9a32fade47b906de908431991440f7c-Abstract-Conference.html.

[49] A. Gu, K. Goel, and C. 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.

[50] D. Guo, C. Xu, N. Duan, J. Yin, and J. J. McAuley. Longcoder: A long-range pre-trained language model for code completion. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 12098–12107. PMLR, 2023. URL https://proceedings.mlr.press/v202/guo23j.html.

[51] D. Guo, Q. Zhu, D. Yang, Z. Xie, K. Dong, W. Zhang, G. Chen, X. Bi, Y. Wu, Y. K. Li, F. Luo, Y. Xiong, and W. Liang. DeepSeek-Coder: When the Large Language Model Meets Programming - The Rise of Code Intelligence, 2024. URL https://doi.org/10.48550/arXiv.2401.14196. arXiv: 2401.14196.

[52] C. Han, Q. Wang, H. Peng, W. Xiong, Y. Chen, H. Ji, and S. Wang. LM-Infinite: Zero-Shot Extreme Length Generalization for Large Language Models. In K. Duh, H. Gómez-Adorno, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3991–4008. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.222. URL https://doi.org/10.18653/v1/2024.naacl-long.222.

[53] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. Distributed Representations. In M. A. Boden, editor, *The Philosophy of Artificial Intelligence*, Oxford readings in philosophy, pages 248–280. Oxford University Press, 1990.

[54] X. Ho, A. D. Nguyen, S. Sugawara, and A. Aizawa. Constructing A multi-hop QA dataset for comprehensive evaluation of reasoning steps. In D. Scott, N. Bel, and C. Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6609–6625. International Committee on Computational Linguistics, 2020. URL https://doi.org/10.18653/v1/2020.coling-main.580.

[55] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780, 1997. doi: 10.1162/NECO.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

[56] C.-P. Hsieh, S. Sun, S. Kriman, S. Acharya, D. Rekesh, F. Jia, and B. Ginsburg. RULER: Whats the Real Context Size of Your Long-Context Language Models? In *The First Conference on Language Modeling, COLM 2024, Philadelphia, Pennsylvania, USA, October 7-9, 2024*, Oct. 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

[57] J. Huang. How Well Can a Long Sequence Model Model Long Sequences? Comparing Architectural Inductive Biases on Long-Context Abilities. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st*

International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025, pages 29–39. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.3/.

[58] J. Huang, P. Lu, and Q. Zeng. Calibrated language models and how to find them with label smoothing. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=soLNj4l2EL.

[59] L. Huang, S. Cao, N. N. Parulian, H. Ji, and L. Wang. Efficient Attentions for Long Document Summarization. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1419–1436. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.112. URL https://doi.org/10.18653/v1/2021.naacl-main.112.

[60] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Dang, A. Yang, R. Men, F. Huang, X. Ren, X. Ren, J. Zhou, and J. Lin. Qwen2.5-Coder Technical Report, 2024. URL https://doi.org/10.48550/arXiv.2409.12186. arXiv: 2409.12186.

[61] D. Hutchins, I. Schlag, Y. Wu, E. Dyer, and B. Neyshabur. Block-Recurrent Transformers. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/d6e0bbb9fc3f4c10950052ec2359355c-Abstract-Conference.html.

[62] K. Irie, I. Schlag, R. Csordás, and J. Schmidhuber. Going Beyond Linear Transformers with Recurrent Fast Weight Programmers. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 7703–7717, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/3f9e3767ef3b10a0de4c256d7ef9805d-Abstract.html.

[63] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive Mixtures of Local Experts. *Neural Comput.*, 3(1):79–87, 1991. doi: 10.1162/NECO.1991.3.1.79. URL https://doi.org/10.1162/neco.1991.3.1.79.

[64] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, A. Iftimie, A. Karpenko, A. T. Passos, A. Neitz, A. Prokofiev, A. Wei, A. Tam, A. Bennett, A. Kumar, A. Saraiva, A. Vallone, A. Duberstein, A. Kondrich, A. Mishchenko, A. Applebaum, A. Jiang, A. Nair, B. Zoph, B. Ghorbani, B. Rossen, B. Sokolowsky, B. Barak, B. McGrew, B. Minaiev, B. Hao, B. Baker, B. Houghton, B. McKinzie, B. Eastman, C. Lugaresi, C. Bassin, C. Hudson, C. M. Li, C. d. Bourcy, C. Voss, C. Shen, C. Zhang, C. Koch, C. Orsinger, C. Hesse, C. Fischer, C. Chan, D. Roberts, D. Kappler, D. Levy, D. Selsam, D. Dohan, D. Farhi, D. Mely, D. Robinson, D. Tsipras, D. Li, D. Oprica, E. Freeman, E. Zhang, E. Wong, E. Proehl, E. Cheung, E. Mitchell, E. Wallace, E. Ritter, E. Mays, F. Wang, F. P. Such, F. Raso, F. Leoni, F. Tsimpourlas, F. Song, F. v. Lohmann, F. Sulit, G. Salmon, G. Parascandolo, G. Chabot, G. Zhao, G. Brockman, G. Leclerc, H. Salman, H. Bao, H. Sheng, H. Andrin, H. Bagherinezhad, H. Ren, H. Lightman, H. W. Chung, I. Kivlichan, I. O'Connell, I. Osband, I. C. Gilaberte, and I. Akkaya. OpenAI o1 System Card. *CoRR*, abs/2412.16720, 2024. doi: 10.48550/ARXIV.2412.16720. URL https://doi.org/10.48550/arXiv.2412.16720. arXiv: 2412.16720.

[65] S. Jelassi, D. Brandfonbrener, S. M. Kakade, and E. Malach. Repeat After Me: Transformers are Better than State Space Models at Copying. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=duRRoGeoQT.

[66] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. L. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7B, 2023. URL https://doi.org/10.48550/arXiv.2310.06825. arXiv: 2310.06825.

[67] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. L. Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mixtral of Experts, 2024. URL https://doi.org/10.48550/arXiv.2401.04088. arXiv: 2401.04088.

[68] H. Jiang, Y. Li, C. Zhang, Q. Wu, X. Luo, S. Ahn, Z. Han, A. Abdi, D. Li, C.-Y. Lin, Y. Yang, and L. Qiu. MInference 1.0: Accelerating Pre-filling for Long-Context LLMs via Dynamic Sparse Attention. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/5dfbe6f5671e82c76841ba687a8a9ecb-Abstract-Conference.html.

[69] M. I. Jordan and D. E. Rumelhart. Internal World Models and Supervised Learning. In L. Birnbaum and G. Collins, editors, *Proceedings of the Eighth International Workshop (ML91), Northwestern University, Evanston, Illinois, USA*, pages 70–74. Morgan Kaufmann, 1991. doi: 10.1016/B978-1-55860-200-7.50018-0. URL https://doi.org/10.1016/b978-1-55860-200-7.50018-0.

[70] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In R. Barzilay and M. Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics, 2017. URL https://doi.org/10.18653/v1/P17-1147.

[71] A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J.-B. Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Kenealy, L. Beyer, X. Zhai, A. Tsitsulin, R. Busa-Fekete, A. Feng, N. Sachdeva, B. Coleman, Y. Gao, B. Mustafa, I. Barr, E. Parisotto, D. Tian, M. Eyal, C. Cherry, J.-T. Peter, D. Sinopalnikov, S. Bhupatiraju, R. Agarwal, M. Kazemi, D. Malkin, R. Kumar, D. Vilar, I. Brusilovsky, J. Luo, A. Steiner, A. Friesen, A. Sharma, A. Sharma, A. M. Gilady, A. Goedeckemeyer, A. Saade, A. Kolesnikov, A. Bendebury, A. Abdagic, A. Vadi, A. György, A. S. Pinto, A. Das, A. Bapna, A. Miech, A. Yang, A. Paterson, A. Shenoy, A. Chakrabarti, B. Piot, B. Wu, B. Shahriari, B. Petrini, C. Chen, C. L. Lan, C. A. Choquette-Choo, C. J. Carey, C. Brick, D. Deutsch, D. Eisenbud, D. Cattle, D. Cheng, D. Paparas, D. S. Sreepathihalli, D. Reid, D. Tran, D. Zelle, E. Noland, E. Huizenga, E. Kharitonov, F. Liu, G. Amirkhanyan, G. Cameron, H. Hashemi, H. Klimczak-Plucinska, H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Ballantyne, I. Szpektor, and I. Nardini. Gemma 3 Technical Report. *CoRR*, abs/2503.19786, 2025. doi: 10.48550/ARXIV.2503.19786. URL https://doi.org/10.48550/arXiv.2503.19786. arXiv: 2503.19786.

[72] A. Katharopoulos, A. Vyas, N. Pappas, and F. 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*, pages 5156–5165. PMLR, 2020. URL http://proceedings.mlr.press/v119/katharopoulos20a.html.

[73] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

[74] N. Kitaev, L. Kaiser, and A. Levskaya. Reformer: The Efficient Transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB.

[75] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[76] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. 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=qrwe7XHTmYb.

[77] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H.-T. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[78] X. Li and D. Roth. Learning question classifiers. In *19th International Conference on Computational Linguistics, COLING 2002, Howard International House and Academia Sinica, Taipei, Taiwan, August 24 - September 1, 2002*, 2002. URL https://aclanthology.org/C02-1150/.

[79] X. L. Li and P. Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.353. URL https://doi.org/10.18653/v1/2021.acl-long.353.

[80] H. Liu, M. Zaharia, and P. Abbeel. RingAttention with Blockwise Transformers for Near-Infinite Context. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=WsRHpHH4s0.

[81] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguistics*, 12: 157–173, 2024. doi: 10.1162/TACL_A_00638. URL https://doi.org/10.1162/tacl_a_00638.

[82] T. Liu, C. Xu, and J. J. McAuley. Repobench: Benchmarking repository-level code autocompletion systems. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=pPjZIOuQuF.

[83] X. Liu, H. Yan, C. An, X. Qiu, and D. Lin. Scaling Laws of RoPE-based Extrapolation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=JO7k0SJ5V6.

[84] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, and A. Roberts. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR, 2023. URL https://proceedings.mlr.press/v202/longpre23a.html.

 https://doi.org/10.52202/085713-0653

[85] I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[86] P. Lu, S. Wang, M. Rezagholizadeh, B. Liu, and I. Kobyzev. Efficient classification of long documents via state-space models. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6559–6565, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.404. URL https://aclanthology.org/2023.emnlp-main.404/.

[87] P. Lu, I. Kobyzev, M. Rezagholizadeh, B. Chen, and P. Langlais. Regla: Refining gated linear attention. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 2884–2898. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.147. URL https://doi.org/10.18653/v1/2025.naacl-long.147.

[88] S. Massaroli, M. Poli, D. Y. Fu, H. Kumbong, R. N. Parnichkun, D. W. Romero, A. Timalsina, Q. McIntyre, B. Chen, A. Rudra, C. Zhang, C. Ré, S. Ermon, and Y. Bengio. Laughing Hyena Distillery: Extracting Compact Recurrences From Convolutions. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/371355cd42caaf83412c3fbef4688979-Abstract-Conference.html.

[89] S. Merity, C. Xiong, J. Bradbury, and R. 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.

[90] T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, and e. al. Gemma: Open Models Based on Gemini Research and Technology. *CoRR*, abs/2403.08295, 2024. doi: 10.48550/ARXIV.2403.08295. URL https://doi.org/10.48550/arXiv.2403.08295. arXiv: 2403.08295.

[91] A. Mohtashami and M. Jaggi. Random-Access Infinite Context Length for Transformers. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/ab05dc8bf36a9f66edbff6992ec86f56-Abstract-Conference.html.

[92] OpenAI. GPT-4 Technical Report, 2023. URL https://doi.org/10.48550/arXiv.2303.08774. arXiv: 2303.08774.

[93] A. Orvieto, S. L. Smith, A. Gu, A. Fernando, Ç. Gülçehre, R. Pascanu, and S. De. Resurrecting Recurrent Neural Networks for Long Sequences. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 26670–26698. PMLR, 2023. URL https://proceedings.mlr.press/v202/orvieto23a.html.

[94] J. Park, J. Park, Z. Xiong, N. Lee, J. Cho, S. Oymak, K. Lee, and D. Papailiopoulos. Can Mamba Learn How To Learn? A Comparative Study on In-Context Learning Tasks. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=GbFluKMmtE.

[95] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1310–1318. JMLR.org, 2013. URL http://proceedings.mlr.press/v28/pascanu13.html.

[96] B. Peng, J. Quesnelle, H. Fan, and E. Shippole. YaRN: Efficient Context Window Extension of Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=wHBfxhZu1u.

[97] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. H. Miller. Language Models as Knowledge Bases? In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1250. URL https://doi.org/10.18653/v1/D19-1250.

[98] M. Poli, S. Massaroli, E. Nguyen, D. Y. Fu, T. Dao, S. Baccus, Y. Bengio, S. Ermon, and C. Ré. Hyena Hierarchy: Towards Larger Convolutional Language Models. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28043–28078. PMLR, 2023. URL https://proceedings.mlr.press/v202/poli23a.html.

[99] M. Poli, A. W. Thomas, E. Nguyen, P. Ponnusamy, B. Deiseroth, K. Kersting, T. Suzuki, B. L. Hie, S. Ermon, C. Ré, C. Zhang, and S. Massaroli. Mechanistic Design and Scaling of Hybrid Architectures. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=GDp7Gyd9nf.

[100] O. Press, N. A. Smith, and M. Lewis. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. 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=R8sQPpGCv0.

[101] Z. Qin, S. Yang, and Y. Zhong. Hierarchically Gated Recurrent Neural Network for Sequence Modeling. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/694be3548697e9cc8999d45e8d16fe1e-Abstract-Conference.html.

[102] Z. Qin, S. Yang, W. Sun, X. Shen, D. Li, W. Sun, and Y. Zhong. HGRN2: Gated Linear RNNs with State Expansion. In *The First Conference on Language Modeling, COLM 2024, Philadelphia, Pennsylvania, USA, October 7-9, 2024*, Oct. 2024. URL https://openreview.net/forum?id=y6SqbJfCSk.

[103] J. W. Rae, A. Potapenko, S. M. Jayakumar, C. Hillier, and T. P. Lillicrap. Compressive Transformers for Long-Range Sequence Modelling. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SylKikSYDH.

[104] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

[105] M. Rivière, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, J. Ferret, P. Liu, P. Tafti, A. Friesen, M. Casbon, S. Ramos, R. Kumar, C. L. Lan, S. Jerome, A. Tsitsulin, N. Vieillard, P. Stanczyk, S. Girgin, N. Momchev,

 https://doi.org/10.52202/085713-0653

M. Hoffman, S. Thakoor, J.-B. Grill, B. Neyshabur, O. Bachem, A. Walton, A. Severyn, A. Parrish, A. Ahmad, A. Hutchison, A. Abdagic, A. Carl, A. Shen, A. Brock, A. Coenen, A. Laforge, A. Paterson, B. Bastian, B. Piot, B. Wu, B. Royal, C. Chen, C. Kumar, C. Perry, C. Welty, C. A. Choquette-Choo, D. Sinopalnikov, D. Weinberger, D. Vijaykumar, D. Rogozinska, D. Herbison, E. Bandy, E. Wang, E. Noland, E. Moreira, E. Senter, E. Eltyshev, F. Visin, G. Rasskin, G. Wei, G. Cameron, G. Martins, H. Hashemi, H. Klimczak-Plucinska, H. Batra, H. Dhand, I. Nardini, J. Mein, J. Zhou, J. Svensson, J. Stanway, J. Chan, J. P. Zhou, J. Carrasqueira, J. Iljazi, J. Becker, J. Fernandez, J. v. Amersfoort, J. Gordon, J. Lipschultz, J. Newlan, J.-y. Ji, K. Mohamed, K. Badola, K. Black, K. Millican, K. McDonell, K. Nguyen, K. Sodhia, K. Greene, L. L. Sjösund, L. Usui, L. Sifre, L. Heuermann, L. Lago, and L. McNealus. Gemma 2: Improving Open Language Models at a Practical Size. *CoRR*, abs/2408.00118, 2024. doi: 10.48550/ARXIV.2408.00118. URL https://doi.org/10.48550/arXiv.2408.00118. arXiv: 2408.00118.

[106] A. Roberts, C. Raffel, and N. Shazeer. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5418–5426. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.437. URL https://doi.org/10.18653/v1/2020.emnlp-main.437.

[107] B. Rozière, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, T. Remez, J. Rapin, A. Kozhevnikov, I. Evtimov, J. Bitton, M. Bhatt, C. Canton-Ferrer, A. Grattafiori, W. Xiong, A. Défossez, J. Copet, F. Azhar, H. Touvron, L. Martin, N. Usunier, T. Scialom, and G. Synnaeve. Code Llama: Open Foundation Models for Code, 2023. URL https://doi.org/10.48550/arXiv.2308.12950. arXiv: 2308.12950.

[108] W. J. Rugh. *Linear system theory*. Prentice-Hall, Inc., 1996.

[109] D. E. Rumelhart. *Parallel distributed processing, 9th Edition*. MIT Pr., 1989. ISBN 978-0-262-68053-0. URL https://www.worldcat.org/oclc/60445750.

[110] P. Sadegh and J. C. Spall. Optimal random perturbations for stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Autom. Control.*, 43(10): 1480–1484, 1998. doi: 10.1109/9.720513.

[111] R. Salakhutdinov and G. E. Hinton. Deep Boltzmann Machines. In D. A. V. Dyk and M. Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009*, volume 5 of *JMLR Proceedings*, pages 448–455. JMLR.org, 2009. URL http://proceedings.mlr.press/v5/salakhutdinov09a.html.

[112] T. Salimans and D. P. Kingma. Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. In D. D. Lee, M. Sugiyama, U. v. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, page 901, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/ed265bc903a5a097f61d3ec064d96d2e-Abstract.html.

[113] I. Schlag, K. Irie, and J. Schmidhuber. Linear Transformers Are Secretly Fast Weight Programmers. In M. Meila and T. Zhang, editors, *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*, pages 9355–9366. PMLR, 2021. URL http://proceedings.mlr.press/v139/schlag21a.html.

[114] J. Shah, G. Bikshandi, Y. Zhang, V. Thakkar, P. Ramani, and T. Dao. FlashAttention-3: Fast and Accurate Attention with Asynchrony and Low-precision. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/7ede97c3e082c6df10a8d6103a2eebd2-Abstract-Conference.html.

[115] U. Shaham, E. Segal, M. Ivgi, A. Efrat, O. Yoran, A. Haviv, A. Gupta, W. Xiong, M. Geva, J. Berant, and O. Levy. SCROLLS: Standardized CompaRison Over Long Language Sequences. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 12007–12021. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.823. URL https://doi.org/10.18653/v1/2022.emnlp-main.823.

[116] U. Shaham, M. Ivgi, A. Efrat, J. Berant, and O. Levy. ZeroSCROLLS: A Zero-Shot Benchmark for Long Text Understanding. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7977–7989. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.536. URL https://doi.org/10.18653/v1/2023.findings-emnlp.536.

[117] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Zhang, Y. K. Li, Y. Wu, and D. Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, 2024. URL https://doi.org/10.48550/arXiv.2402.03300. arXiv: 2402.03300.

[118] F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, D. Das, and J. Wei. Language Models are Multilingual Chain-of-Thought Reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=fR3wGCk-IXp.

[119] J. C. Spall. *Introduction to stochastic search and optimization - estimation, simulation, and control*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 2003. ISBN 978-0-471-33052-3. doi: 10.1002/0471722138.

[120] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014. doi: 10.5555/2627435.2670313. URL https://dl.acm.org/doi/10.5555/2627435.2670313.

[121] J. Su, M. H. M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/J.NEUCOM.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

[122] Y. Sun, L. Dong, B. Patra, S. Ma, S. Huang, A. Benhaim, V. Chaudhary, X. Song, and F. Wei. A Length-Extrapolatable Transformer. In A. Rogers, J. L. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14590–14604. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.816. URL https://doi.org/10.18653/v1/2023.acl-long.816.

[123] I. Sutskever, J. Martens, and G. E. Hinton. Generating Text with Recurrent Neural Networks. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 1017–1024, 2011. URL https://icml.cc/2011/papers/524_icmlpaper.pdf.

[124] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H.-T. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/55053683268957697aa39fba6f231c68-Abstract.html.

[125] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler. Long Range Arena : A Benchmark for Efficient Transformers. 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=qVyeW-grC2k.

[126] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler. Efficient Transformers: A Survey. *ACM Comput. Surv.*, 55(6):109:1–109:28, 2023. doi: 10.1145/3530811.

[127] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. LLaMA: Open and Efficient Foundation Language Models, 2023. URL https://doi.org/10.48550/arXiv.2302.13971. arXiv: 2302.13971.

[128] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023. URL https://doi.org/10.48550/arXiv.2307.09288. arXiv: 2307.09288.

[129] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is All you Need. In I. Guyon, U. v. Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[130] J. Wang, D. Paliotta, A. May, A. M. Rush, and T. Dao. The Mamba in the Llama: Distilling and Accelerating Hybrid Models. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/723933067ad315269b620bc0d2c05cba-Abstract-Conference.html.

[131] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma. Linformer: Self-Attention with Linear Complexity, 2020. URL https://arxiv.org/abs/2006.04768. arXiv: 2006.04768.

[132] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

[133] X. Wang, L. Ma, J. Huang, P. Lu, P. Parthasarathi, X.-W. Chang, B. Chen, and Y. Cui. Resona: Improving context copying in linear recurrence models with retrieval. In *The Second Conference on Language Modeling, COLM 2025, Montréal, Canada, September 3-6, 2025*, Sept. 2025. URL https://openreview.net/forum?id=4mxQmpnawk.

[134] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned Language Models are Zero-Shot Learners. 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=gEZrGCozdqR.

[135] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent Abilities of Large Language Models. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=yzkSU5zdwD.

[136] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In

S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[137] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. v. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-Art Natural Language Processing. In Q. Liu and D. Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-DEMOS.6. URL https://doi.org/10.18653/v1/2020.emnlp-demos.6.

[138] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis. Efficient Streaming Language Models with Attention Sinks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

[139] M. Xu, X. Men, B. Wang, Q. Zhang, H. Lin, X. Han, and W. Chen. Base of RoPE Bounds Context Length. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/9f12dd32d552f3ad9eaa0e9dfec291be-Abstract-Conference.html.

[140] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, and Z. Fan. Qwen2 Technical Report, 2024. URL https://doi.org/10.48550/arXiv.2407.10671. arXiv: 2407.10671.

[141] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 Technical Report, 2024. URL https://doi.org/10.48550/arXiv.2412.15115. arXiv: 2412.15115.

[142] A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, C. Li, D. Liu, J. Tu, J. Zhou, J. Lin, K. Lu, M. Xue, R. Lin, T. Liu, X. Ren, and Z. Zhang. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement, 2024. URL https://doi.org/10.48550/arXiv.2409.12122. arXiv: 2409.12122.

[143] S. Yang, B. Wang, Y. Shen, R. Panda, and Y. Kim. Gated Linear Attention Transformers with Hardware-Efficient Training. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=ia5XvxFUJT.

[144] S. Yang, B. Wang, Y. Zhang, Y. Shen, and Y. Kim. Parallelizing Linear Transformers with the Delta Rule over Sequence Length. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/d13a3eae72366e61dfdc7eea82eeb685-Abstract-Conference.html.

[145] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In E. Riloff,

 https://doi.org/10.52202/085713-0653

D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics, 2018. URL https://doi.org/10.18653/v1/d18-1259.

[146] Z. Ye, K. Xia, Y. Fu, X. Dong, J. Hong, X. Yuan, S. Diao, J. Kautz, P. Molchanov, and Y. C. Lin. LongMamba: Enhancing Mamba's Long-Context Capabilities via Training-Free Receptive Field Enlargement. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=fMbLszVO1H.

[147] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontañón, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed. Big Bird: Transformers for Longer Sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H.-T. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html.

[148] Q. Zeng, J. Huang, P. Lu, G. Xu, B. Chen, C. Ling, and B. Wang. ZETA: Leveraging Z-order Curves for Efficient Top-k Attention. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=j9VVzueEbG.

[149] X. Zhang, Y. Chen, S. Hu, Z. Xu, J. Chen, M. K. Hao, X. Han, Z. L. Thai, S. Wang, Z. Liu, and M. Sun. nftyBench: Extending Long Context Evaluation Beyond 100K Tokens. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15262–15277. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.814. URL https://doi.org/10.18653/v1/2024.acl-long.814.

[150] Y. Zhang, S. Yang, R.-J. Zhu, Y. Zhang, L. Cui, Y. Wang, B. Wang, F. Shi, B. Wang, W. Bi, P. Zhou, and G. Fu. Gated Slot Attention for Efficient Linear-Time Sequence Modeling. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/d3f39e51f5f634fb16cc3e658f8512b9-Abstract-Conference.html.

[151] Y. Zhang, S. Yang, R.-J. Zhu, Y. Zhang, L. Cui, Y. Wang, B. Wang, F. Shi, B. Wang, W. Bi, P. Zhou, and G. Fu. Gated Slot Attention for Efficient Linear-Time Sequence Modeling. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/d3f39e51f5f634fb16cc3e658f8512b9-Abstract-Conference.html.

[152] Z. Zhang, Y. Sheng, T. Zhou, T. Chen, L. Zheng, R. Cai, Z. Song, Y. Tian, C. Ré, C. W. Barrett, Z. Wang, and B. Chen. H2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6ceefa7b15572587b78ecfcebb2827f8-Abstract-Conference.html.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims come with experimental support as well as theoretical proofs when necessary.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide this in the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/085713-0653

Justification: We provide proof to all these in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all these details in full.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We mention which code bases we base our experiments off of.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide these details in our experimental section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We show these within our plots.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

https://doi.org/10.52202/085713-0653

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide this in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have complied with the NeurIPS Ethics Guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide this in the Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not introduce anything within this work that would be interpreted as having possibility of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The authors cite the original paper that produced used code packages and dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

https://doi.org/10.52202/085713-0653

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowd-sourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowd-sourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We do not use LLMs in any non-standard or novel manner.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-0653

A Proofs

A.1 Proof of Lemma 4.1

Lemma 4.1. Let $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ be a matrix with $\|\boldsymbol{B}\|_2 = \sigma_B$, and let $\boldsymbol{x} \in \mathbb{R}^d$ be a vector such that each entry of $\boldsymbol{x}$ satisfies $|x_i| \le \sigma_x$. The upper bound for $\|\boldsymbol{Bx}\|_2$ is:

$$\|\boldsymbol{Bx}\|_2 \le \sigma_B \cdot \sigma_x \cdot \sqrt{d}.$$

Proof. For any vector $\boldsymbol{x} \in \mathbb{R}^d$, it follows that:

$$\|\boldsymbol{Bx}\|_2 \le \|\boldsymbol{B}\|_2 \cdot \|\boldsymbol{x}\|_2 .$$

Substituting $\|\boldsymbol{B}\|_2 = \sigma_B$, we obtain:

$$\|\boldsymbol{Bx}\|_2 \le \sigma_B \cdot \|\boldsymbol{x}\|_2 .$$

And

$$\|\boldsymbol{x}\|_2 \le \sqrt{\sum_{i=1}^{d} \sigma_x^2} = \sqrt{d} \cdot \sigma_x.$$

Substituting the bound on $\|\boldsymbol{x}\|_2$ into the inequality for $\|\boldsymbol{Bx}\|_2$, we have the norm of logit vector $u \in \mathbb{R}^d$:

$$\|\boldsymbol{u}\|_2 = \|\boldsymbol{Bx}\|_2 \le \sigma_B \cdot \|\boldsymbol{x}\|_2 \le \sigma_B \cdot \sqrt{d} \cdot \sigma_x.$$

□

A.2 Proof of Theorem 4.2

Theorem 4.2. Assume the transition matrix $\boldsymbol{\Lambda}$ is diagonal with eigenvalues $\lambda_i \sim \mathrm{Uniform}[\lambda_{\min}, \lambda_{\max}]$ for $0 < \lambda_{\min} < \lambda_{\max} < 1$. Suppose the system evolves as

$$\boldsymbol{h}_t = \boldsymbol{\Lambda h}_{t-1} + \boldsymbol{Bx}_t, \tag{10}$$

where $\boldsymbol{x}_t \sim \mathcal{N}(0, I)$ and $\boldsymbol{B}$ is a weight matrix whose rows are independently sampled as $\boldsymbol{b} \sim \mathcal{N}(0, \frac{1}{\sqrt{d}}\mathbf{I})$. Then, in the limit $t \to \infty$, the expected squared norm of the hidden state converges to

$$\mathbb{E}[\|\boldsymbol{h}_\infty\|^2] = \frac{1}{2(\lambda_{\max} - \lambda_{\min})} \log\left(\frac{1-\lambda_{\min}^2}{1-\lambda_{\max}^2}\right) \cdot \mathbb{E}[\|\boldsymbol{Bx}\|^2]. \tag{11}$$

Proof. We begin by unrolling the recurrence:

$$\boldsymbol{h}_t = \sum_{i=0}^{t-1} \boldsymbol{\Lambda}^i \boldsymbol{Bx}_{t-i}. \tag{12}$$

Assuming stationarity and independence of the inputs $\boldsymbol{x}_t$, the expected squared norm at steady state is

$$\mathbb{E}[\|\boldsymbol{h}_\infty\|^2] = \sum_{i=0}^{\infty} \mathbb{E}[\|\boldsymbol{\Lambda}^i \boldsymbol{Bx}\|^2]. \tag{13}$$

Consider the case of a single unit with eigenvalue $\lambda \in [\lambda_{\min}, \lambda_{\max}]$. The contribution of this unit is:

$$\mathbb{E}[h^2] = \sum_{i=0}^{\infty} \lambda^{2i} \mathbb{E}[\|\boldsymbol{bx}\|^2] = \frac{\sigma^2}{1-\lambda^2}, \tag{14}$$

where $\mathbb{E}[\|\boldsymbol{bx}\|^2] = \mathbb{E}_b[\mathbb{E}_x[(bu)^2|b]] = \mathbb{E}_b[\|b\|^2] = \sigma^2$ is the contribution from the corresponding row of $\boldsymbol{B}$, and $\boldsymbol{B}$ is a weight matrix whose rows are independently sampled as $\boldsymbol{b} \sim \mathcal{N}(0, \frac{1}{\sqrt{d}}\mathbf{I})$.

With $\lambda \sim \text{Uniform}[\lambda_{\min}, \lambda_{\max}]$, where $0 \leq \lambda_{\min} < \lambda_{\max} < 1$, the expected contribution over all units is

$$\mathbb{E}[\|\boldsymbol{h}_\infty\|^2] = d \cdot \mathbb{E}_\lambda \left[\frac{\sigma^2}{1-\lambda^2}\right] = \sigma^2 d \cdot \frac{1}{\lambda_{\max} - \lambda_{\min}} \int_{\lambda_{\min}}^{\lambda_{\max}} \frac{1}{1-\lambda^2}\, d\lambda. \tag{15}$$

Evaluating the integral:

$$\int_{\lambda_{\min}}^{\lambda_{\max}} \frac{1}{1-\lambda^2}\, d\lambda = \frac{1}{2} \log\left(\frac{1-\lambda_{\min}^2}{1-\lambda_{\max}^2}\right). \tag{16}$$

Hence,

$$\mathbb{E}[\|\boldsymbol{h}_\infty\|^2] = \sigma^2 d \cdot \frac{1}{2(\lambda_{\max} - \lambda_{\min})} \log\left(\frac{1-\lambda_{\min}^2}{1-\lambda_{\max}^2}\right). \tag{17}$$

Since $\mathbb{E}[\|\boldsymbol{B}\boldsymbol{x}\|^2] = d \cdot \sigma^2$, we obtain the final expression:

$$\mathbb{E}[\|\boldsymbol{h}_\infty\|^2] = \frac{1}{2(\lambda_{\max} - \lambda_{\min})} \log\left(\frac{1-\lambda_{\min}^2}{1-\lambda_{\max}^2}\right) \cdot \mathbb{E}[\|\boldsymbol{B}\boldsymbol{x}\|^2]. \tag{18}$$

□

A.3 Proof of Corollary 4.3 and Corollary 4.4

Corollary 4.3 [Norm of `Mamba` State] Suppose the diagonal entries of $\boldsymbol{\Lambda}$ are independently drawn from a uniform distribution on $[0, \lambda]$, a moderate discretized step value Δ and the system evolves as $\boldsymbol{h}_t = \boldsymbol{\Lambda}\boldsymbol{h}_{t-1} + \overline{\boldsymbol{B}}\boldsymbol{x}_t = \text{diag}\left(\exp(-\Delta\alpha)\right)\boldsymbol{h}_{t-1} + \Delta\boldsymbol{B}\boldsymbol{x}_t$. Then the convergence rate ρ of the expected squared norm of the limiting state satisfies $\mathcal{O}\left(\frac{\Delta}{2\lambda}\log\left(\frac{1}{1-\lambda^2}\right)\right)$.

Proof. Given Theorem 4.2, the convergence rate ρ of `Mamba` state can be estimated as $\lambda_{min} \to 0$:

$$\rho = \lim_{t\to\infty} \frac{\mathbb{E}[\|\boldsymbol{h}_t\|^2]}{\mathbb{E}[\|\overline{\boldsymbol{B}}\boldsymbol{x}\|^2]} = \lim_{t\to\infty} \frac{\Delta\mathbb{E}[\|\boldsymbol{h}_t\|^2]}{\mathbb{E}[\|\boldsymbol{B}\boldsymbol{x}\|^2]} = \frac{\Delta}{2\lambda}\log\left(\frac{1}{1-\lambda^2}\right) \tag{19}$$

□

Corollary 4.4 [Norm of `Mamba2` State] Suppose $\boldsymbol{\Lambda} = \lambda\boldsymbol{I} = \exp(-\Delta\alpha)\boldsymbol{I}$ is a scalar multiple of the identity matrix, where $\lambda \in (0, 1)$, a moderate discretized step value Δ and the system evolves as $\boldsymbol{h}_t = \boldsymbol{\Lambda}\boldsymbol{h}_{t-1} + \overline{\boldsymbol{B}}\boldsymbol{x}_t = \exp(-\Delta\alpha) \odot \boldsymbol{h}_{t-1} + \Delta\boldsymbol{B}\boldsymbol{x}_t$. Then the convergence rate ρ of the expected squared norm of the limiting state can be estimated as $\mathcal{O}(\frac{\Delta\cdot\lambda}{1-\lambda})$.

Proof. Given Theorem 4.2, the convergence rate ρ of `Mamba2` state can be estimated as $\delta = |\lambda_{max} - \lambda_{min}| \to 0$:

$$\rho = \lim_{\substack{t\to\infty \\ \delta\to 0}} \frac{\mathbb{E}[\|\boldsymbol{h}_t\|^2]}{\mathbb{E}[\|\overline{\boldsymbol{B}}\boldsymbol{x}\|^2]} = \lim_{\substack{t\to\infty \\ \delta\to 0}} \frac{\Delta\mathbb{E}[\|\boldsymbol{h}_t\|^2]}{\mathbb{E}[\|\boldsymbol{B}\boldsymbol{x}\|^2]} = \lim_{\delta\to 0} \frac{\Delta}{2\delta}\log\left(\frac{1-\lambda_{min}^2}{1-(\lambda_{min}+\delta)^2}\right) \tag{20}$$

Let $\lambda_{\min} = \lambda, \quad \lambda_{\max} = \lambda + \delta, \quad \delta \to 0$

Substitute into the expression:

$$\rho = \lim_{\delta\to 0} \frac{\Delta}{2\delta}\log\left(\frac{1-\lambda^2}{1-(\lambda+\delta)^2}\right) = \lim_{\delta\to 0} \frac{\Delta}{2\delta}\log\left(\frac{1-\lambda^2}{1-\lambda^2-2\lambda\delta-\delta^2}\right) \tag{21}$$

Let $\lambda = \lambda_{\min}$, $\delta = \lambda_{\max} - \lambda$, then:

https://doi.org/10.52202/085713-0653

$$= \lim_{\delta \to 0} \frac{\Delta}{2\delta} \log \left(\frac{1-\lambda^2}{1-(\lambda+\delta)^2} \right) \tag{22}$$

$$= \lim_{\delta \to 0} \frac{\Delta}{2\delta} \log \left(\frac{1-\lambda^2}{1-\lambda^2-2\lambda\delta-\delta^2} \right) \tag{23}$$

$$= \lim_{\delta \to 0} \frac{\Delta}{2\delta} \log \left(1 + \frac{2\lambda\delta+\delta^2}{1-\lambda^2} \right) \tag{24}$$

$$\approx \lim_{\delta \to 0} \frac{\Delta}{2\delta} \cdot \frac{2\lambda\delta+\delta^2}{1-\lambda^2} \tag{25}$$

$$= \frac{\Delta\lambda}{1-\lambda^2} \tag{26}$$

□

B Elaboration on Theorem 4.2

The divergent convergence behavior is irrespective of spectral assumptions. We provide an elaboration on Theorem 4.2 and corresponding convergence analysis without imposing any strict distribution assumptions.

B.1 Setup and notation

Let $d, m \in \mathbb{N}$. For $i = 1, \ldots, d$ let $\lambda_i \in (0,1)$ be i.i.d. samples from a density $p(\lambda)$ supported on $[\lambda_{\min}, \lambda_{\max}] \subset (0,1)$. Denote

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^{d \times d}.$$

Consider the linear system

$$\mathbf{h}_t = \Lambda \mathbf{h}_{t-1} + \mathbf{B}\mathbf{x}_t, \qquad t \geq 1, \tag{27}$$

where $\{\mathbf{x}_t\}$ are i.i.d. $\mathcal{N}(0, I_m)$ and $\mathbf{B} \in \mathbb{R}^{d \times m}$ has i.i.d. rows $\mathbf{b}_1, \ldots, \mathbf{b}_d$, each distributed as

$$\mathbf{b}_i \sim \mathcal{N}\big(0, \Sigma_B\big),$$

with $\Sigma_B \in \mathbb{R}^{m \times m}$ a (given) covariance matrix.

Assume $\mathbf{h}_0 = 0$ (or any initial condition that decays under Λ). We are interested in the steady-state expected squared norm

$$\mathbb{E}\big[\|\mathbf{h}_\infty\|^2\big] := \lim_{t \to \infty} \mathbb{E}\big[\|\mathbf{h}_t\|^2\big],$$

where the expectation is over the driving noise $\{\mathbf{x}_t\}$ and the random matrix $\mathbf{B}$ and the random eigenvalues $\{\lambda_i\}$.

B.2 Theorem (Expected state norm under general spectral distribution)

Theorem B.1. *Under the assumptions above, the limit $\mathbb{E}[\|\mathbf{h}_\infty\|^2]$ exists and equals*

$$\boxed{\mathbb{E}\big[\|\mathbf{h}_\infty\|^2\big] = \mathbb{E}\big[\|\mathbf{B}\mathbf{x}\|^2\big] \cdot \int_{\lambda_{\min}}^{\lambda_{\max}} \frac{p(\lambda)}{1-\lambda^2}\, d\lambda,} \tag{28}$$

where $\mathbf{x} \sim \mathcal{N}(0, I_m)$ is independent of $\mathbf{B}$ and λ, and the expectation on the left of the product is taken over $\mathbf{B}$ and $\mathbf{x}$.

Proof. Because Λ is diagonal and $\mathbf{x}_t \sim \mathcal{N}(0, I_m)$ i.i.d., the process equation 27 is Gaussian with zero mean for all t. Let

$$\Sigma_t := \mathbb{E}[\mathbf{h}_t \mathbf{h}_t^\top] \in \mathbb{R}^{d \times d}$$

be the (time-t) covariance of the hidden state. From equation 27 we have the Lyapunov-type recursion [108, 2]

$$\Sigma_t = \Lambda \Sigma_{t-1} \Lambda + \mathbb{E}[\mathbf{B}\mathbf{x}_t \mathbf{x}_t^\top \mathbf{B}^\top].$$

Since $\mathbf{x}_t\mathbf{x}_t^\top$ has expectation I_m and is independent of $\mathbf{B}$ and Λ, the driving covariance is

$$Q := \mathbb{E}[\mathbf{B}\mathbf{x}_t\mathbf{x}_t^\top\mathbf{B}^\top] = \mathbb{E}[\mathbf{B}\mathbf{B}^\top].$$

Because Λ is diagonal, the steady-state covariance $\Sigma_\infty := \lim_{t\to\infty}\Sigma_t$ is also diagonal; denote its diagonal entries by $s_i := (\Sigma_\infty)_{ii}$, $i = 1, \ldots, d$. The scalar recurrence for each diagonal entry is

$$s_i = \lambda_i^2 s_i + Q_{ii},$$

hence (since $|\lambda_i| < 1$)

$$s_i = \frac{Q_{ii}}{1-\lambda_i^2}.$$

Therefore the steady-state expected squared norm equals

$$\mathbb{E}\big[\|\mathbf{h}_\infty\|^2\big] = \mathbb{E}\big[\mathrm{trace}(\Sigma_\infty)\big] = \mathbb{E}\left[\sum_{i=1}^{d}\frac{Q_{ii}}{1-\lambda_i^2}\right].$$

Because the pairs (Q_{ii}, λ_i) are i.i.d. across i and rows of $\mathbf{B}$ are independent of the λ_i's, we have for a generic row $\mathbf{b} \in \mathbb{R}^m$

$$\mathbb{E}\big[\|\mathbf{h}_\infty\|^2\big] = d\,\mathbb{E}\left[\frac{\|\mathbf{b}\|^2}{1-\lambda^2}\right] = d\,\mathbb{E}[\|\mathbf{b}\|^2]\,\mathbb{E}\left[\frac{1}{1-\lambda^2}\right],$$

where λ is a generic draw from $p(\lambda)$ and independence between $\mathbf{b}$ and λ was used to factor the expectation.

Observe that

$$\mathbb{E}\big[\|\mathbf{B}\mathbf{x}\|^2\big] = \mathbb{E}_{\mathbf{B}}\mathbb{E}_{\mathbf{x}}\big[\|\mathbf{B}\mathbf{x}\|^2\big] = \mathbb{E}_{\mathbf{B}}\big[\mathrm{trace}(\mathbf{B}^\top\mathbf{B})\big] = \mathbb{E}_{\mathbf{B}}\Big[\sum_{i=1}^{d}\|\mathbf{b}_i\|^2\Big] = d\,\mathbb{E}[\|\mathbf{b}\|^2].$$

Combining the last two displayed equalities yields equation 28:

$$\mathbb{E}\big[\|\mathbf{h}_\infty\|^2\big] = \mathbb{E}\big[\|\mathbf{B}\mathbf{x}\|^2\big]\,\mathbb{E}\left[\frac{1}{1-\lambda^2}\right] = \mathbb{E}\big[\|\mathbf{B}\mathbf{x}\|^2\big]\,\int_{\lambda_{\min}}^{\lambda_{\max}}\frac{p(\lambda)}{1-\lambda^2}\,d\lambda.$$

This completes the proof. □

B.3 Evaluation of $\mathbb{E}\big[\|\mathbf{B}\mathbf{x}\|^2\big]$ in the isotropic row case

If each row $\mathbf{b}_i \sim \mathcal{N}(0, \sigma_B^2 I_m)$ (i.i.d. across rows), then

$$\mathbb{E}[\|\mathbf{b}\|^2] = \mathrm{trace}(\sigma_B^2 I_m) = m\sigma_B^2, \qquad \mathbb{E}\big[\|\mathbf{B}\mathbf{x}\|^2\big] = d\,m\,\sigma_B^2.$$

In the special (informal) normalization used in the statement above, where each row has covariance $\Sigma_B = (1/\sqrt{d})I_m$, one has $\sigma_B^2 = 1/\sqrt{d}$ and hence

$$\mathbb{E}\big[\|\mathbf{B}\mathbf{x}\|^2\big] = dm\frac{1}{\sqrt{d}} = md^{1/2}.$$

B.4 Asymptotic analysis of the integral

Define

$$I(\lambda_{\min}, \lambda_{\max}) := \int_{\lambda_{\min}}^{\lambda_{\max}}\frac{p(\lambda)}{1-\lambda^2}\,d\lambda = \mathbb{E}\left[\frac{1}{1-\lambda^2}\right].$$

(1) As $\lambda_{\max} \to 1^-$. Near $\lambda = 1$ we have the expansion $1 - \lambda^2 = (1-\lambda)(1+\lambda) \approx 2(1-\lambda)$. Suppose p is continuous at $\lambda = 1$ and $p(1) > 0$. For λ close to 1,

$$\frac{p(\lambda)}{1-\lambda^2} \sim \frac{p(1)}{2}\cdot\frac{1}{1-\lambda}.$$

Hence for $\lambda_{\max}$ sufficiently close to 1,

$$I(\lambda_{\min}, \lambda_{\max}) = \int_{\lambda_{\min}}^{\lambda_{\max}}\frac{p(\lambda)}{1-\lambda^2}\,d\lambda \approx \frac{p(1)}{2}\int_{\lambda_{\min}}^{\lambda_{\max}}\frac{1}{1-\lambda}\,d\lambda = \frac{p(1)}{2}\log\Big(\frac{1-\lambda_{\min}}{1-\lambda_{\max}}\Big).$$

Therefore

$$I(\lambda_{\min}, \lambda_{\max}) \sim -\frac{p(1)}{2}\log(1-\lambda_{\max}) \quad \text{as } \lambda_{\max} \to 1^-,$$

and in particular $I(\lambda_{\min}, \lambda_{\max}) \to +\infty$ with logarithmic divergence.

https://doi.org/10.52202/085713-0653

(2) As $\lambda_{\max} \to 0^+$. When $\lambda_{\max}$ is small, $1 - \lambda^2 \approx 1$, so the integrand is approximately $p(\lambda)$. If p is continuous near 0 with $p(0) > 0$, then

$$I(\lambda_{\min}, \lambda_{\max}) \approx \int_{\lambda_{\min}}^{\lambda_{\max}} p(\lambda)\, d\lambda \approx p(0) \cdot (\lambda_{\max} - \lambda_{\min}).$$

If we also take $\lambda_{\min} = 0$ or consider the leading-order scaling in $\lambda_{\max}$, then

$$I(\lambda_{\min}, \lambda_{\max}) \sim p(0)\, \lambda_{\max} \quad \text{as } \lambda_{\max} \to 0^+,$$

i.e., I vanishes linearly with $\lambda_{\max}$.

Combining the prefactor $\mathbb{E}[\|\mathbf{Bx}\|^2]$ from equation 28 with the asymptotics above yields

$$\boxed{\mathbb{E}\big[\|\mathbf{h}_\infty\|^2\big] \sim \begin{cases} -\dfrac{p(1)}{2}\, \mathbb{E}\big[\|\mathbf{Bx}\|^2\big]\ \log(1 - \lambda_{\max}), & \lambda_{\max} \to 1^-, \\ p(0)\, \mathbb{E}\big[\|\mathbf{Bx}\|^2\big]\ \lambda_{\max}, & \lambda_{\max} \to 0^+. \end{cases}}$$

C Additional Experimental Details and Results

C.1 Technical Details

All experiments were conducted on a single machine with 2 NVIDIA RTX4080 16GB GPUs. Experiments were run in an environment using CUDA version 12.6 and PyTorch 2.6.0.

C.2 Constant Scaling Language Modeling Perplexity

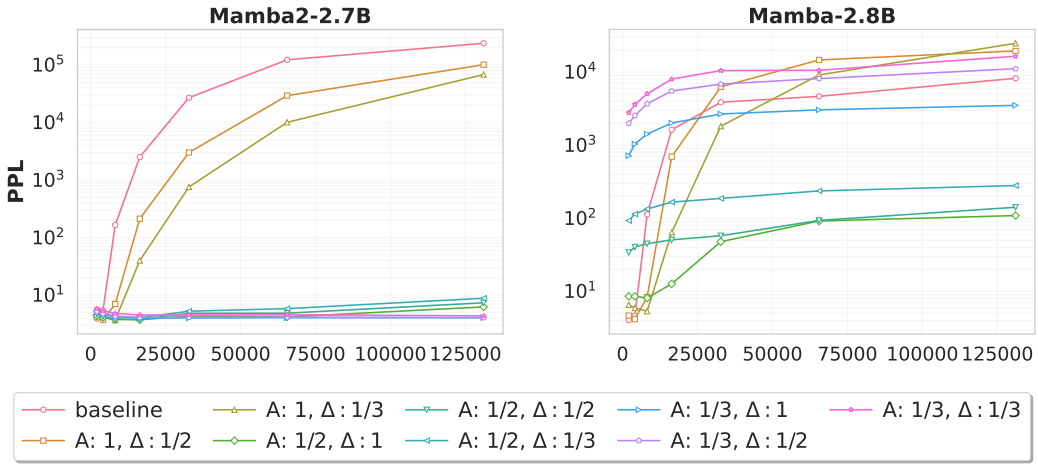

Figure 5: Language modeling perplexity on ProofPile after applying a constant scaling factor to either $\boldsymbol{A}$ or Δ_t. The red line with 'o' mark indicates the baseline, where neither $\boldsymbol{A}$ nor Δ_t are scaled.

https://doi.org/10.52202/085713-0653

C.3 `MambaExtend` Calibration

Here, we give an overview of the calibration functions we use within our `MambaExtend`-based experiments. Each of the described methods replace the calibration function CF within Algorithm 1. In our explicit implementation for calibrating scaling factors for $\boldsymbol{A}$, we use the same hyperparameters as Azizi et al. [3].

Calibration via back-propagation. To train the un-frozen calibration parameters on a calibration set, we apply a back-propagation algorithm to find the optimal scaling factors. This is described in Algorithm 2.

Algorithm 2 Calibration via back-propagation

Input: Frozen model $\mathcal{M}$, calibration set $\mathcal{C}$, initial scaling factors $\boldsymbol{S}$. Learning rate η, perturbation magnitude c, iterations K
Output: Learned scaling factors $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L] \in \mathbb{R}_+^{d_s \times L}$
$\texttt{optimizer} = \mathrm{Adam}(\boldsymbol{S}, \eta)$
for $k \leq K$ **do**
    $\mathcal{L} = \mathsf{eval}\left(\mathcal{M}_{c \times \boldsymbol{S}^+}, \mathcal{C}\right)$
    $\mathcal{L}.\mathsf{backward}\left(\right)$
    $\texttt{optimizer}.\mathsf{step}\left(\right)$
    $\boldsymbol{S} \leftarrow \mathsf{clamp}\left(\boldsymbol{S}, 0.001\right)$
end for
return $\boldsymbol{S}$

Calibration via zeroth-order optimization. Zeroth-order optimization offers an efficient yet noisier method for calibration, as it relies solely on forward passes to approximate gradients. Specifically, this is a multi-iteration process in which, at each iteration, the scaling factors are randomly perturbed using a random variable δ sampled from a Rademacher distribution. The magnitude of the perturbation and the learning rate for the updates are controlled by the hyper-parameters c and η, respectively. We employ the two-sided variant of the simultaneous perturbation stochastic approximation method (SPSA) [110], which obtains gradient approximations by applying both positive and negative perturbations to the parameters simultaneously. The two-sided SPSA approach yields gradient estimates with lower variance than the one-sided version, thus enhancing accuracy, especially in noisy environments [119]. This is described in Algorithm 3.

Algorithm 3 Calibration via zeroth-order optimization

Input: Frozen model $\mathcal{M}$, calibration set $\mathcal{C}$, perturbation magnitude c, iterations K
Output: Learned scaling factors $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L] \in \mathbb{R}_+^{d_s \times L}$
for $k \leq K$ **do**
    $\delta \in \mathbb{R}^{d_s \times L} \sim \mathrm{Radamacher}()$
    $\boldsymbol{S}^+ = \boldsymbol{S} + c \times \delta$
    $\boldsymbol{S}^- = \boldsymbol{S} - c \times \delta$
    $\ell^+ = \mathsf{eval}(\mathcal{M}_{c \times \boldsymbol{S}^+}, \mathcal{C})$
    $\ell^- = \mathsf{eval}(\mathcal{M}_{c \times \boldsymbol{S}^-}, \mathcal{C})$
    $\hat{\nabla}_{\boldsymbol{S}} = (\ell^+ - \ell^-)/(2 \cdot c \cdot \delta)$
    $\boldsymbol{S} \leftarrow \boldsymbol{S} - \eta \cdot \hat{\nabla}_{\boldsymbol{S}}$
    $\boldsymbol{S} \leftarrow \mathsf{clamp}\left(\boldsymbol{S}, 0.001\right)$
end for
return $\boldsymbol{S}$

C.3.1 Language Modeling Perplexity

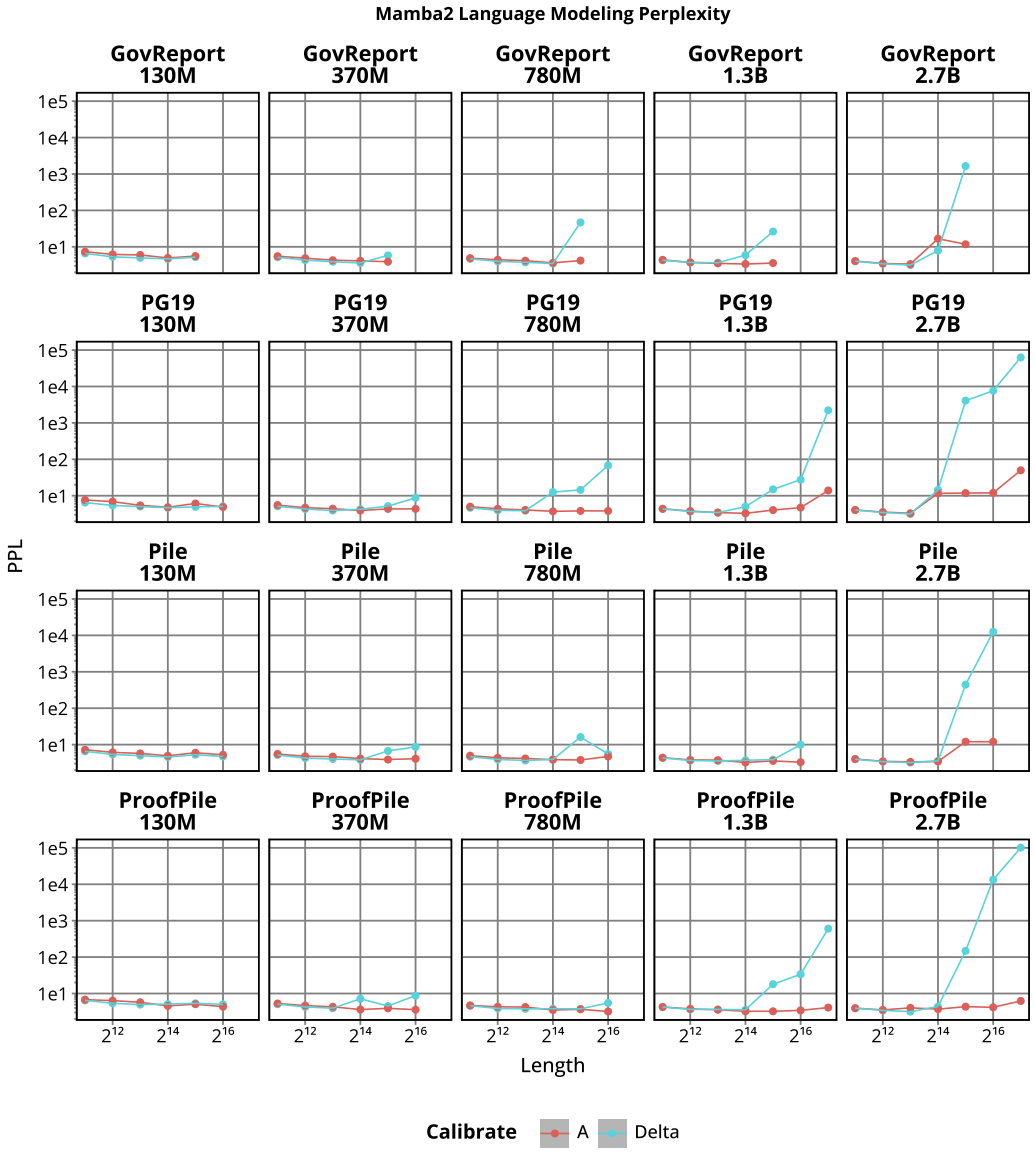

Figure 6: Language Model Perplexity performance of `Mamba2` models by calibrating scaling factors for either $\log(A)$ (red lines) or Δ_t (cyan lines). Perplexities are reported across various datasets (GovReport, PG19, ProofPile, Pile) as well as model sizes.

https://doi.org/10.52202/085713-0653

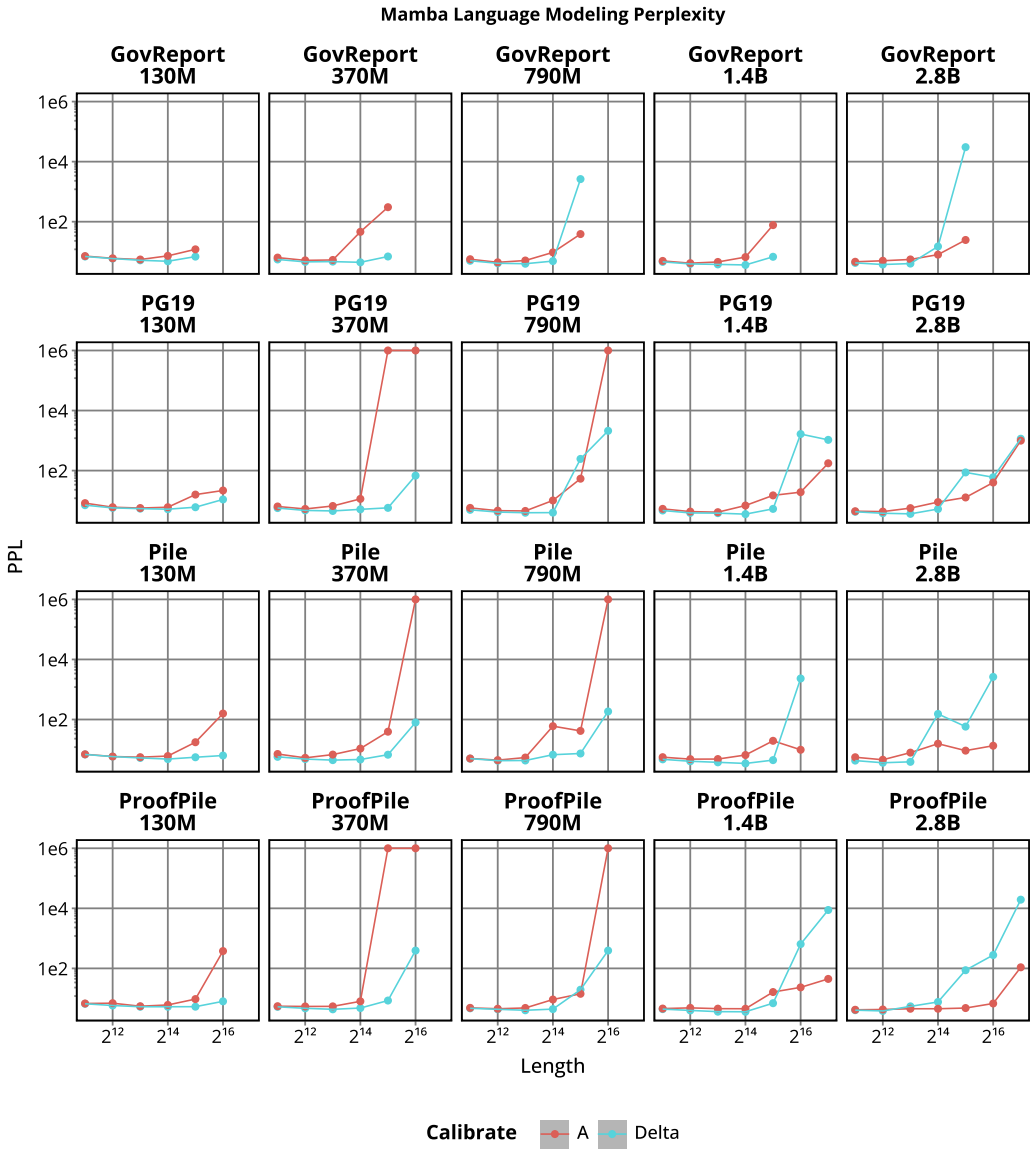

Figure 7: Language Model Perplexity performance of `Mamba` models by calibrating scaling factors for either $\log(A)$ (red lines) or Δ_t (cyan lines). Perplexities are reported across various datasets (GovReport, PG19, ProofPile, Pile) as well as model sizes.

https://doi.org/10.52202/085713-0653

C.3.2 Passkey Retrieval

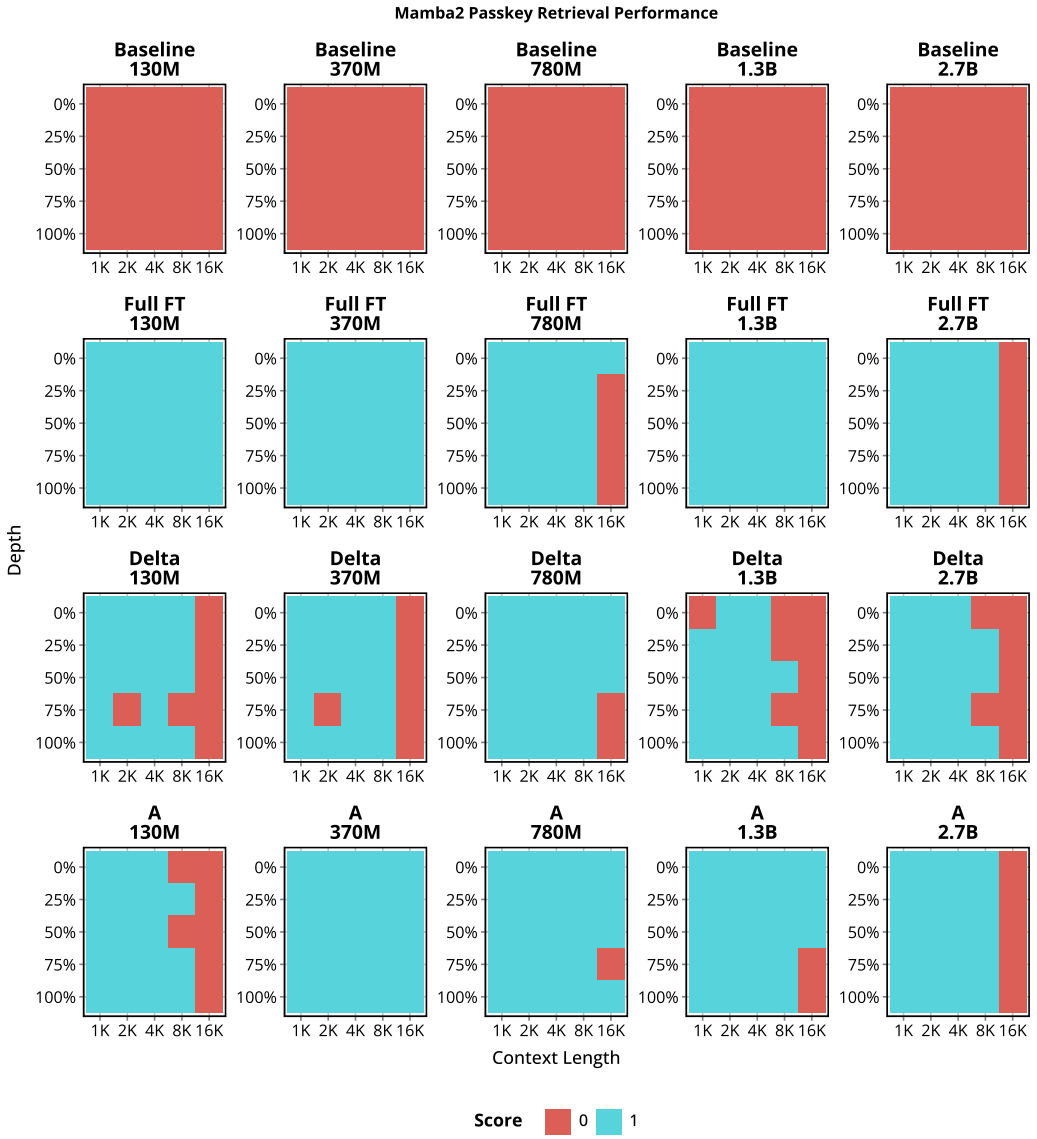

Figure 8: Passkey Retrieval performance of `Mamba2` models by calibrating scaling factors for either $\log(\boldsymbol{A})$ or Δ_t. Blue squares mean that the model was able to solve all examples of the given evaluation length/depth pair after tuning scaling factors, while red squares means that at least one mistake was made, i.e. an incorrect passage was retrieved.

https://doi.org/10.52202/085713-0653

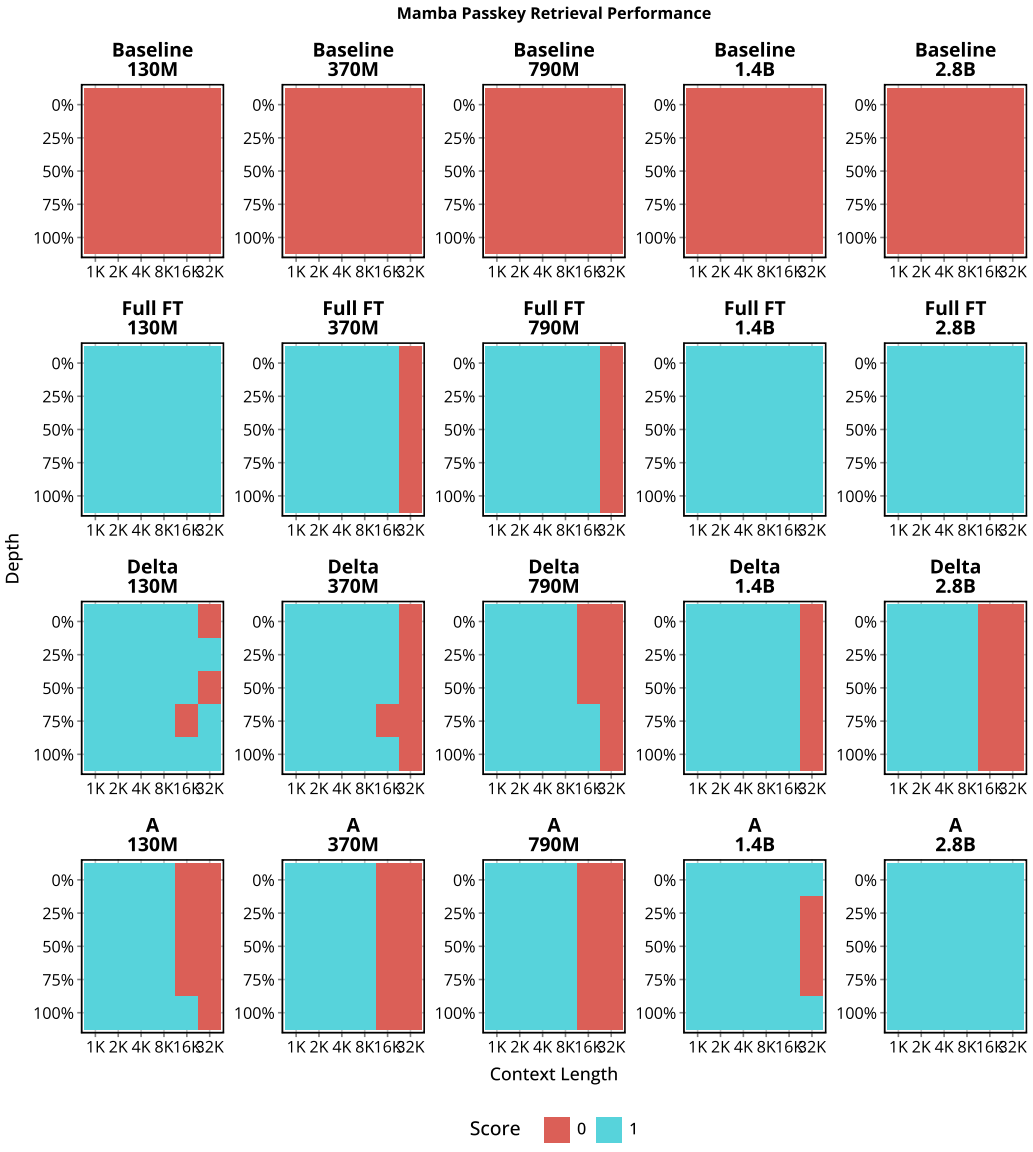

Figure 9: Passkey Retrieval performance of `Mamba` models by calibrating scaling factors for either $\log(\boldsymbol{A})$ or Δ_t. Blue squares mean that the model was able to solve all examples of the given evaluation length/depth pair after tuning scaling factors, while red squares means that at least one mistake was made, i.e. an incorrect passage was retrieved.

https://doi.org/10.52202/085713-0653

C.3.3 LongBench

We evaluate the following tasks from `LongBench` (Table 4). Due to our pre-training on an English dataset, we choose to use only the English language tasks included in the benchmark.

Table 4: Tasks from LongBench on which we evaluate.

Task	Context Type	Average Length	Metric	Data Samples
QASPERQA [23]	Science	3619	F1	200
HOTPOTQA [145]	Wikipedia	9151	F1	200
2WIKIMULTIQA [54]	Wikipedia	4887	F1	200
TREC [78]	Web Questions	5117	Accuracy	200
TRIVIAQA [70]	Wikipedia/Web	8209	F1	200
LCC [50]	Github	1235	Edit Similarity	500
REPOBENCH-P [82]	Github Repositories	4206	Edit Similarity	500

C.4 Pre-Trained Model Checkpoints Used

We use the official pre-trained model checkpoints of `Mamba` from the Hugging Face model Hub, found at `https://huggingface.co/state-spaces`. :

- `state-spaces/mamba-130m`
- `state-spaces/mamba-370m`
- `state-spaces/mamba-790m`
- `state-spaces/mamba-1.4b`
- `state-spaces/mamba-2.8b`
- `state-spaces/mamba2-130m`
- `state-spaces/mamba2-370m`
- `state-spaces/mamba2-780m`
- `state-spaces/mamba2-1.3b`
- `state-spaces/mamba2-2.7b`

We use the original `Mamba`[5] implementations for adjusting scaling parameters.

[5] `https://github.com/state-spaces/mamba`

https://doi.org/10.52202/085713-0653

D Broader Impacts

This work explores a novel method for length generalization of Mamba-based language models. While the direct usage of such models can entail potential broader risks within AI-based systems if potentially trained to scale, these risks do not stem directly from the methods and analysis presented within the paper. As such, there are no risks that are deemed significant and worthy of further discussion.

E Limitations

The primary limitation of our current work is that it is focused on Mamba-style models; as such, the methodology requires adaptation to similar models that utilize state-transition dynamics. However, many analogies exist between how information is written to the state and read out from the memory, presenting a potential avenue for use in other such models.

Another potential limitation is the lack of instruction-tuned models that are available for direct use, limiting the set of experiments and evaluations on which can be adequately conducted.