
Mamba-Modulation

On the Length Generalization of Mamba

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 The quadratic complexity of the attention mechanism in Transformer models has
2 motivated the development of alternative architectures with sub-quadratic scaling,
3 such as state-space models. Among these, Mamba has emerged as a leading
4 architecture, achieving state-of-the-art results across a range of language modeling
5 tasks. However, Mamba’s performance significantly deteriorates when applied to
6 contexts longer than those seen during pre-training, revealing a sharp sensitivity to
7 context length extension. Through detailed analysis, we attribute this limitation
8 to the out-of-distribution behavior of its state-space dynamics, particularly within
9 the parameterization of the state transition matrix \mathbf{A} . Unlike recent works which
10 attribute this sensitivity to the vanished accumulation of discretization time steps,
11 $\exp(-\sum_{t=1}^N \Delta_t)$, we establish a connection between state convergence behavior
12 as the input length approaches infinity and the spectrum of the transition matrix
13 \mathbf{A} , offering a well-founded explanation of its role in length extension. Next, to
14 overcome this challenge, we propose an approach that applies spectrum scaling
15 to pre-trained Mamba models to enable robust long-context generalization by
16 selectively modulating the spectrum of \mathbf{A} matrices in each layer. We show that this
17 can significantly improve performance in settings where simply modulating Δ_t
18 fails, validating our insights and providing avenues for better length generalization
19 of state-space models with structured transition matrices.

20 1 Introduction

21 In the new age of deep learning, the Transformers [54] architecture has spurred a new age of research
22 into large language models (LLMs) [16, 58, 14, 40] that has largely dominated the space of natural
23 language processing (NLP) research since their introduction. Their surprising capabilities and rapid
24 development have led to their wide application across various domains, including chatbots, intelligent
25 agents, code assistants, etc. However, the Transformer comes with various deficiencies, which has
26 led to research into alternative paradigms that seek to resolve these outstanding concerns. One of
27 these competitors, Mamba [21, 12], is based off the state-space model (SSM) paradigm from control
28 theory [23, 24] that have enabled the training of recurrent models that have overcome the sequential
29 bottleneck of traditional models [47, 28, 9].

30 A primary motivation for Mamba and its successors is the claim of length extrapolation, whereby a
31 model that is initially trained on a limited context length (e.g. 2048 tokens in each training sequence)
32 is capable of generalizing to longer sequences at test time (i.e. without further training) due to a
33 more efficient inference-time token processing methodology. However, various works [29, 15, 30]
34 have brought about challenges to this claim. Meanwhile, a key component in the Transformer is the
35 position embedding, for which Rotary position embeddings (RoPE) [51] has been a popular choice
36 and applied in many LLMs. Various works have studied RoPE [57, 38] and shown it to be intuitive to

manipulate to extend the context window within Transformers [8, 7, 18, 42], whereas no equivalent method yet exists for Mamba-style models. A common explanation for the ability to extend this context length is through avoiding out-of-distribution (OOD) rotation angles [38, 26] 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 [62, 6, 2] have meanwhile made attempts at exploring how to conduct length generalization with Mamba models. A shared feature of all of these works is the focus on a specific input-dependent parameter, Δ , which is used to discretize the underlying state-space model and control for the decay within the state as well as the contribution of the new input to the state. Their observations rely on the implicit notion that since the size of the time-step will influence the state, one can use it as a proxy to filter out (or ignore) parts of the input, or scale it in order to influence the long-term information decay within the model state. However, despite the compelling intuition behind such a notion, there remains a gap in truly understanding if this is well motivated.

In this work, we attempt to build a better fundamental 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 when input length goes to infinity. From this, we identify two ways in which this process can be controlled: either through the state-transition matrix A , or through the discretization time-steps Δ . We then motivate why controlling for or adjusting A is a more reasonable process. Through experiments on standard long-context extension settings, such as long-context language modeling and passkey retrieval, we demonstrate empirically how scaling A is more effective compared to scaling Δ , in the case of both Mamba and Mamba2 models. Broadly, 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 A as opposed to the more common practice of scaling Δ is a more effective proposition.
- III) We demonstrate on a series of long-context generalization tasks that such an intuition holds empirically on both Mamba and Mamba2 models, highlighting the potential benefits of using 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 [53]. Early work in efficient Transformers [34, 63, 5, 55, 10] attempted to reduce the computational complexity of attention by inducing greater sparsity. Additional work has explored the use of linear attention [33, 59, 60, 64] to remove the softmax activation that induces this quadratic complexity. Furthermore, hardware optimizations for more efficient computation [12, 11, 49, 36] as well as inference-time acceleration methods [56] to reduce the computational and memory complexity of Transformers. However, a broader class of linear recurrent models [24, 21, 41, 4, 44, 45], 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 [52, 38, 57], 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 [7, 18, 42, 8] 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, 43], 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 [29].

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 [6, 62, 2] 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. [6] 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. [62] use the value of the time-steps to filter out tokens. Azizi et al. [2] 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. [22] 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. [41] further demonstrated how the eigenvalues have a specific influence on state decay as well as long-term dynamics during training. Beck et al. [4] further bound the state of the recurrence, implicitly bounding the spectrum as well. Finally, Grazzi et al. [20] 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) [24] and Mamba [21] are inspired by the continuous system, which maps a 1-D function or sequence $\mathbf{x}(t) \in \mathbb{R}^{d_m}$ to an output $\mathbf{y}(t) \in \mathbb{R}^{d_m}$ through a hidden state $\mathbf{h}(t) \in \mathbb{R}^{d_h}$. The system uses evolution parameters $\mathbf{A} \in \mathbb{R}^{d_h \times d_h}$, \mathbf{B} , and $\mathbf{C} \in \mathbb{R}^{d_h \times d_m}$, creating a continuous system whose dynamics are governed by

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

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 $\Delta(x_t)$ to transform the continuous parameters \mathbf{A} , \mathbf{B} to discrete parameters $\bar{\mathbf{A}}_t$ and $\bar{\mathbf{B}}_t$. We follow the official implementation of Mamba [21]:

$$\mathbf{h}_{t+1} = \bar{\mathbf{A}}_t \mathbf{h}_t + \bar{\mathbf{B}}_t \mathbf{x}_t, \quad \mathbf{y}_t = \mathbf{C}_t \mathbf{h}_t, \quad (2)$$

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

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

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

$$\Delta_t = \text{softplus}(\text{Linear}_\Delta(x_t)), \quad \mathbf{B}_t = \text{Linear}_B(x_t), \quad \mathbf{C}_t = \text{Linear}_C(x_t) \quad (4)$$

where Linear_Δ , Linear_B and Linear_C are regular linear projections, \odot is the Hadamard product, \otimes is the outer product, $\Delta_t \in \mathbb{R}_+^d$ and $A = \text{diag}(\alpha_1, \dots, \alpha_d)$ with each $\alpha > 0$. Mamba makes Δ , B and C to be input dependent, such that at each time-step unique transition matrices can be used to update the system (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 $(\Delta_t, \bar{A}_t, \bar{B}_t, \bar{C}_t)$, enabling the ability to solve problems that require selective processing of the sequence. In order to maintain computational efficiency A is restricted to having a diagonal structure such that only the diagonal elements of these matrices need to be stored. Mamba2 [12] 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:

$$Y = MX, \quad M_{i,j} = C_i \left(\prod_{t=j+1}^i \bar{A}_t \right) \bar{B}_j \quad (5)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_L \end{bmatrix} = \begin{bmatrix} C_1 \bar{B}_1 & 0 & \cdots & 0 \\ C_2 \bar{A}_2 \bar{B}_1 & C_2 \bar{B}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ C_L \prod_{t=2}^L \bar{A}_t \bar{B}_1 & C_L \prod_{t=3}^L \bar{A}_t \bar{B}_2 & \cdots & C_L \bar{B}_L \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_L \end{bmatrix} \quad (6)$$

In this formulation, each output y_t is computed as a weighted sum of all inputs, with each weight involving a product of transition matrices $\prod_{t=j+1}^L \bar{A}_t$. 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.

$$\prod_{t=j+1}^L \bar{A}_t = \prod_{t=j+1}^L \exp(-A \Delta_t) = \exp \left(-A \sum_{t=j+1}^L \Delta_t \right) \quad (7)$$

Previous work [6, 62, 2] has primarily focused on analyzing the discretization step Δ_t for long context, particularly the vanishing effect of the accumulated term $\exp(-\sum_{t=1}^N \Delta_t)$ when N is large and propose different solutions to overcome this Out-of-Distribution (OOD) issue. For instance, Azizi et al. [2] propose applying scalar values $s < 1$ 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 Mamba’s 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 begin by visualizing the spectrum of the continuous transition matrix $\Lambda = \text{diag}(\exp(-A))$ of Mamba models. The $\exp(-A)$ parameterization guarantees that all values are bounded between 0 and 1. We stack all 64 layers row

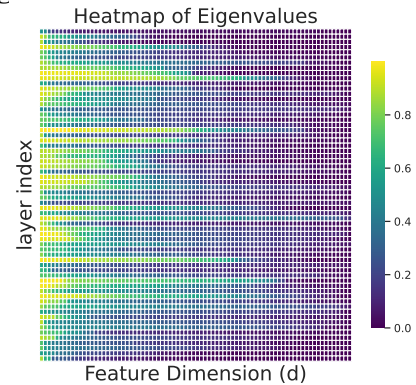


Figure 1: Heatmap of eigenvalues of the transition matrix $\text{diag}(\exp(-A))$ of mamba2-2.7b.

by row, and for each row, we rank the eigenvalues in descending order. The eigenvalue 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.

4.2 State Convergence in SSMs for Long Contexts

The previous section presented the numerical spectrum of the transition matrix $\exp(-\mathbf{A})$. 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 $\mathbf{B} \in \mathbb{R}^{d \times d}$ be a matrix with $\|\mathbf{B}\|_2 = \sigma_B$, and let $\mathbf{x} \in \mathbb{R}^d$ be a vector such that each entry of \mathbf{x} satisfies $|x_i| \leq \sigma_x$. The upper bound for $\|\mathbf{B}\mathbf{x}\|_2$ is:*

$$\|\mathbf{B}\mathbf{x}\|_2 \leq \sigma_B \cdot \sigma_x \cdot \sqrt{d}. \quad (8)$$

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: $\mathbf{h}_t = \mathbf{\Lambda}\mathbf{h}_{t-1} + \mathbf{B}\mathbf{x}_t$ where \mathbf{x}_t is the input vector at time step t , and \mathbf{B} is a weight matrix whose rows are independently sampled as $\mathbf{b} \sim \mathcal{N}(0, \frac{1}{2d}\mathbf{I})$. Then, as $t \rightarrow \infty$, the expected squared norm of the state \mathbf{h}_t converges to:*

$$\mathbb{E}[\|\mathbf{h}_\infty\|^2] = \frac{1}{2(\lambda_{\max} - \lambda_{\min})} \log \left(\frac{1 - \lambda_{\min}^2}{1 - \lambda_{\max}^2} \right) \cdot \mathbb{E}[\|\mathbf{B}\mathbf{x}\|^2]. \quad (9)$$

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

The preceding theorem and corollaries provide insight into the asymptotic convergence behavior of Mamba states as the input sequence length grows with different eigenvalues. If $\lambda \rightarrow 1$, or $\lambda \rightarrow 0$, then

$$\lim_{\lambda \rightarrow 1^-} \rho = \infty, \quad \lim_{\lambda \rightarrow 0^+} \rho = 0 \quad (10)$$

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 increases—leading to state explosion or vanishing, respectively. While tuning the discretization step Δ 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 model’s ability to generalize over input length.

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 B provides more specific implementation details and design choices.

5.1 A Simple Case Analysis on Constant Scaling

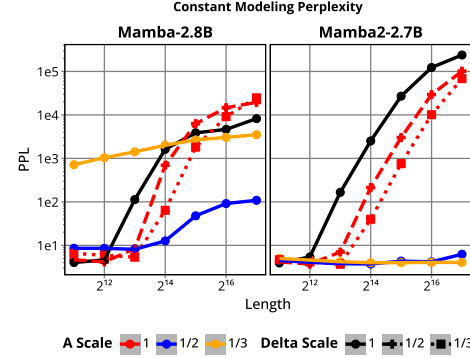


Figure 2: Language modeling perplexity on ProofPile after applying a constant scaling factor to either \mathbf{A} or Δ_t . The solid black line indicates the baseline, where no scaling is used. Red lines indicate Δ_t is scaled, while the other colors indicate \mathbf{A} was scaled.

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.

To confirm our intuition, we first attempt a simple comparison between the effects of scaling Δ_t and \mathbf{A} . Here, across all layers, we use a fixed, constant-valued scaling factor. We evaluate language modeling perplexity on the ProofPile dataset [19], following Peng et al. [42], 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 \mathbf{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

5.2 Adapting MambaExtend to Scale \mathbf{A}

Given our observations and analysis regarding the relationship between \mathbf{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 $\{s_i\}_{i=1}^L$ which can be used to adjust the discretizations steps Δ_t . In

general, 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 B.3 describes the implementation in further detail.

Although the original MambaExtend learns scaling factors only for Δ_t , their methodology is adaptable to usage with \mathbf{A} instead; given the shared dimensionality for both \mathbf{A} and Δ_t , the scaling factors can be directly used for calibrating \mathbf{A} . Furthermore, this means that tuning scaling factors for \mathbf{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 \mathbf{A} for long-context generalization. The following sections evaluates the performance and efficiency of tuning these scaling factors for \mathbf{A} on a number of standard settings for evaluating long-context generalization of models. As a baseline, we compare directly with the original MambaExtend.

Algorithm 1 MambaExtend methodology.

- 1: **Input:** Model \mathcal{M} , calibration set \mathcal{C} and function CF
 - 2: **Output:** Scaling factors $\mathbf{S} = [s_1, \dots, s_L] \in \mathbb{R}_+^{d_a \times L}$
 - 3: **for** $i \leq L$ **do**
 - 4: $s_i \leftarrow U(0, 1)$
 - 5: **end for**
 - 6: $\mathbf{S} \leftarrow \text{CF}(\mathbf{S}, \mathcal{C}, \mathcal{M})$
 - 7: **return** \mathbf{S}
-

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 \mathcal{A} . In this task, we use the black-box zeroth-order calibration method suggested by Azizi et al. [2]; 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 [19], PG19 [46] and GovReport [31].

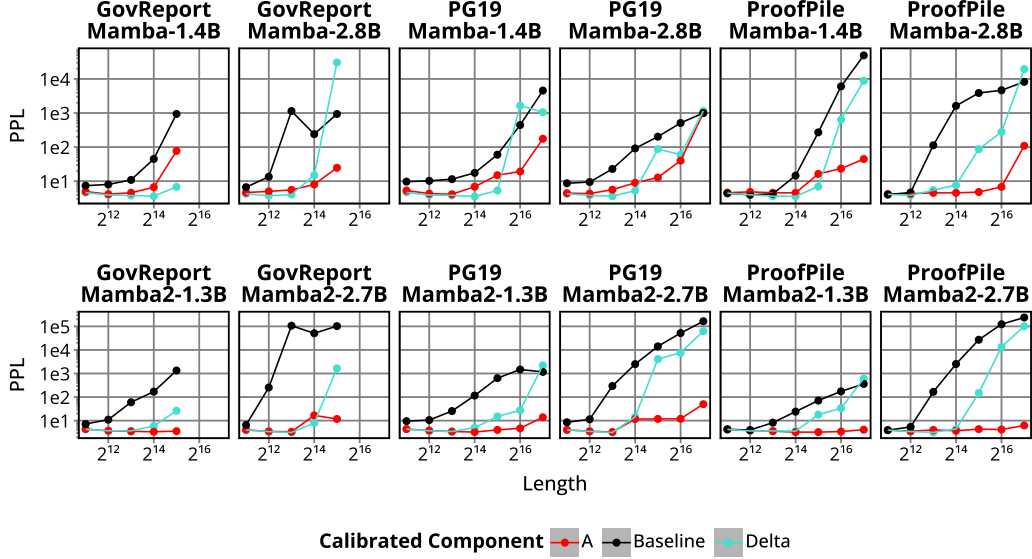


Figure 3: Model perplexity by calibrating scaling factors for either $\log(\mathcal{A})$ (red) or Δ_t (turquoise), across different datasets and sizes. **Baseline** means the base model with no calibration.

In particular, scaling \mathcal{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 \mathcal{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 \mathcal{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 [17]. 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. 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. [6], in particular, the task comprises of a 5-digit code embedded at a random sequence depth within samples from the WikiText-103 dataset [39]. 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.

Our results are visualized in Figure 4 and Appendix B.3.2. In particular, we see very consistent results similar to our language modeling perplexity results; for Mamba, smaller models appear to fare

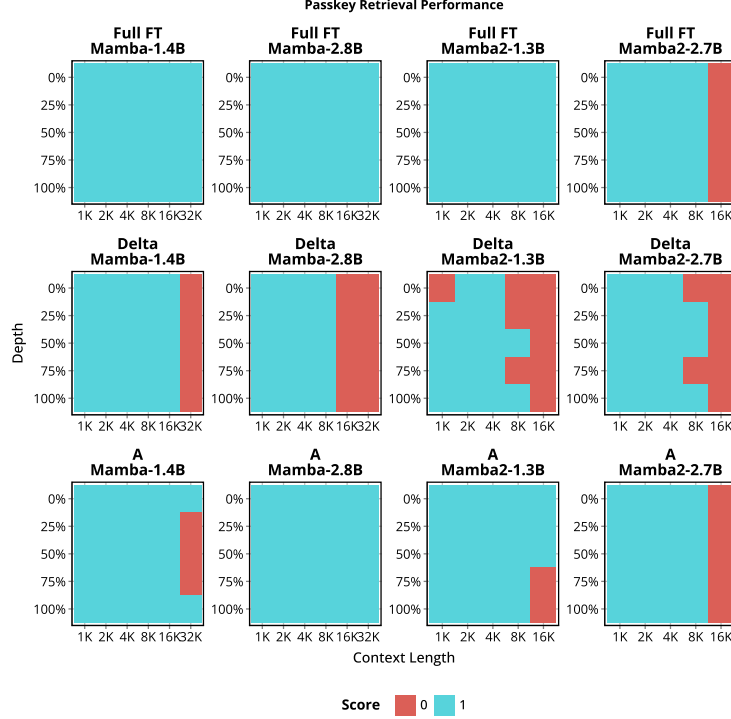


Figure 4: Passkey Retrieval performance of Mamba models by calibrating scaling factors for either $\log(\mathbf{A})$ or Δ_t . Blue squares mean that the model was able to solve all examples of the given evaluation length/depth pair after tuning scaling factors, while red squares means that at least one mistake was made, i.e., an incorrect passage was retrieved.

slightly better when trained to scale Δ_t , but as the models get larger, learning to scale \mathbf{A} closes the gap and eventually exceeds the performance of scaling Δ_t . Similarly, for Mamba2 models, scaling \mathbf{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 \mathbf{A} is as effective despite fewer parameters being trained, whereas scaling Δ_t does observe a drop-off in performance.

6.3 LongBench

LongBench [3] 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 \mathbf{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 1 shows results on Mamba2-2.7B. In particular, we show that we can increase performance by over 6% through the calibrated scaling of \mathbf{A} , with a relative improvement of nearly 10% compared to if the scaling was instead calibrated for Δ_t .

Table 1: Results on LongBench [3].

| Model | Strategy | Qasper | HotpotQA | 2WikiMultihopQA | TREC | TriviaQA | LCC | RepoBench-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 \mathbf{A} | 12.90 | 5.69 | 11.18 | 24.32 | 10.49 | 23.36 | 16.91 | 14.98 |

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

Table 2: Comparison of PG19 perplexity at varying lengths. Cases where scaling \mathbf{A} leads to the lowest perplexity are **bolded** and underlined when second best. If the best method does not involve scaling \mathbf{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 \mathbf{A} | 44.68 | 53.46 | 59.56 | 63.51 | 75.86 | 114.44 |
| Calibrated Scaling \mathbf{A} | <u>5.31</u> | <u>4.31</u> | <u>4.13</u> | <u>6.88</u> | <u>14.94</u> | 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 \mathbf{A} | 28.80 | 33.93 | 39.80 | 69.37 | 162.77 | 355.77 |
| Calibrated Scaling \mathbf{A} | <u>4.44</u> | <u>4.31</u> | <u>5.63</u> | <u>8.94</u> | 12.75 | <u>40.00</u> |
| 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 \mathbf{A} | 11.12 | 11.83 | 12.47 | 12.71 | 12.85 | <u>13.22</u> |
| Calibrated Scaling \mathbf{A} | <u>4.38</u> | <u>3.78</u> | 3.44 | 3.28 | 4.03 | 4.72 |

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 [62], 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 [6] 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 availability¹ and methodology², 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 \mathbf{A} , namely constant scaling as well as the calibrated scaling based on MambaExtend.

Our results in Table 2 compare the effectiveness of these different methods on perplexity on the PG19 dataset. We note that in all cases, the calibrated scaling of \mathbf{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 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 \mathbf{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 \mathbf{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 \mathbf{A} for length generalization.

¹LongMamba did not release their tuning code: <https://github.com/GATECH-EIC/LongMamba>

²DeciMamba only modified Mamba CUDA kernels: <https://github.com/assafbk/DeciMamba>

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] 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>.
- [3] 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>.
- [4] 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.
- [5] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The Long-Document Transformer, 2020. URL <https://arxiv.org/abs/2004.05150>. arXiv: 2004.05150.
- [6] 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>.
- [7] 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/141z7j5/ntkaware_scaled_rope_allows_llama_models_to_have/.
- [8] 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.
- [9] K. Cho, B. v. Merriënboer, 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>.
- [10] K. M. Choromanski, V. Likhoshervstov, 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>.
- [11] 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>.

- [12] 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>.
- [13] 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>.
- [14] 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.
- [15] 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>.
- [16] 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.
- [17] 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.
- [18] 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/.
- [19] 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.

- [20] 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>.
- [21] 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>.
- [22] 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>.
- [23] 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>.
- [24] 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=uYLFoz1v1AC>.
- [25] 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>.
- [26] 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>.
- [27] 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>.
- [28] 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>.
- [29] C.-P. Hsieh, S. Sun, S. Krizan, S. Acharya, D. Rekesh, F. Jia, and B. Ginsburg. RULER: What’s 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>.
- [30] 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*

- 501 *International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE,*
502 *January 19-24, 2025*, pages 29–39. Association for Computational Linguistics, 2025. URL
503 <https://aclanthology.org/2025.coling-main.3/>.
- 504 [31] L. Huang, S. Cao, N. N. Parulian, H. Ji, and L. Wang. Efficient Attentions for Long Document
505 Summarization. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Belt-
506 agy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021*
507 *Conference of the North American Chapter of the Association for Computational Linguistics:*
508 *Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1419–1436.
509 Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.112.
510 URL <https://doi.org/10.18653/v1/2021.naacl-main.112>.
- 511 [32] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer. Triviaqa: A large scale distantly supervised
512 challenge dataset for reading comprehension. In R. Barzilay and M. Kan, editors, *Proceedings*
513 *of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017,*
514 *Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association
515 for Computational Linguistics, 2017. URL <https://doi.org/10.18653/v1/P17-1147>.
- 516 [33] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are RNNs: Fast Au-
517 toregressive Transformers with Linear Attention. In *Proceedings of the 37th International*
518 *Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume
519 119 of *Proceedings of Machine Learning Research*, pages 5156–5165. PMLR, 2020. URL
520 <http://proceedings.mlr.press/v119/katharopoulos20a.html>.
- 521 [34] N. Kitaev, L. Kaiser, and A. Levskaya. Reformer: The Efficient Transformer. In *8th Interna-*
522 *tional Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30,*
523 *2020*. OpenReview.net, 2020. URL <https://openreview.net/forum?id=rkgNKKHtvB>.
- 524 [35] X. Li and D. Roth. Learning question classifiers. In *19th International Conference on Com-*
525 *putational Linguistics, COLING 2002, Howard International House and Academia Sinica,*
526 *Taipei, Taiwan, August 24 - September 1, 2002*, 2002. URL <https://aclanthology.org/C02-1150/>.
- 528 [36] H. Liu, M. Zaharia, and P. Abbeel. RingAttention with Blockwise Transformers for Near-Infinite
529 Context. In *The Twelfth International Conference on Learning Representations, ICLR 2024,*
530 *Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL [https://openreview.net/](https://openreview.net/forum?id=WsRHpHH4s0)
531 [forum?id=WsRHpHH4s0](https://openreview.net/forum?id=WsRHpHH4s0).
- 532 [37] T. Liu, C. Xu, and J. J. McAuley. Repobench: Benchmarking repository-level code auto-
533 completion systems. In *The Twelfth International Conference on Learning Representa-*
534 *tions, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL [https://openreview.net/](https://openreview.net/forum?id=pPjZIOuQuF)
535 [forum?id=pPjZIOuQuF](https://openreview.net/forum?id=pPjZIOuQuF).
- 536 [38] X. Liu, H. Yan, C. An, X. Qiu, and D. Lin. Scaling Laws of RoPE-based Extrapolation.
537 In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna,*
538 *Austria, May 7-11, 2024*. OpenReview.net, 2024. URL [https://openreview.net/](https://openreview.net/forum?id=J07k0SJ5V6)
539 [forum?id=J07k0SJ5V6](https://openreview.net/forum?id=J07k0SJ5V6).
- 540 [39] S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer Sentinel Mixture Models. In *5th*
541 *International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26,*
542 *2017, Conference Track Proceedings*. OpenReview.net, 2017. URL [https://openreview.](https://openreview.net/forum?id=Byj72udxe)
543 [net/forum?id=Byj72udxe](https://openreview.net/forum?id=Byj72udxe).
- 544 [40] OpenAI. GPT-4 Technical Report, 2023. URL [https://doi.org/10.48550/arXiv.2303.](https://doi.org/10.48550/arXiv.2303.08774)
545 [08774](https://doi.org/10.48550/arXiv.2303.08774). arXiv: 2303.08774.
- 546 [41] A. Orvieto, S. L. Smith, A. Gu, A. Fernando, Ç. Gülçehre, R. Pascanu, and S. De. Res-
547 urrecting Recurrent Neural Networks for Long Sequences. In A. Krause, E. Brunskill,
548 K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on*
549 *Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202
550 of *Proceedings of Machine Learning Research*, pages 26670–26698. PMLR, 2023. URL
551 <https://proceedings.mlr.press/v202/orvieto23a.html>.

- [42] 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>.
- [43] 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>.
- [44] 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*. URL http://papers.nips.cc/paper_files/paper/2023/hash/694be3548697e9cc8999d45e8d16fe1e-Abstract-Conference.html.
- [45] 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>.
- [46] 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>.
- [47] D. E. Rumelhart. *Parallel distributed processing, 9th Edition*. MIT Pr., 1989. ISBN 978-0-262-68053-0. URL <https://www.worldcat.org/oclc/60445750>.
- [48] 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.
- [49] 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/7ede97c3e082c6df10a8d6103a2eabd2-Abstract-Conference.html.
- [50] 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.
- [51] 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>.
- [52] 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>.
- [53] 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>.

- [54] 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>.
- [55] 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.
- [56] 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>.
- [57] 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.
- [58] 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.
- [59] 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>.
- [60] 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.
- [61] 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, 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>.
- [62] 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=fMbLszV01H>.
- [63] 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>.

- 655 [64] Y. Zhang, S. Yang, R.-J. Zhu, Y. Zhang, L. Cui, Y. Wang, B. Wang, F. Shi, B. Wang, W. Bi,
656 P. Zhou, and G. Fu. Gated Slot Attention for Efficient Linear-Time Sequence Modeling. In
657 A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang,
658 editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural*
659 *Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December*
660 *10 - 15, 2024*, 2024. URL [http://papers.nips.cc/paper_files/paper/2024/hash/](http://papers.nips.cc/paper_files/paper/2024/hash/d3f39e51f5f634fb16cc3e658f8512b9-Abstract-Conference.html)
661 [d3f39e51f5f634fb16cc3e658f8512b9-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2024/hash/d3f39e51f5f634fb16cc3e658f8512b9-Abstract-Conference.html).

662 A Proofs

663 A.1 Proof of Lemma 4.1

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

$$\|\mathbf{B}\mathbf{x}\|_2 \leq \sigma_B \cdot \sigma_x \cdot \sqrt{d}.$$

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

$$\|\mathbf{B}\mathbf{x}\|_2 \leq \|\mathbf{B}\|_2 \cdot \|\mathbf{x}\|_2.$$

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

$$\|\mathbf{B}\mathbf{x}\|_2 \leq \sigma_B \cdot \|\mathbf{x}\|_2.$$

668 And

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

669 Substituting the bound on $\|\mathbf{x}\|_2$ into the inequality for $\|\mathbf{B}\mathbf{x}\|_2$, we have the norm of logit vector
670 $\mathbf{u} \in \mathbb{R}^d$:

$$\|\mathbf{u}\|_2 = \|\mathbf{B}\mathbf{x}\|_2 \leq \sigma_B \cdot \|\mathbf{x}\|_2 \leq \sigma_B \cdot \sqrt{d} \cdot \sigma_x.$$

671 □

672 A.2 Proof of Theorem 4.2

673 **Theorem 4.2.** Assume the transition matrix $\mathbf{\Lambda}$ is diagonal with eigenvalues $\lambda_i \sim$
674 $\text{Uniform}[\lambda_{\min}, \lambda_{\max}]$ for $0 < \lambda_{\min} < \lambda_{\max} < 1$. Suppose the system evolves as

$$\mathbf{h}_t = \mathbf{\Lambda} \mathbf{h}_{t-1} + \mathbf{B} \mathbf{x}_t, \quad (11)$$

675 where $\mathbf{x}_t \sim \mathcal{N}(0, \mathbf{I})$ and \mathbf{B} is a weight matrix whose rows are independently sampled as $\mathbf{b} \sim$
676 $\mathcal{N}(0, \frac{1}{\sqrt{d}} \mathbf{I})$. Then, in the limit $t \rightarrow \infty$, the expected squared norm of the hidden state converges to

$$\mathbb{E}[\|\mathbf{h}_\infty\|^2] = \frac{1}{2(\lambda_{\max} - \lambda_{\min})} \log \left(\frac{1 - \lambda_{\min}^2}{1 - \lambda_{\max}^2} \right) \cdot \mathbb{E}[\|\mathbf{B}\mathbf{x}\|^2]. \quad (12)$$

677 *Proof.* We begin by unrolling the recurrence:

$$\mathbf{h}_t = \sum_{i=0}^{t-1} \mathbf{\Lambda}^i \mathbf{B} \mathbf{x}_{t-i}. \quad (13)$$

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

$$\mathbb{E}[\|\mathbf{h}_\infty\|^2] = \sum_{i=0}^{\infty} \mathbb{E}[\|\mathbf{\Lambda}^i \mathbf{B} \mathbf{x}\|^2]. \quad (14)$$

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

$$\mathbb{E}[h^2] = \sum_{i=0}^{\infty} \lambda^{2i} \mathbb{E}[\|\mathbf{b}\mathbf{x}\|^2] = \frac{\sigma^2}{1 - \lambda^2}, \quad (15)$$

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

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

$$\mathbb{E}[\|\mathbf{h}_\infty\|^2] = d \cdot \mathbb{E}_\lambda \left[\frac{\sigma^2}{1 - \lambda^2} \right] = \sigma^2 d \cdot \frac{1}{\lambda_{\max} - \lambda_{\min}} \int_{\lambda_{\min}}^{\lambda_{\max}} \frac{1}{1 - \lambda^2} d\lambda. \quad (16)$$

684 Evaluating the integral:

$$\int_{\lambda_{\min}}^{\lambda_{\max}} \frac{1}{1-\lambda^2} d\lambda = \frac{1}{2} \log \left(\frac{1-\lambda_{\min}^2}{1-\lambda_{\max}^2} \right). \quad (17)$$

685 Hence,

$$\mathbb{E}[\|\mathbf{h}_{\infty}\|^2] = \sigma^2 d \cdot \frac{1}{2(\lambda_{\max} - \lambda_{\min})} \log \left(\frac{1-\lambda_{\min}^2}{1-\lambda_{\max}^2} \right). \quad (18)$$

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

$$\mathbb{E}[\|\mathbf{h}_{\infty}\|^2] = \frac{1}{2(\lambda_{\max} - \lambda_{\min})} \log \left(\frac{1-\lambda_{\min}^2}{1-\lambda_{\max}^2} \right) \cdot \mathbb{E}[\|\mathbf{B}\mathbf{x}\|^2]. \quad (19)$$

687 \square

688 A.3 Proof of Corollary 4.3 and Corollary 4.4

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

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

$$\rho = \lim_{t \rightarrow \infty} \frac{\mathbb{E}[\|\mathbf{h}_t\|^2]}{\mathbb{E}[\|\mathbf{B}\mathbf{x}\|^2]} = \lim_{t \rightarrow \infty} \frac{\Delta \mathbb{E}[\|\mathbf{h}_t\|^2]}{\mathbb{E}[\|\mathbf{B}\mathbf{x}\|^2]} = \frac{\Delta}{2\lambda} \log \left(\frac{1}{1-\lambda^2} \right) \quad (20)$$

694 \square

695 **Corollary 4.4** [Norm of Mamba2 State] Suppose $\mathbf{\Lambda} = \lambda \mathbf{I} = \exp(-\Delta\alpha)\mathbf{I}$ is a scalar multiple of the
 696 identity matrix, where $\lambda \in (0, 1)$, a moderate discretized step value Δ and the system evolves as
 697 $\mathbf{h}_t = \mathbf{\Lambda}\mathbf{h}_{t-1} + \mathbf{B}\mathbf{x}_t = \exp(-\Delta\alpha) \odot \mathbf{h}_{t-1} + \Delta\mathbf{B}\mathbf{x}_t$. Then the convergence rate ρ of the expected
 698 squared norm of the limiting state can be estimated as $\mathcal{O}\left(\frac{\Delta \cdot \lambda}{1-\lambda}\right)$.

699 *Proof.* Given Theorem 4.2, the convergence rate ρ of Mamba2 state can be estimated as $\delta =$
 700 $|\lambda_{\max} - \lambda_{\min}| \rightarrow 0$:

$$\rho = \lim_{\delta \rightarrow 0} \frac{\mathbb{E}[\|\mathbf{h}_t\|^2]}{\mathbb{E}[\|\mathbf{B}\mathbf{x}\|^2]} = \lim_{\delta \rightarrow 0} \frac{\Delta \mathbb{E}[\|\mathbf{h}_t\|^2]}{\mathbb{E}[\|\mathbf{B}\mathbf{x}\|^2]} = \lim_{\delta \rightarrow 0} \frac{\Delta}{2\delta} \log \left(\frac{1-\lambda_{\min}^2}{1-(\lambda_{\min} + \delta)^2} \right) \quad (21)$$

701 Let $\lambda_{\min} = \lambda$, $\lambda_{\max} = \lambda + \delta$, $\delta \rightarrow 0$

702 Substitute into the expression:

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

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

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

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

$$= \lim_{\delta \rightarrow 0} \frac{\Delta}{2\delta} \log \left(1 + \frac{2\lambda\delta + \delta^2}{1 - \lambda^2} \right) \quad (25)$$

$$\approx \lim_{\delta \rightarrow 0} \frac{\Delta}{2\delta} \cdot \frac{2\lambda\delta + \delta^2}{1 - \lambda^2} \quad (26)$$

$$= \frac{\Delta\lambda}{1 - \lambda^2} \quad (27)$$

704

□

705 B Additional Experimental Details and Results

706 B.1 Technical Details

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

709 B.2 Constant Scaling Language Modeling Perplexity

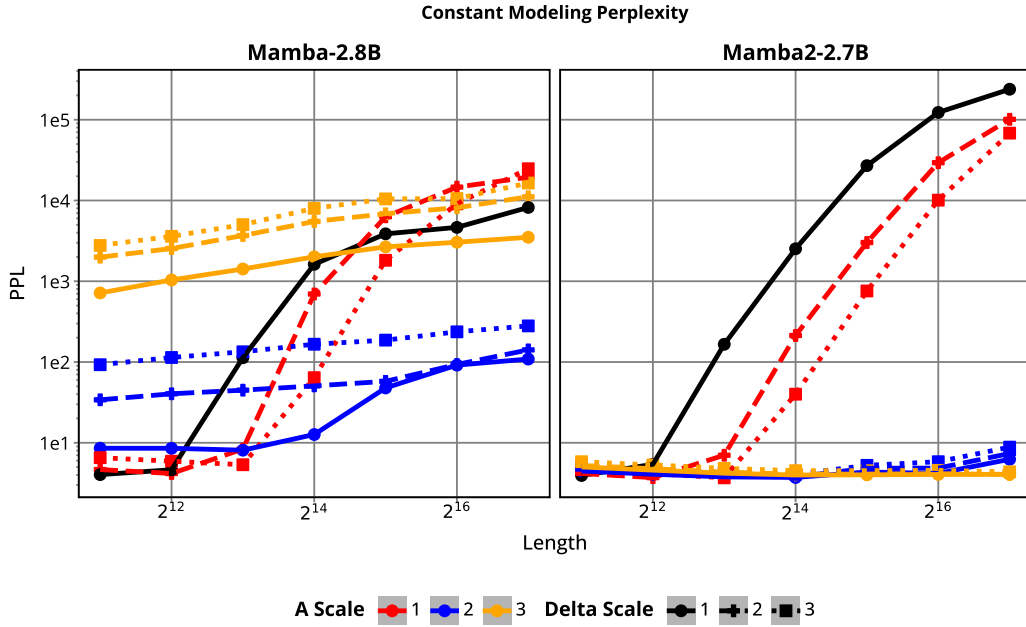


Figure 5: Language modeling perplexity on ProofPile after applying a constant scaling factor to either A or Δ_t . The solid black line indicates the baseline, where neither A nor Δ_t are scaled. Red lines indicate cases where Δ_t is scaled, while the other solid lines indicate where A was scaled.

710 B.3 MambaExtend Calibration

711 Here, we give an overview of the calibration functions we use within our MambaExtend-based
 712 experiments. Each of the described methods replace the calibration function **CF** within Algorithm 1.
 713 In our explicit implementation for calibrating scaling factors for \mathcal{A} , we use the same hyperparameters
 714 as Azizi et al. [2].

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

Algorithm 2 Calibration via back-propagation

```

1: Input: Frozen model  $\mathcal{M}$ , calibration set  $\mathcal{C}$ , initial scaling factors  $\mathbf{S}$ . Learning rate  $\eta$ , perturbation
   magnitude  $c$ , iterations  $K$ 
2: Output: Learned scaling factors  $\mathbf{S} = [s_1, \dots, s_L] \in \mathbb{R}_+^{d_s \times L}$ 
3: for  $k \leq K$  do
4:    $\delta \in \mathbb{R}^{d_s \times L} \sim \text{Radamacher}()$ 
5:    $\mathbf{S}^+ = \mathbf{S} + c \times \delta$ 
6:    $\mathbf{S}^- = \mathbf{S} - c \times \delta$ 
7:    $\ell^+ = \text{eval}(\mathcal{M}_{c \times \mathbf{S}^+}, \mathcal{C})$ 
8:    $\ell^- = \text{eval}(\mathcal{M}_{c \times \mathbf{S}^-}, \mathcal{C})$ 
9:    $\hat{\nabla}_{\mathbf{S}} = (\ell^+ - \ell^-) / (2 \cdot c \cdot \delta)$ 
10:   $\mathbf{S} \leftarrow \mathbf{S} - \eta \cdot \hat{\nabla}_{\mathbf{S}}$ 
11:   $\mathbf{S} \leftarrow \text{clamp}(\mathbf{S}, 0.001)$ 
12: end for
13: return  $\mathbf{S}$ 

```

718 **Calibration via zeroth-order optimization.** Zeroth-order optimization offers an efficient yet
 719 noisier method for calibration, as it relies solely on forward passes to approximate gradients. Specif-
 720 ically, this is a multi-iteration process in which, at each iteration, the scaling factors are randomly
 721 perturbed using a random variable δ sampled from a Rademacher distribution. The magnitude
 722 of the perturbation and the learning rate for the updates are controlled by the hyper-parameters c
 723 and η , respectively. We employ the two-sided variant of the simultaneous perturbation stochastic
 724 approximation method (SPSA) [48], which obtains gradient approximations by applying both positive
 725 and negative perturbations to the parameters simultaneously. The two-sided SPSA approach yields
 726 gradient estimates with lower variance than the one-sided version, thus enhancing accuracy, especially
 727 in noisy environments [50]. This is described in Algorithm 3.

Algorithm 3 Calibration via zeroth-order optimization

```

1: Input: Frozen model  $\mathcal{M}$ , calibration set  $\mathcal{C}$ , perturbation magnitude  $c$ , iterations  $K$ 
2: Output: Learned scaling factors  $\mathbf{S} = [s_1, \dots, s_L] \in \mathbb{R}_+^{d_s \times L}$ 
3: for  $k \leq K$  do
4:    $\delta \in \mathbb{R}^{d_s \times L} \sim \text{Radamacher}()$ 
5:    $\mathbf{S}^+ = \mathbf{S} + c \times \delta$ 
6:    $\mathbf{S}^- = \mathbf{S} - c \times \delta$ 
7:    $\ell^+ = \text{eval}(\mathcal{M}_{c \times \mathbf{S}^+}, \mathcal{C})$ 
8:    $\ell^- = \text{eval}(\mathcal{M}_{c \times \mathbf{S}^-}, \mathcal{C})$ 
9:    $\hat{\nabla}_{\mathbf{S}} = (\ell^+ - \ell^-) / (2 \cdot c \cdot \delta)$ 
10:   $\mathbf{S} \leftarrow \mathbf{S} - \eta \cdot \hat{\nabla}_{\mathbf{S}}$ 
11:   $\mathbf{S} \leftarrow \text{clamp}(\mathbf{S}, 0.001)$ 
12: end for
13: return  $\mathbf{S}$ 

```

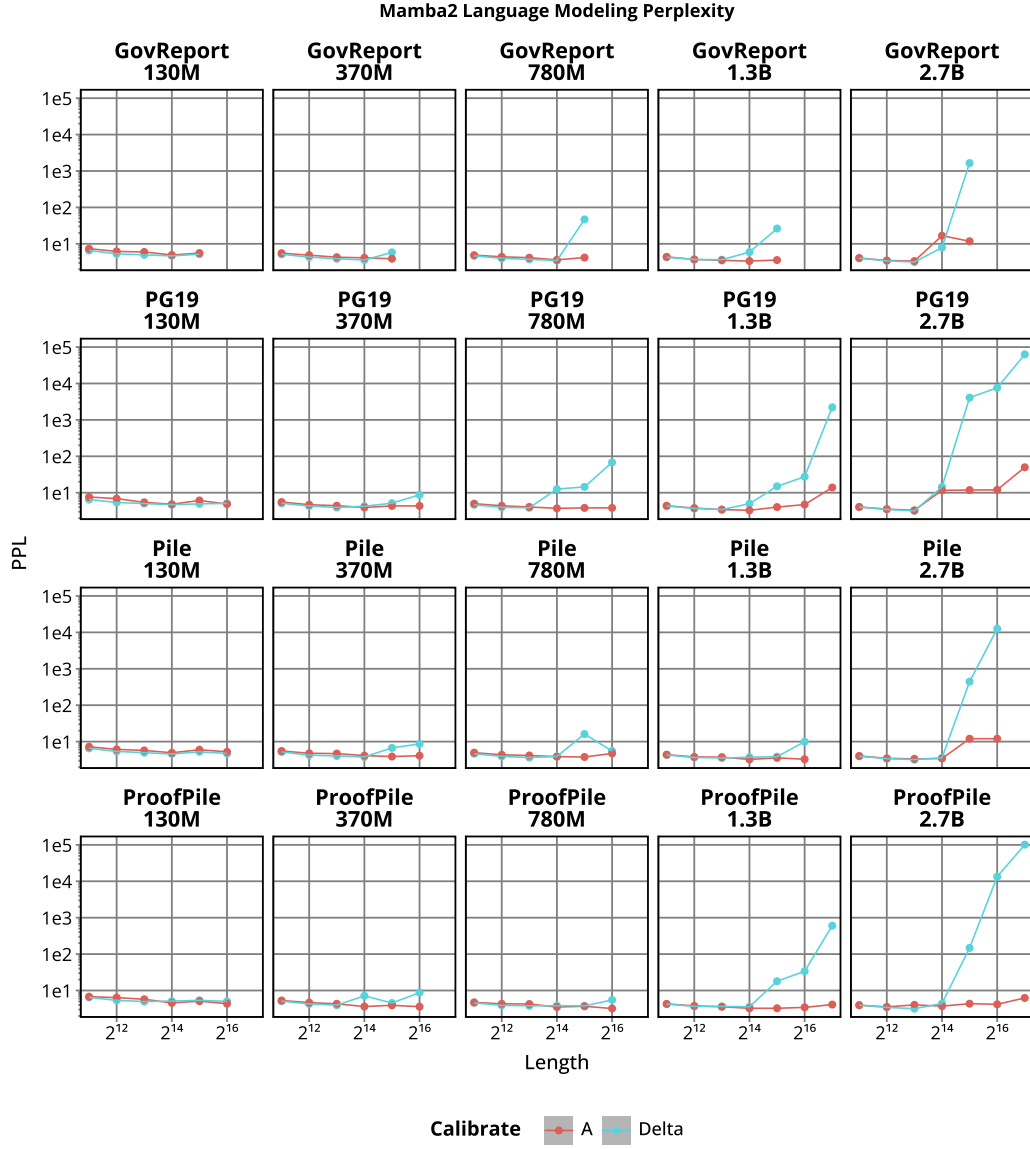


Figure 6: Language Model Perplexity performance of Mamba2 models by calibrating scaling factors for either $\log(A)$ (red lines) or Δ_t (cyan lines). Perplexities are reported across various datasets (GovReport, PG19, ProofPile, Pile) as well as model sizes.

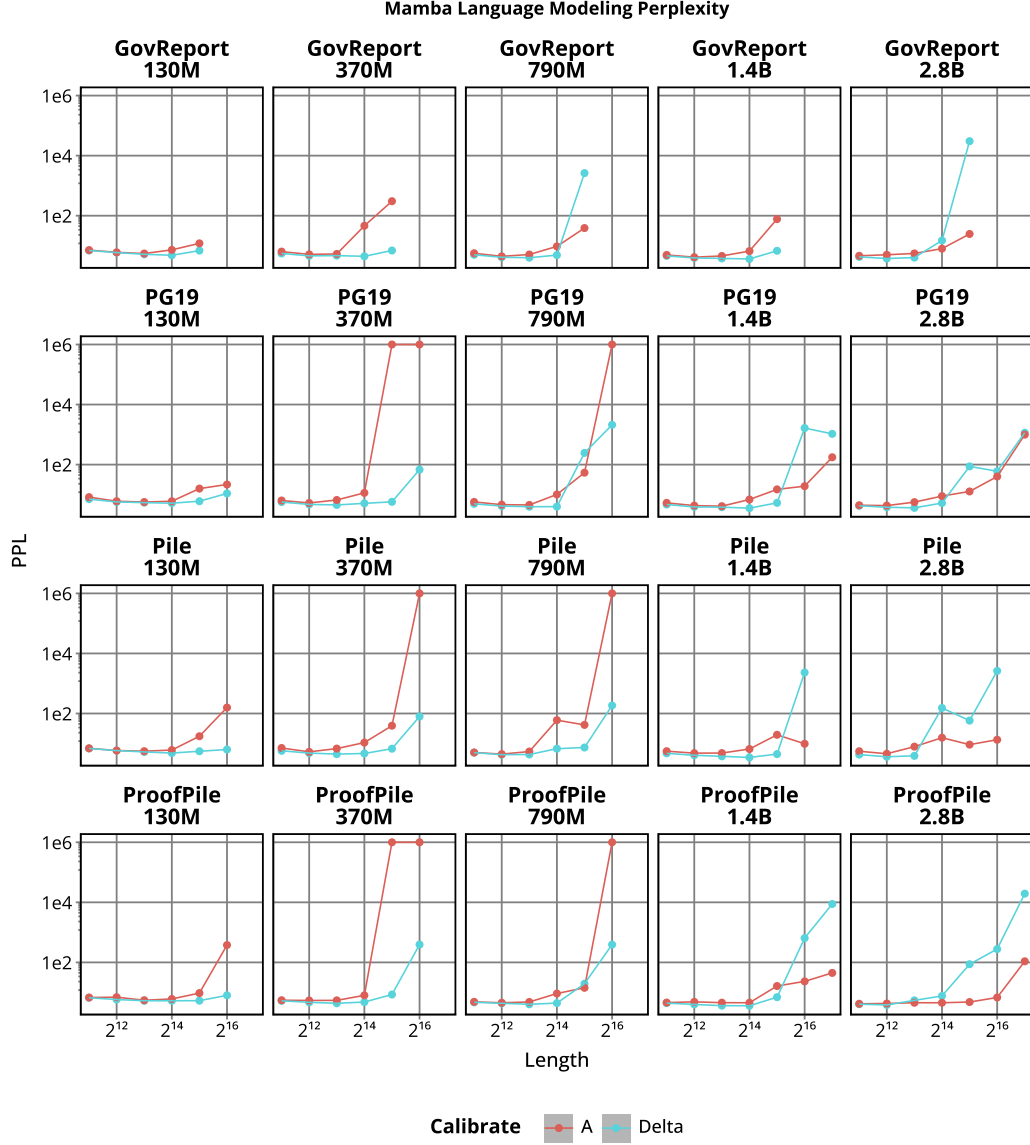


Figure 7: Language Model Perplexity performance of Mamba models by calibrating scaling factors for either $\log(A)$ (red lines) or Δ_t (cyan lines). Perplexities are reported across various datasets (GovReport, PG19, ProofPile, Pile) as well as model sizes.

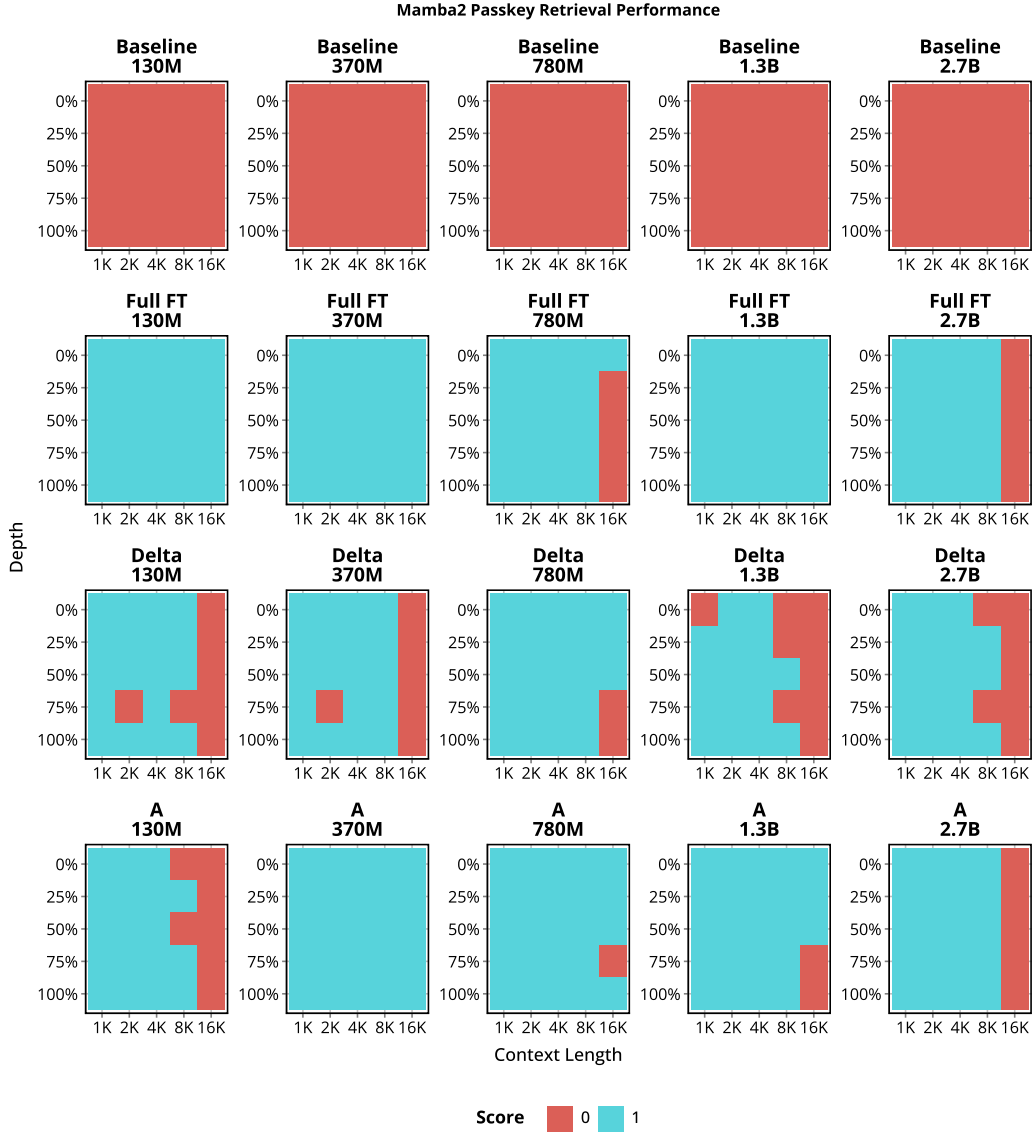


Figure 8: Passkey Retrieval performance of Mamba2 models by calibrating scaling factors for either $\log(\mathcal{A})$ or Δ_t . Blue squares mean that the model was able to solve all examples of the given evaluation length/depth pair after tuning scaling factors, while red squares means that at least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

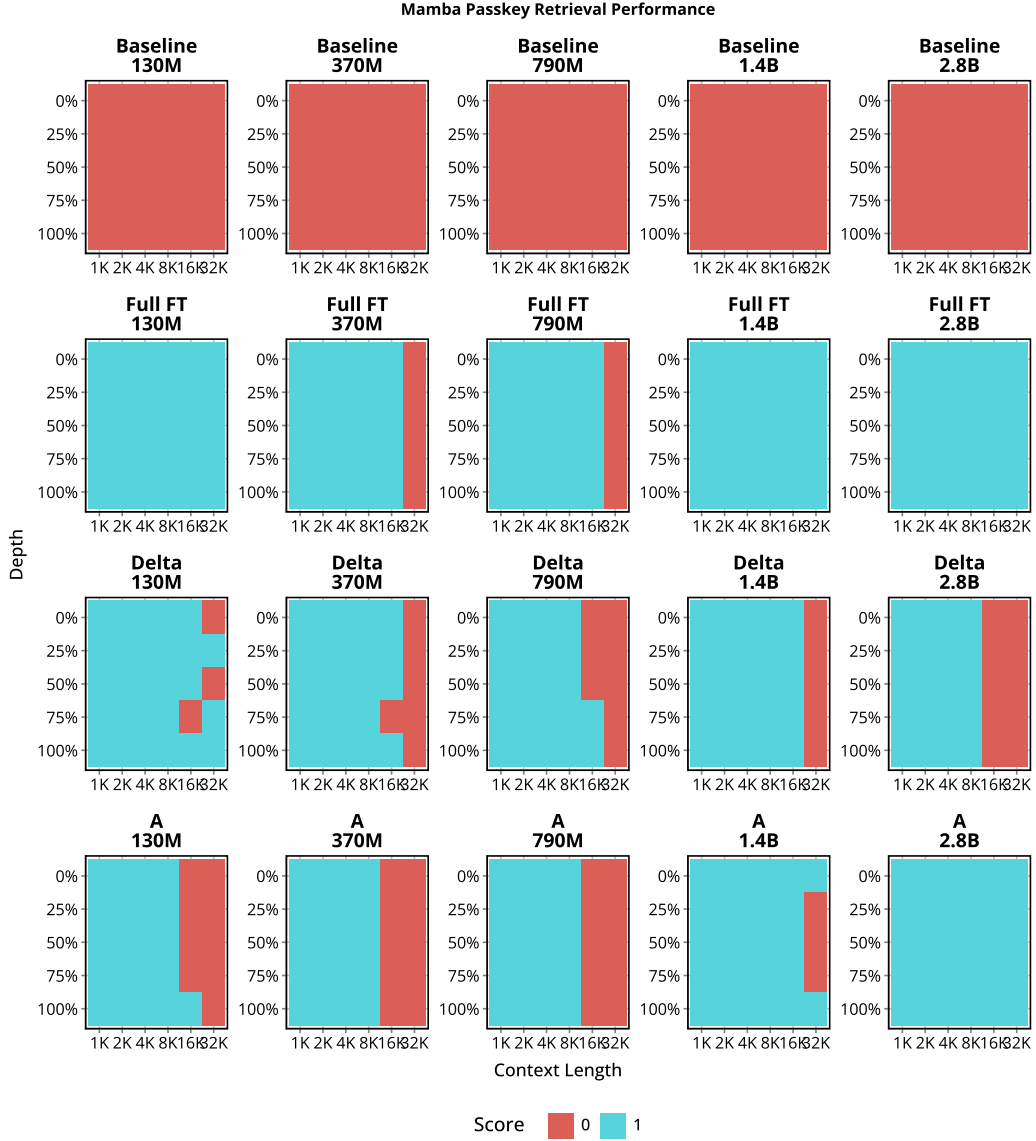


Figure 9: Passkey Retrieval performance of Mamba models by calibrating scaling factors for either $\log(A)$ or Δ_t . Blue squares mean that the model was able to solve all examples of the given evaluation length/depth pair after tuning scaling factors, while red squares means that at least one mistake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

730 B.3.3 LongBench

731 We evaluate the following tasks from LongBench (Table 3). Due to our pre-training on an English
 732 dataset, we choose to use only the English language tasks included in the benchmark.

Table 3: Tasks from LongBench on which we evaluate.

| Task | Context Type | Average Length | Metric | Data Samples |
|-------------------|---------------------|----------------|-----------------|--------------|
| QASPERQA [13] | Science | 3619 | F1 | 200 |
| HOTPOTQA [61] | Wikipedia | 9151 | F1 | 200 |
| 2WIKIMULTIQA [27] | Wikipedia | 4887 | F1 | 200 |
| TREC [35] | Web Questions | 5117 | Accuracy | 200 |
| TRIVIAQA [32] | Wikipedia/Web | 8209 | F1 | 200 |
| LCC [25] | Github | 1235 | Edit Similarity | 500 |
| REPOBENCH-P [37] | Github Repositories | 4206 | Edit Similarity | 500 |

733 B.4 PRE-TRAINED MODEL CHECKPOINTS USED

734 We use the official pre-trained model checkpoints of Mamba from the Hugging Face model Hub ³ :

- 735 • state-spaces/mamba-130m
- 736 • state-spaces/mamba-370m
- 737 • state-spaces/mamba-790m
- 738 • state-spaces/mamba-1.4b
- 739 • state-spaces/mamba-2.8b
- 740 • state-spaces/mamba2-130m
- 741 • state-spaces/mamba2-370m
- 742 • state-spaces/mamba2-780m
- 743 • state-spaces/mamba2-1.3b
- 744 • state-spaces/mamba2-2.7b

³<https://github.com/state-spaces/mamba>

745 **C Broader Impacts**

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

751 **D Limitations**

752 A potential limitation of our work is the application on Mamba-based models. As such, we qualify
753 our claims to only apply to such models. Further investigations can focus on similar linear recurrent
754 models or to hybrid models that combine both attention and recurrence.

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: [\[Yes\]](#)

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, hyper-parameters, 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.

- 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

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

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

975 11. Safeguards

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

979 Answer: [NA]

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

982 Guidelines:

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

993 12. Licenses for existing assets

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

997 Answer: [Yes]

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

1000 Guidelines:

- 1001 • The answer NA means that the paper does not use existing assets.
- 1002 • The authors should cite the original paper that produced the code package or dataset.
- 1003 • The authors should state which version of the asset is used and, if possible, include a
 1004 URL.
- 1005 • The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- 1006 • For scraped data from a particular source (e.g., website), the copyright and terms of
 1007 service of that source should be provided.
- 1008 • If assets are released, the license, copyright information, and terms of use in the
 1009 package should be provided. For popular datasets, `paperswithcode.com/datasets`
 1010 has curated licenses for some datasets. Their licensing guide can help determine the
 1011 license of a dataset.
- 1012 • For existing datasets that are re-packaged, both the original license and the license of
 1013 the derived asset (if it has changed) should be provided.
- 1014 • If this information is not available online, the authors are encouraged to reach out to
 1015 the asset's creators.

1016 13. New assets

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 rigor, or originality of the research, declaration is not required.

1069 Answer: [NA]
1070 Justification: We do not use LLMs in any non-standard or novel manner.
1071 Guidelines:
1072 • The answer NA means that the core method development in this research does not
1073 involve LLMs as any important, original, or non-standard components.
1074 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
1075 for what should or should not be described.