
The Mamba in the Llama: Distilling and Accelerating Hybrid Models

Junxiong Wang^{1*} Daniele Paliotta^{2,3*} Avner May³ Alexander M. Rush¹ Tri Dao^{3,4}
¹Cornell University ²University of Geneva ³Together AI ⁴Princeton University

Abstract

Linear RNN architectures, like Mamba, can be competitive with Transformer models in language modeling while having advantageous deployment characteristics. Given the focus on training large-scale Transformer models, we consider the challenge of converting these pretrained models for deployment. We demonstrate that it is feasible to distill large Transformers into linear RNNs by reusing the linear projection weights from attention layers with academic GPU resources. The resulting hybrid model, which incorporates a quarter of the attention layers, achieves performance comparable to the original Transformer in chat benchmarks and outperforms open-source hybrid Mamba models trained from scratch with trillions of tokens in both chat benchmarks and general benchmarks. Moreover, we introduce a hardware-aware speculative decoding algorithm that accelerates the inference speed of Mamba and hybrid models. Overall we show how, with limited computation resources, we can remove many of the original attention layers and generate from the resulting model more efficiently. Our top-performing model, distilled from Llama3-8B-Instruct, achieves a 29.61 length-controlled win rate on AlpacaEval 2 against GPT-4 and 7.35 on MT-Bench, surpassing the best 8B scale instruction-tuned linear RNN model. We also find that the distilled model has natural length extrapolation, showing almost perfect accuracy in the needle-in-a-haystack test at 20x the distillation length. Code and pre-trained checkpoints are open-sourced at <https://github.com/jxiw/MambaInLlama> and https://github.com/itsdaniele/speculative_mamba.

1 Introduction

While Transformers [73] have been an essential architecture in deep learning and have driven the success of large language models such as GPT [9], Llama [71], and Mistral [37], they are prohibitively slow for very long sequence generation due to their quadratic complexity with respect to sequence length and large key-value (KV) cache requirement. Recent linear RNN models (Mamba [26], Mamba2 [18], GLA [79], RWKV [55], RetNet [68], Griffin [19]) beat Transformers in controlled experiments at small to medium scale, although the best Transformers still significantly outperform these models on downstream tasks. We note that the training times of linear RNN models are similar to those of highly optimized Transformers [79], and therefore scaling up either of these models requires substantial computational resources.

*Equal Contribution. Order determined by coin flip. Correspondence to: junxiong@cs.cornell.edu and daniele.paliotta@unige.ch

The primary benefit of linear RNN models (Mamba [26], Mamba2 [18]) is that they have faster inference (5× higher throughput) than Transformers. Efficient inference is emerging as a critical need for LLM systems such as new applications currently bottlenecked by the large KV cache of Transformers, e.g. reasoning over multiple long documents [30, 65, 56] and files in large codebases [61, 42]). Emerging workflows with agents [81, 77] also require large-batch inference to explore more trajectories and long-context to model complex environments.

These properties motivate the goal of distilling a large pretrained Transformer model into a linear RNN in order to generate as efficiently as possible. The technical challenges are two-fold: how to map pretrained Transformer weights to linear RNN weights for distillation, and how to adapt best-practice Transformer inference techniques, such as speculative decoding, to the new architecture. We make the following contributions:

- We show that by reusing weights from attention layers, it is possible to distill a large transformer into a large hybrid-linear RNN with minimal additional compute while preserving much of its generation quality. We propose a modified Mamba architecture that can be directly initialized from the attention block of a pretrained model.
- We propose a multistage distillation approach that mirrors the standard LLM pipeline combining progressive distillation, supervised fine-tuning [39], and directed preference optimization [58]. This approach shows better perplexity and downstream evaluation compared with vanilla distillation.
- We develop a hardware-aware speculative sampling algorithm and a fast kernel for speculative decoding on Mamba and hybrid architectures. We achieve a throughput of over 300 tokens/second for a Mamba 7B model. Additionally, we show that speculative decoding can be effectively applied to our hybrid architecture.

Our experiments distill different large-scale open chat LLMs, Zephyr-7B [72], Llama-3 8B [21] to linear RNN models (hybrid Mamba and Mamba2), using only 20B tokens of training. Results show that the distilled approach matches the teacher model in standard Chat benchmarks [84, 43]. We also show that it performs on par or better with all similarly sized pretrained-from-scratch Mamba models including Mamba 7B models [52, 26] trained from scratch with 1.2T tokens or NVIDIA Hybrid Mamba2 models [74] trained from scratch with 3.5T tokens in multiple tasks (e.g., MMLU [34], TruthfulQA [47]) in the LM evaluation [25]. Concurrent with this work, MOHAWK [6] distills a Mamba-2 variant based on the Phi-1.5 architecture with limited computation costs and performance loss.

2 From Transformer to Mamba

2.1 Relationship Between Attention and Linear RNNs

We begin by reviewing multihead attention to clarify the shapes of intermediate objects. Notationally, we use explicit subscripts for the sequence position instead of matrix representation, to better highlight similarities between the two models.

Attention is computed in parallel for multiple differently parameterized heads. Each head takes sequence o with hidden size D as an argument and computes,

$$\begin{aligned} \mathbf{Q}_t &= \mathbf{W}^Q o_t, & \mathbf{K}_t &= \mathbf{W}^K o_t, & \mathbf{V}_t &= \mathbf{W}^V o_t & \text{for all } t, \\ \alpha_1 \dots \alpha_T &= \text{softmax}([\mathbf{m}_{1,t} \mathbf{Q}_t^\top \mathbf{K}_1 \dots \mathbf{m}_{T,t} \mathbf{Q}_t^\top \mathbf{K}_T] / \sqrt{D}) & \mathbf{y}_t &= \sum_{s=1}^t \alpha_s \mathbf{V}_s \\ \text{where } o_t &\in \mathbb{R}^{D \times 1}, & \mathbf{W} &\in \mathbb{R}^{N \times D} & \mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t &\in \mathbb{R}^{N \times 1} & m_{s,t} = \mathbf{1}(s \leq t) \end{aligned}$$

Recent work has argued that linear RNNs can be serious competitors to attention in large language models. Several different linear RNN formulations have been proposed with similar formulations. For now, we leave the shapes of the parameters $\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t$ abstract, and note that linear RNNs all take the following form that maps a 1-dimensional sequence to another through an implicit matrix-valued hidden state \mathbf{h} .

$$\mathbf{h}_t = \mathbf{A}_t \mathbf{h}_{t-1} + \mathbf{B}_t x_t, \quad y_t = \mathbf{C}_t^\top \mathbf{h}_t \tag{1}$$

Linear RNNs have several computational advantages over attention. During training, all \mathbf{y}_t values can be computed more efficiently than attention since there is no softmax non-linearity. During inference, each next \mathbf{y}_t can be computed serially without requiring a cache.

Despite the superficially different form, there is a natural relationship between linear RNNs and attention. *Linearizing* the attention formula by removing the softmax yields:

$$y_t = \sum_{s=1}^t \alpha_s v_s = \frac{1}{\sqrt{D}} \sum_{s=1}^t m_{s,t} \mathbf{Q}_t^\top \mathbf{K}_s v_s = \frac{1}{\sqrt{D}} \mathbf{Q}_t^\top \sum_{s=1}^t m_{s,t} \mathbf{K}_s v_s$$

This implies that there exists a linear RNN form of linearized attention, specifically:

$$\begin{aligned} \mathbf{h}_t &= m_{t-1,t} \mathbf{h}_{t-1} + \mathbf{K}_t v_t & y_t &= \frac{1}{\sqrt{D}} \mathbf{Q}_t^\top \mathbf{h}_t \\ &\downarrow \\ \mathbf{h}_t &= \mathbf{A}_t \mathbf{h}_{t-1} + \mathbf{B}_t x_t, & y_t &= \mathbf{C}_t^\top \mathbf{h}_t \\ \mathbf{A}_t &= m_{t-1,t}, \mathbf{B}_t = \mathbf{W}^K \mathbf{o}_t, \mathbf{C}_t = \mathbf{W}^Q \mathbf{o}_t, \mathbf{x}_t = \mathbf{W}^V \mathbf{o}_t \end{aligned}$$

Note though that this version uses a hidden state of size $\mathbf{h} \in \mathbb{R}^{N \times 1}$. Effectively tracking only one scalar over time per hidden dimension. Naively applying this transformation leads to poor results. The issue is that linearizing attention produces a degraded representation of the original model, as the softmax nonlinearity is critical to attention.

The key to improving these models is to increase the capacity of the linear hidden state to better capture long-term structure. For example, previous work has shown the use of kernel methods to improve this approximation [63, 36, 83]. These approaches expand the size of the hidden state representation to \mathbf{h} to $\mathbb{R}^{N \times N'}$ to better match the modeling capacity of softmax.

2.2 Distilling to an Expanded Linear RNN

To design a effective distilled linear RNN, we aim to stay as close as possible to the original Transformer parameterization, while also expanding the capacity of the linear RNN in an efficient manner. We will *not* attempt to have the new model capture the exact original attention function, but instead use the linearized form as a starting point for distillation.

Specifically, we adapt the parameterization from Mamba, [26] to increase the hidden state size, while initializing from the attention representation. Mamba uses a continuous time state-space model (SSM) to parameterize a linear RNN at run time, described by the differential equation,

$$\mathbf{h}'(k) = \mathbf{A} \mathbf{h}(k) + \mathbf{B}(k) \mathbf{x}(k) \quad \mathbf{y}(k) = \mathbf{C}(k) \mathbf{h}(k)$$

Where \mathbf{A} is a diagonal matrix and other values are continuous signals. To apply this formulation to a discrete-time problem like language modeling, we use a neural network to produce a sequence of sampling intervals Δ_t and samples of the signals at these time steps. Given these sampling intervals, and T samples of \mathbf{B}, \mathbf{C} , Mamba approximates the continuous-time equation using a linear RNN as a discretization. We use an overbar to indicate the discrete-time form, which is reconstructed dynamically.

$$\overline{\mathbf{A}}_{1:T}, \overline{\mathbf{B}}_{1:T}, \overline{\mathbf{C}}_{1:T} = \text{Discretize}(\mathbf{A}, \mathbf{B}_{1:T}, \mathbf{C}_{1:T}, \Delta_{1:T})$$

Algorithm 1 Attention-Initialized Mamba

- 1: **Shapes:** B - Batch, L - Length, D - embed size,
 - 2: $N = D/\text{Heads}$, N' - expand
 - 3: **Input:** \mathbf{o}_t : (B, D)
 - 4: **Output:** output: (B, D)
 - 5: **New Params:** MLP, \mathbf{A}
 - 6: **for** each head $\mathbf{W}^k, \mathbf{W}^q, \mathbf{W}^v, \mathbf{W}^o : (N, D)$
 - 7: expanding grouped KVs **do**
 - 8: **Head Parameter:** $\mathbf{A} : (N, N')$
 - 9: **for** all positions t :
 - 10: $\mathbf{x}_t : (B, N) \leftarrow \mathbf{W}^V \mathbf{o}_t$
 - 11: $\mathbf{B}_t : (B, N) \leftarrow \mathbf{W}^K \mathbf{o}_t$
 - 12: $\mathbf{C}_t : (B, N) \leftarrow \mathbf{W}^Q \mathbf{o}_t$
 - 13: $\Delta_t : (B, N') \leftarrow \text{MLP}(\mathbf{x}_t)$
 - 14: $\overline{\mathbf{A}}_{1:T}, \overline{\mathbf{B}}_{1:T}, \overline{\mathbf{C}}_{1:T} : (B, N, N') \leftarrow$
 - $\text{DISC}(\mathbf{A}, \mathbf{B}, \mathbf{C}, \Delta)$
 - 15: $\mathbf{y} \leftarrow \text{LINEARRNN}(\overline{\mathbf{A}}, \overline{\mathbf{B}}, \overline{\mathbf{C}}, \mathbf{x})$
 - 16: output \leftarrow output + $\mathbf{W}^{O^\top} \mathbf{y}$
 - 17: **return** output
-

In this simplest case, with $N' = 1$ and an identity discretization, this approach recovers the linear attention to linear RNN conversion discussed in the previous section. The benefit of Mamba is that with $N' > 1$ the continuous-time parameterization allows the model to learn significantly richer functions, without many more parameters or decreased efficiency. Specifically the only additional learned parameters will be the sampling rate Δ and the dynamic \mathbf{A} . These new parameters will control the constructed linear RNN through the discretization function yielding the new matrix valued linear RNN. Specifically, we take in the same $\mathbf{B}_t, \mathbf{C}_t \in \mathbb{R}^{N \times 1}$ and $\Delta_t \in \mathbb{R}^{N'}$, but output $\bar{\mathbf{B}}_t, \bar{\mathbf{C}}_t \in \mathbb{R}^{N' \times N \times 1}$, effectively increasing the hidden size by a factor of N' over the naive linear attention.

A core contribution of Mamba [26, 18] is to demonstrate a hardware-aware factorization of this algorithm. Implementing the algorithm naively would be prohibitively slow as the new expanded parameters are quite large. Their approach fuses discretization, state expansion, and applying the linear RNN into a single kernel, which circumvents fully materializing the discrete parameters. This allows for large N' with relatively small efficiency costs.

2.3 Attention-to-Mamba Initialization and Hybrid Stepwise Training

Our full approach is shown in Algorithm 1. This algorithm feeds the standard $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ heads from attention directly into the Mamba discretization, and then applies the resulting linear RNN. As noted above, this can be seen as roughly initializing with linearized attention and allowing the model to learn richer interactions through the expanded hidden state.

Figure 1 shows the resulting architecture. Our version directly replaces Transformer attention heads directly with fine-tune linear RNN layers. We keep the Transformer MLP layers as is and do not train them. This approach also requires processing additional components like grouped query attention that shares keys and values across heads. We note that this architecture differs from the architecture used in many Mamba systems, which combines MLP-SSM layers and uses a single head.

This initialization allows us to replace any attention block with a linear RNN block. We experiment with *hybrid* models where we keep every n attention layers. Empirically we found that replacing layers in a stepwise manner was the most effective strategy, i.e. we first keep every 2 layers, distill, and then every 4, and continue distillation.

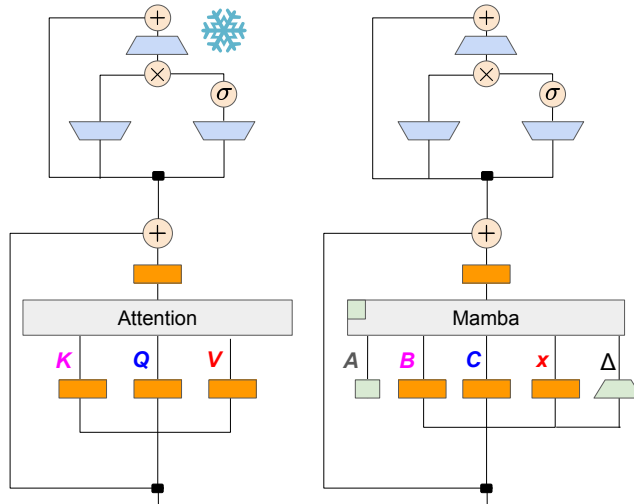


Figure 1: Transferring Transformer to Mamba. Weights, in orange, are initialized from the Transformer (Linear projections for \mathbf{Q}, \mathbf{K} , and \mathbf{V} are initialized using linear projection for \mathbf{C}, \mathbf{B} , and \mathbf{X} respectively). We replace individual attention heads with Mamba heads, and then finetune Mamba blocks while freezing the MLP blocks. Shapes are kept mainly the same. Weights in green are added. New parameters are introduced for the learned \mathbf{A} and Δ parameters.

3 Knowledge Distillation for Aligned LMs

Knowledge distillation (KD) [35] serves as a compression technique aimed at training a smaller network that mimics the behavior of a larger teacher network. After initializing the model from the Transformer parameters, we aim to distill it to perform on par with the original language model. We assume that most of the knowledge from the transformer is maintained in the MLP layers which were transferred from the original model, and focus on distilling the fine-tuning and alignment steps of the LLM. During this stage, the MLP layers are kept frozen and the Mamba layers are trained as in Figure 1.

Supervised Fine-Tuning We first apply knowledge distillation to redo the supervised fine-tuning (SFT) stage of language model adaptation. During this stage, an LLM is trained to maximize the likelihood of a response y given an input prompt x , i.e. $p(y | x)$. The task looks similar to conditional generation.

There are two common approaches for distillation in this setting. One method is to use word-level KL-Divergence. In this setting, the full probability distribution of the student model $p(\cdot; \theta)$ is trained to match the full distribution of the teacher model $p(\cdot; \theta_T)$ by minimizing the KL divergence over the entire set of next possible tokens at position t . The second method is sequence-level knowledge distillation (SeqKD) [39]. SeqKD suggests a simple method for distillation on this style of task, by replacing the ground truth text $y_{1..t}$ with the teacher generation output $\hat{y}_{1..t}$, also known as pseudo-labels.

$$\mathcal{L}(\theta) = - \sum_{t=1}^T \alpha \log p(\hat{y}_{t+1} | \hat{y}_{1:t}, x, \theta) + \beta \text{KL} [p(\cdot | \hat{y}_{1:t}, x, \theta) || p(\cdot | \hat{y}_{1:t}, x, \theta_T)] \quad (2)$$

Here θ is trainable parameters of the student model and α and β control the weights of sequence and word loss term respectively.

Preference Optimization The second stage of instruction-tuning for LLMs is to align them to a set of user preferences. During this stage, a set of desired preference pairs is used to improve the model’s output. The objective is to produce outputs y to prompts x that maximize a reward model r while maintaining close to a reference model. Typically the reference model is chosen to be the model after supervised fine-tuning. For distillation, we can conveniently utilize the original teacher, i.e.

$$\max_{\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim p(y|x;\theta)} [r_{\phi}(x, y)] - \beta \text{KL} [p(y | x; \theta) || \pi(y | x; \theta_T)] \quad (3)$$

This preference model is defined by a reward function $r_{\phi}(x, y)$ dependent on the method used. Previous research utilizing AI feedback has primarily focused on employing reinforcement learning methods, such as proximal policy optimization (PPO) [64], to optimize ϕ concerning this reward. Recently, methods using direct preference optimization (DPO) [58] have been effective at optimizing this objective with direct gradient updates. Specifically, DPO shows that, if we have access to preferred y_w and dispreferred y_l outputs for a given prompt x , we can reformulate this optimization problem as,

$$\pi_{\theta} = \max_{\theta} \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \log \sigma \left(\beta \log \frac{p(y_w|x; \theta)}{p(y_w|x; \theta_T)} - \beta \log \frac{p(y_l|x; \theta)}{p(y_l|x; \theta_T)} \right). \quad (4)$$

This optimization can be performed at the sequence level by scoring the preferred and dispreferred outputs of the model with the teacher and student and then backpropagating to the student. As far as we are aware this is the first use of DPO as a distillation objective.

4 Speculative Decoding Algorithms For Linear RNNs

The main goal of the linear RNN formulation is to improve decoding efficiency. For both attention and linear RNNs, the serial dependency of autoregressive generation inherently bottlenecks efficiency. Systems cannot utilize all available compute, as they need to wait for the generation of previous tokens to proceed [67, 41, 11, 76, 10]. *Speculative decoding* has emerged as a method for breaking this bottleneck by spending extra compute to speculate

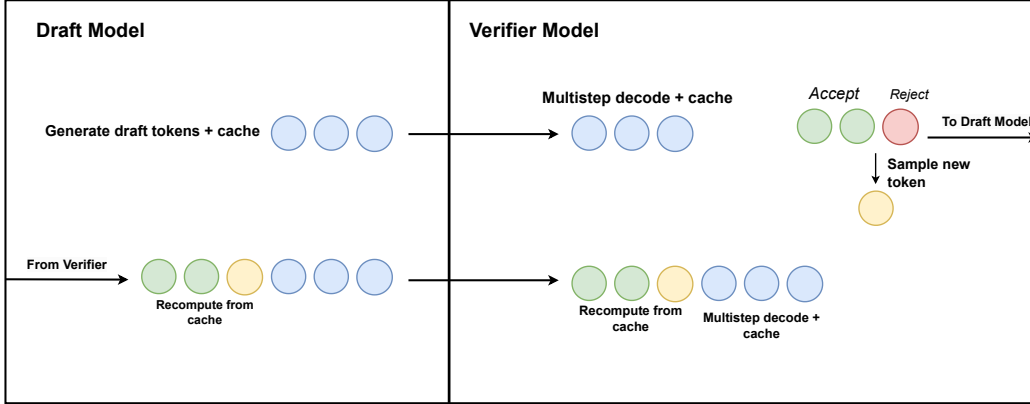


Figure 2: Multi-Step RNN Speculative Decoding. *Left (top)*: The draft model generates the set of blue draft tokens sequentially. The draft tokens are then verified. *Right (top)*: Verification uses the multistep kernel, without materializing the intermediate states. The last token is rejected and replaced with the true best token. Note, that even though more tokens are generated we cannot advance the hidden state cache. *Left (bottom)* The draft model can now generate more blue draft tokens from the current tokens, resulting in six total. *Right (bottom)* When the new draft is verified, the multi-step kernel returns both the hidden state after the yellow token and the final hidden state, since verification will fall between those positions. on future generations. In this section, we consider approaches for applying this technique to large Mamba models, which can then be applied to the distilled models.

4.1 Challenges in RNN Speculation

Speculative decoding uses two models: a draft model, θ_D , and a verification model, θ_V . The fast draft model produces potential future completions, $y^* = \arg \max_{y_{1:T}} p(y_1, \dots, y_T; \theta_D)$, and the larger verification model checks that these are top ranking at each time step, i.e. checking $p(y_t^* | y_{1:t-1}^*; \theta_V)$. The longer a chain before a verification failure the faster the output. If a partial chain matches, we can rewind to the last match.

Attention-based models are particularly amenable to speculation, as they are slow at generation due to sequential nature, but fast at verification due to their ability to check multiple tokens in parallel. Linear RNN models like Mamba have significantly different performance characteristics that make them less amenable to speculative decoding. Sequential decoding using recurrent-style sampling is already significantly faster than attention. Like attention, there are parallel modes for models like Mamba which are used at training. These are efficient, but are tuned for extremely long sequences. In addition, they rely on hardware-aware optimizations, such as avoiding materializing intermediate states. These properties make it difficult to use for speculation for relatively short chains when it is unknown when a conflict will occur.

An additional challenge arises from caching states in RNN models. The state of an attention model is represented by the key-value cache, $\mathbf{K}_{1:t}, \mathbf{V}_{1:t}$; whereas the state of an RNN model is simply \mathbf{h}_t . To be competitive with attention this single RNN state needs to be very large. During speculation, we need to rewind to a previous state at time step t' . For attention, this is simply $\mathbf{K}_{1:t'}, \mathbf{V}_{1:t'}$; however, for RNNs this would require caching all $\mathbf{h}_{1:t}$ which would require a large memory overhead.

4.2 Multi-Step Linear RNN Speculation

We propose a new algorithm for linear RNN speculative decoding using hardware-aware multi-step generation. The core to the approach is the generation kernel that computes,

$$y_{j:k}, \mathbf{h}_j, \mathbf{h}_k \leftarrow \text{MULTISTEP}(\mathbf{h}_i, y_{1:n}, i, j, k; \mathbf{A}, \mathbf{B}, \mathbf{C}, \Delta)$$

Where i is the starting hidden state, $i \leq j \leq k$, and $j \dots k$ is the range of \mathbf{y} outputs needed. The kernel is hardware-aware because it avoids materializing key terms off of the fast GPU

memory. Specifically, it avoids instantiating most $\mathbf{h}_{1:n}$ as well as the discrete-time linear RNN parameters. This kernel is aimed to target the issues presented above. Specifically, it can save a snapshot of the state \mathbf{h}_j before evaluating the draft tokens. This allows recomputing the correct state on the fly after a token is rejected. The assumption is that decoding is bottlenecked by memory and not by compute, as we can compute multiple steps of decoding with very little overhead over single-step decoding.

Algorithm 2 and Figure 2 show the full algorithm. The approach maintains only one RNN hidden state in cache for verification and advances it lazily based on the success of the multi-step kernel. Since the distilled models contain transformer layers, we also extend speculative decoding to Attention/RNN hybrid architectures. In this setting, the RNN layers perform verification according to Algorithm 2, while the transformer layers simply perform parallel verification.

Note that if the draft model is a Mamba or hybrid model, the speculation part of the algorithm gets more complicated, as the draft model needs to recompute the state for the tokens accepted in the previous iteration. This is done similarly to the verifier model, by caching older entries and recomputing on the fly during the next round of speculation.

Algorithm 2 Multi-Step Linear RNN Speculation

```

function VERIFY( $y_{1:k}, j, \mathbf{h}_i$ )
  //  $y_{1:k}$  are draft,  $j$  is last verified,
  //  $\mathbf{h}_i$  is a cached state with  $i \leq j$ 
   $y'_{j:k}, \mathbf{h}_j, \mathbf{h}_k$  ← MULTISTEP( $\mathbf{h}_i, y_{1:k}, i, j, k; \theta_v$ )
   $k' \leftarrow$  FIRSTCONFLICT( $y_{j:k}, y'_{j:k}$ )
  return  $k', \mathbf{h}_k$  if  $k' = k$  else  $\mathbf{h}_j$ 

function SPECULATE( $K$ )
  //  $K$  tokens are drafted per step
   $\mathbf{h}_{\text{cache}} \leftarrow \mathbf{h}_0$ 
   $j \leftarrow 0$ 
  while  $y_j$  is not end do
     $k \leftarrow j + K$ 
     $y_{j+1:k} \leftarrow$  arg max  $p(y_{j+1:k} | y_{1:j}, \theta_D)$ 
     $j, \mathbf{h}_{\text{cache}} \leftarrow$  VERIFY( $y_{1:k}, j, \mathbf{h}_{\text{cache}}$ )
  return  $y_{1:j}$ 

```

4.3 Speculation Analysis and Hardware Specific Optimization

To verify the effectiveness of this approach we run the speculation using Mamba 7B and Mamba 2.8B as target models. Results are shown in Table 1. Figure 3 shows the performance characteristics of the Multi-Step kernel itself.

Model Size	GPU	K	# Gen. Tokens	Throughput (toks/s)	Speedup
2.8B	3090	3	3.01	259	2.3x
2.8B	3090	4	3.28	289	2.6x
2.8B	H100	3	4.04	389	1.71x
2.8B	H100	4	3.9	421	1.85x
7B	3090	3	3.19	109	2.1x
7B	3090	4	3.56	110	2.1x
7B	H100	3	3.28	271	1.95x
7B	H100	4	3.6	272	2x

Table 1: Speedup results for speculative decoding with pure Mamba models. The 2.8B verifier uses a 130M Mamba draft. The 7B verifier uses a Llama3 1B draft we trained. Data is from The Pile. K is number of draft tokens produced, # Gen includes an additional token from the last verifier logits.

Speedup on H100 GPUs. A naive implementation of our algorithm already shows strong performance on Ampere GPUs as shown in Table 1. However, achieving strong performance on H100 GPUs is much more challenging. This is mainly due to GEMM operations being much faster, which makes the overhead incurred from the caching and recomputation operations more visible. In practice, the naive implementation of our algorithm, with several different kernel

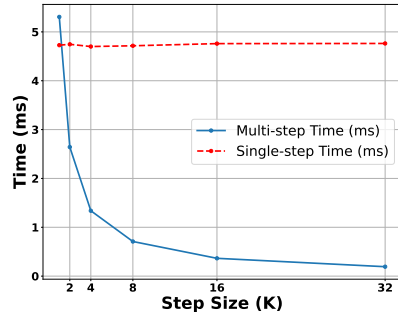


Figure 3: Performance of the multi-step SSM kernel for generating 32 tokens.

calls, achieves a decent speedup on 3090 GPUs (1.5x for Mamba 2.8B with 60% acceptance rate) but no speedup at all on H100s.

We optimized our implementation by fusing kernels, and by adapting the implementation to easily allow caching and recomputing old steps. Specifically, the verifier model performs i) recomputation of previous steps from the cache, ii) multistep decoding for the new sequence of draft tokens and iii) caching within a single kernel ². For the draft model, recomputation, decoding and caching are also fused in a single kernel. The resulting implementations archives speedups on H100s GPUs, as shown in Table 1.

5 Results

5.1 Experimental Setup

Target models. We perform experiments using two LLM chat models: Zephyr-7B [72], which is a chat fine-tuned Mistral 7B [37], Llama-3 Instruct 8B [21]. For the linear RNN models, we use hybrid versions of Mamba and Mamba2 with 50%, 25%, 12.5%, and 0% attention layers. We refer to 0% as a pure Mamba model. Mamba2 is a variant architecture of Mamba that is designed to be more targeted to recent GPU architectures. Zephyr-Mamba refers to a distillation from Zephyr [72], while Llama3-Mamba / Llama3-Mamba2 indicates distillation from Llama-3 instruct 8B [71]. Strictly speaking, our distilled Mamba-Zephyr is a subquartic model, since Zephyr/Mistral-8B uses sliding window attention architecture. Our distilled Mamba-Zephyr (50%) has the similar architecture as Samba [60].

Training. Distillation does not require any language modeling pretraining data, but instead uses the post-training process to adapt the new model. We use a three-stage process. In the first stage, we use UltraChat [20] and UltraFeedback [17] as seed prompts and use the teacher model to generate pseudo-labels. The student model is trained in one epoch using the loss \mathcal{L} in Eq 2 with $\alpha = 1$ and $\beta = 0.1$. Models are trained using AdamW optimizer with $\beta = (0.9, 0.98)$ with a batch size 64. We use a linear learning rate warm-up (for the first 500 steps) followed by cosine annealing. In the second stage, we use supervised finetuning with our model on the GenQA [12], InfinityInstruct [3] and OpenHermes 2.5 [70] datasets using SFT in one epoch, with the same hyperparameters as Zephyr [72]. In the final stage, for models distilled from Zephyr, we do distilled alignment with our model using DPO on the UltraFeedback [17] dataset which is consistent with teacher model. While models distilled from Llama-3 instructed 8B, we use datasets from SimPO [51] and Zephyr [72]. We only freeze Gated MLP (FFN) in the first stage, while in the second and final stage all parameters are trained ³. The total distillation process for each hybrid model (e.g., Mamba-Llama3 (50% att)) takes less than five days in 8x80G A100.

Baselines. In addition to the core Transformer architectures, the main baselines we compare against are other large-scale linear RNN models. We compare with both pure SSM architectures, such as TRI Mamba 7B [52] trained with 1.2T tokens and Falcon Mamba 7B⁴ trained with more than 5T tokens, hybrid SSM architectures, such as Nvidia Hybrid Mamba 2 [74] trained with 3.7T tokens, and other linear hybrid RNN models, such as Recurrent Gemma-9B Instruct [8, 19].

After the release of the new SoTA transformer models at the 8B and 3B scales, Llama-3.1 and Llama-3.2, we have streamlined the distillation process and are now distilling using the larger Llama-3.1 70B teacher model while initializing models with similarly sized 3B and 8B scales, respectively. We distill our model on the GenQA [12] and InfinityInstruct [3] datasets, resulting in Mamba-Llama3.2-3B, Mamba2-Llama3.2-3B, Mamba-Llama3.1-8B, and Mamba2-Llama3.1-8B. Additionally, we perform further DPO on top of these models using the same dataset as before, resulting in Mamba-Llama3.2-3B-dpo, Mamba2-Llama3.2-3B-dpo, Mamba-Llama3.1-8B-dpo, and Mamba2-Llama3.1-8B-dpo. The distillation phase takes eight days on 8xA100 and four days on 8xH100.

²Additionally, we implement the convolutional part of the Mamba block using a circular buffer which allows us to keep track of the old entries and include them in the convolution when they are needed for recomputation.

³We freeze the MLP layers in the first stage because we want to produce a model similar to the initialization model. However, in the end-to-end distillation, we only focus on the KL loss, so training all parameters (not freezing the MLP layers) will give better results.

⁴<https://huggingface.co/tiiuae/falcon-mamba-7b>

5.2 Evaluation on Chat Benchmarks

We evaluate our models using both single-turn, AlpacaEval [43] and multi-turn chat benchmarks, MT-Bench [84]. These benchmarks assess the model’s ability to follow instructions and respond to challenging prompts across a wide variety of domains.

Model (% Att)	Size	Align	MT-Bench (score)	MT-Bench (Round 1)	MT-Bench (Round 2)	AlpacaEval (LC win %)	AlpacaEval (win %)
Zephyr	7B	DPO	7.34	-	-	13.20 _{0.96}	10.99 _{0.96}
Mamba-Zephyr (50%)	7B	DPO	7.31	-	-	20.66 _{0.74}	16.69 _{1.10}
Mamba-Zephyr (25%)	7B	DPO	7.03	-	-	17.16 _{0.69}	13.11 _{1.00}
Mamba-Zephyr (12.5%)	7B	DPO	6.40	-	-	15.32 _{0.66}	12.96 _{1.02}
Llama-3.1-Instruct	8B	RLHF	8.0	-	-	20.9	21.8
Mamba-Llama3.1 (50%)	8B		7.7	8.0	7.3	18.97 _{1.23}	21.22 _{1.23}
Mamba2-Llama3.1 (50%)	8B		7.6	8.1	7.0	18.99 _{1.24}	21.55 _{1.24}
Mamba-Llama3.2 (50%)	3B		6.9	7.6	6.1	13.57 _{1.08}	15.54 _{1.08}
Mamba2-Llama3.2 (50%)	3B		6.5	7.1	5.8	12.61 _{1.05}	14.34 _{1.05}
Llama-3-Instruct	8B	RLHF	8.00	-	-	22.90 _{1.26}	22.60 _{1.26}
Mamba-Llama3 (50%)	8B	DPO	7.35	7.82	6.88	29.61 _{1.31}	26.69 _{1.31}
Mamba-Llama3 (25%)	8B	DPO	6.86	7.56	6.15	25.85 _{1.26}	22.50 _{1.26}
Mamba-Llama3 (12.5%)	8B	DPO	6.46	6.91	6.01	20.76 _{1.16}	17.93 _{1.16}
Mamba2-Llama3 (50%)	8B	DPO	7.32	7.93	6.70	26.78 _{1.26}	22.69 _{1.26}
Mamba2-Llama3 (25%)	8B	DPO	6.74	7.24	6.24	22.75 _{1.18}	19.01 _{1.18}
Mamba2-Llama3 (12.5%)	8B	DPO	6.48	6.83	6.13	20.25 _{1.13}	16.88 _{1.13}
Mamba2-Llama3 (0%)	8B	DPO	5.64	6.16	5.11	14.49 _{0.93}	10.88 _{0.93}
Falcon Mamba Instruct	7B	SFT	6.40	7.25	5.55	4.04 _{0.45}	2.15 _{0.45}
GPT-3.5-turbo	-	RLHF	7.94	-	-	22.70	14.10
GPT-4o	-	RLHF	-	-	-	57.46 _{1.47}	51.33 _{1.47}

Table 2: Chat benchmark results for open-access and proprietary models on MT-Bench and AlpacaEval. MT-Bench scores model responses using GPT-4. AlpacaEval version two measures the win-loss rate between baseline models and GPT-4 scored by GPT-4 Turbo.

Table 2 shows the performance of our models on chat benchmarks compared with large transformer models. The distilled hybrid Mamba model (50%) achieves a similar score in the MT-benchmark as the teacher model, and slightly better than the teacher model on the AlpacaEval benchmark in both LC win rate and overall win rate. The distilled hybrid Mamba (25% and 12.5%) performance is slightly worse than that of the teacher models in the MT benchmark but still surpasses some large transformers even with more parameters in AlpacaEval. The distilled pure (0%) model does degrade significantly in accuracy. Notably, the distilled hybrid model performs better than Falcon Mamba, which was trained from scratch with more than 5T tokens.

5.3 Evaluation on General Benchmarks

Zero Shot Evaluation. We utilize the open-source LM Evaluation Harness library [25] (branch big-refactor) to assess 10 tasks, with the following evaluation metrics: WinoGrande (WG) accuracy [62], PIQA (PQ) accuracy [7], HellaSwag (HS) normalized accuracy [82], ARC-Easy and ARC-Challenge (AE and AC) accuracy and normalized accuracy, [15], MMLU (MM), accuracy [33], OpenBookQA (OB) normalized accuracy [54], TruthFulQA (TQ) accuracy [46], PubMedQA (PM) accuracy [38], and RACE (RA), accuracy [40]. Each task is evaluated by analyzing the probability assigned by the model to each potential answer choice.

Table 3 shows zero shot evaluation in LM Eval benchmark for Mamba and Mamba2 distilled from different teacher models. Both hybrid Mamba-Llama3 and Mamba2-Llama3 models, distilled from the Llama-3 Instruct 8B, perform better compared to the open-source TRI Mamba and Nvidia Mamba models trained from scratch. Performance degrades with more linear RNN layers, but is still competitive at 25% to models trained from scratch.

5.4 Hybrid speculative decoding

Setup We perform speculative decoding using the distilled hybrid models. We run experiments using both Hybrid Mamba 50% and Hybrid Mamba 25% as main models. For the draft models, we train 2 and 4-layer Transformer Draft models on the OpenHermes2.5

Model (% Att)	WG	PI	HS	AE	AC	MM	OB	TQ	PM	RA	AVG
TRI Mamba-7B	71.42	81.01	77.93	77.53	46.67	33.39	46.20	32.09	72.30	37.99	57.65
Nvidia Hybrid Mamba-8B	71.27	79.65	77.68	77.23	47.70	51.46	42.80	38.72	69.80	39.71	59.60
Llama-3.1-8B-Instruct	73.88	80.79	79.21	81.78	55.20	68.12	43.20	42.67	75.20	44.78	64.48
Llama3.1-Mamba (50%)	72.77	79.33	75.91	82.24	53.84	62.13	42.80	40.02	72.00	42.11	62.32
Llama3.1-Mamba-DPO (50%)	73.80	80.41	77.36	84.01	56.57	63.50	44.20	46.07	74.40	43.44	64.38
Llama3.1-Mamba2 (50%)	71.74	78.89	75.36	82.20	52.65	61.01	41.60	40.31	72.60	42.11	61.85
Llama3.1-Mamba2-DPO (50%)	74.11	80.03	79.69	84.81	59.73	59.74	44.00	50.22	74.60	46.12	65.31
Llama-3.2-3B-Instruct	67.48	75.68	70.43	74.07	45.90	60.43	36.00	38.01	69.60	40.67	57.83
Llama3.2-Mamba (50%)	67.32	77.31	70.37	77.65	48.38	54.48	39.40	42.02	66.40	40.29	58.36
Llama3.2-Mamba-DPO (50%)	67.40	77.31	72.56	79.97	52.65	55.09	41.60	48.53	70.00	43.64	60.88
Llama3.2-Mamba2 (50%)	66.06	76.01	69.13	76.68	46.67	53.12	38.80	34.78	63.80	39.81	56.49
Llama3.2-Mamba2-DPO (50%)	67.32	77.69	74.45	80.26	54.10	52.47	42.40	50.28	65.40	43.44	60.78
Mamba-Zephyr (50%)	68.82	80.36	76.91	81.40	55.63	55.43	42.60	41.99	72.60	42.20	61.79
Mamba-Llama3 (50%)	68.98	78.02	78.43	74.45	51.96	57.81	44.00	47.69	73.00	38.56	61.30
Mamba-Llama3 (25%)	62.83	78.07	75.00	74.28	47.35	53.50	40.00	43.64	65.40	36.94	57.70
Mamba-Llama3 (12.5%)	59.75	75.08	71.71	70.58	43.60	49.81	41.40	41.41	62.40	34.45	55.02
Mamba2-Llama3 (50%)	71.51	81.45	79.47	78.83	58.19	55.70	44.20	57.74	72.4	38.85	63.84
Mamba2-Llama3 (25%)	64.80	78.73	77.7	76.35	52.47	53.71	42.40	55.33	64.80	39.23	60.55
Mamba2-Llama3 (12.5%)	63.38	76.82	73.14	75.84	50.26	50.78	39.60	50.00	65.80	36.46	58.21
Mamba2-Llama3 (0%)	58.56	76.82	70.75	74.12	47.95	45.19	39.00	40.20	62.20	32.63	54.74

Table 3: Evaluation on LM Eval benchmark for Mamba and Mamba2 distilled from Llama-3 Instruct 8B.

dataset [70], for approximately 3 full epochs, following the “shrink and fine-tune” approach from [66]. Specifically, we initialize the draft layers using layers from the Zephyr-7B model (we take layers at indices [0, 31] for the 2-layer model and [0, 10, 20, 31] for the 4-layer model), and the embeddings and language model head also from the Zephyr-7B model [72]. We perform loss masking on the prompt, thus only considering next token prediction loss (cross-entropy) on the chat continuations from the training set. Speculative decoding experiments are run on a single NVIDIA RTX 3090 on data from OpenHermes2.5.

Draft Model	K	Target Model (% Att)	# Gen. Tokens	Speedup
2 layers	4	Mamba-Zephyr (50%)	2.48	1.8x
	4	Mamba-Zephyr (25%)	2.64	1.88x
4 layers	4	Mamba-Zephyr (50%)	3	1.81x
	4	Mamba-Zephyr (25%)	3	1.8x
4 layers	3	Mamba-Llama3 (50%)	2.7	1.6x
4 layers	4	Mamba-Llama3 (50%)	3.6	1.58x

Table 4: Performance metrics for different draft and target model configurations for $K = 4$ on data from OpenHermes2.5. # Gen is the average number of generated tokens per speculative decoding step and includes an additional token from the last verifier logits.

Results Table 4 shows results for hybrid speculative decoding with, using both the Zephyr and Llama hybrid models with different configurations. For both the 50% and 25% distilled models, we achieve speedups of over 1.8x on the Zephyr-Hybrid compared to the non-speculative baseline. We also show that the 4-layer draft model we trained achieves a higher acceptance rate, but it adds some additional overhead due to the increased draft model size. For the Llama-hybrid models, the speedups are more modest since the draft model is larger due to the large embedding table of Llama 3. In subsequent work, we will focus on making these draft models smaller.

6 Conclusion

We consider the problem of maintaining LLM abilities while increasing decoding speed through a combination of distillation and speculative decoding. We first show that a Transformer LLM can be used to effectively initialize a Mamba linear RNN model while maintaining original abilities. We then show that through a combination of distillation on supervised instructions and preferences, we can improve the model’s ability with relatively little compute. Finally, we show that the Mamba model can be significantly sped up at inference time through the use of a hardware-aware speculative decoding method. The full model nears LLM chat accuracy, and is accelerated with speculative decoding. We believe these results show that transformer knowledge can be transferred effectively to other architectures, opening up the potential for customizing the inference profile of LLMs beyond optimizing attention.

Acknowledgement

We thank Together AI for providing compute for some of the experiments. This work has benefited from helpful discussions with Albert Gu at CMU, François Fleuret and Vincent Micheli at the University of Geneva, Albert Tseng and Wen-Ding Li at Cornell University.

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. *arXiv preprint arXiv:2312.04927*, 2023.
- [2] S. Arora, S. Eyuboglu, M. Zhang, A. Timalsina, S. Alberti, D. Zinsley, J. Zou, A. Rudra, and C. Ré. Simple linear attention language models balance the recall-throughput tradeoff. *arXiv preprint arXiv:2402.18668*, 2024.
- [3] BAAI. Infinity instruct. *arXiv preprint arXiv:2406.XXXX*, 2024.
- [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. *arXiv preprint arXiv:2405.04517*, 2024.
- [5] N. Bhendawade, I. Belousova, Q. Fu, H. Mason, M. Rastegari, and M. Najibi. Speculative streaming: Fast llm inference without auxiliary models. *arXiv preprint arXiv:2402.11131*, 2024.
- [6] A. Bick, K. Y. Li, E. P. Xing, J. Z. Kolter, and A. Gu. Transformers to ssms: Distilling quadratic knowledge to subquadratic models. *arXiv preprint arXiv:2408.10189*, 2024.
- [7] Y. Bisk, R. Zellers, J. Gao, Y. Choi, et al. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439, 2020.
- [8] A. Botev, S. De, S. L. Smith, A. Fernando, G.-C. Muraru, R. Haroun, L. Berrada, R. Pascanu, P. G. Sessa, R. Dadashi, et al. Recurrentgemma: Moving past transformers for efficient open language models. *arXiv preprint arXiv:2404.07839*, 2024.
- [9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [10] T. Cai, Y. Li, Z. Geng, H. Peng, J. D. Lee, D. Chen, and T. Dao. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*, 2024.
- [11] C. Chen, S. Borgeaud, G. Irving, J.-B. Lespiau, L. Sifre, and J. Jumper. Accelerating Large Language Model Decoding with Speculative Sampling, 2023.
- [12] J. Chen, R. Qadri, Y. Wen, N. Jain, J. Kirchenbauer, T. Zhou, and T. Goldstein. Genqa: Generating millions of instructions from a handful of prompts. *arXiv preprint arXiv:2406.10323*, 2024.
- [13] Z. Chen, X. Yang, J. Lin, C. Sun, J. Huang, and K. C.-C. Chang. Cascade speculative drafting for even faster llm inference. *arXiv preprint arXiv:2312.11462*, 2023.
- [14] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.
- [15] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*, 2018.

- [16] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [17] G. Cui, L. Yuan, N. Ding, G. Yao, W. Zhu, Y. Ni, G. Xie, Z. Liu, and M. Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.
- [18] T. Dao and A. Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.
- [19] S. De, S. L. Smith, A. Fernando, A. Botev, G. Cristian-Muraru, A. Gu, R. Haroun, L. Berrada, Y. Chen, S. Srinivasan, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*, 2024.
- [20] N. Ding, Y. Chen, B. Xu, Y. Qin, S. Hu, Z. Liu, M. Sun, and B. Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, 2023.
- [21] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [22] D. Fu, S. Arora, J. Grogan, I. Johnson, E. S. Eyuboglu, A. Thomas, B. Spector, M. Poli, A. Rudra, and C. Ré. Monarch mixer: A simple sub-quadratic gemm-based architecture. *Advances in Neural Information Processing Systems*, 36, 2024.
- [23] D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, and C. Ré. Hungry hungry hippos: Towards language modeling with state space models. In *The Eleventh International Conference on Learning Representations*, 2022.
- [24] Y. Fu, P. Bailis, I. Stoica, and H. Zhang. Break the sequential dependency of llm inference using lookahead decoding. *arXiv preprint arXiv:2402.02057*, 2024.
- [25] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac’h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou. A framework for few-shot language model evaluation, 12 2023. URL <https://zenodo.org/records/10256836>.
- [26] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.
- [27] A. Gu, K. Goel, and C. Ré. Efficiently Modeling Long Sequences with Structured State Spaces. *arXiv preprint arXiv:2111.00396*, 2021.
- [28] A. Gu, K. Goel, A. Gupta, and C. Ré. On the Parameterization and Initialization of Diagonal State Space Models. *Advances in Neural Information Processing Systems*, 35: 35971–35983, 2022.
- [29] A. Gu, B. Rozière, H. Leather, A. Solar-Lezama, G. Synnaeve, and S. I. Wang. Crux-eval: A benchmark for code reasoning, understanding and execution. *arXiv preprint arXiv:2401.03065*, 2024.
- [30] M. Guo, J. Ainslie, D. Uthus, S. Ontanon, J. Ni, Y.-H. Sung, and Y. Yang. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916*, 2021.
- [31] A. Gupta, A. Gu, and J. Berant. Diagonal State Spaces are as Effective as Structured State Spaces. *Advances in Neural Information Processing Systems*, 35:22982–22994, 2022.
- [32] Z. He, Z. Zhong, T. Cai, J. D. Lee, and D. He. Rest: Retrieval-based speculative decoding. *arXiv preprint arXiv:2311.08252*, 2023.

- [33] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*, 2020.
- [34] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [35] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [36] K. Irie, I. Schlag, R. Csordás, and J. Schmidhuber. Going beyond linear transformers with recurrent fast weight programmers. *Advances in neural information processing systems*, 34:7703–7717, 2021.
- [37] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. I. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- [38] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu. PubMedQA: A Dataset for Biomedical Research Question Answering. *arXiv preprint arXiv:1909.06146*, 2019.
- [39] Y. Kim and A. M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, 2016.
- [40] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy. RACE: Large-scale Reading Comprehension Dataset From Examinations. *arXiv preprint arXiv:1704.04683*, 2017.
- [41] Y. Leviathan, M. Kalman, and Y. Matias. Fast Inference from Transformers via Speculative Decoding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19274–19286. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/leviathan23a.html>.
- [42] R. Li, L. B. Allal, Y. Zi, N. Muennighoff, D. Kocetkov, C. Mou, M. Marone, C. Akiki, J. Li, J. Chim, et al. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*, 2023.
- [43] X. Li, T. Zhang, Y. Dubois, R. Taori, I. Gulrajani, C. Guestrin, P. Liang, and T. B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2023.
- [44] O. Lieber, B. Lenz, H. Bata, G. Cohen, J. Osin, I. Dalmedigos, E. Safahi, S. Meirum, Y. Belinkov, S. Shalev-Shwartz, et al. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.
- [45] B. Y. Lin. ZeroEval: A Unified Framework for Evaluating Language Models, July 2024. URL <https://github.com/yuchenlin/ZeroEval>.
- [46] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
- [47] S. Lin, J. Hilton, and O. Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, 2022.
- [48] X. Liu, L. Hu, P. Bailis, I. Stoica, Z. Deng, A. Cheung, and H. Zhang. Online speculative decoding. *arXiv preprint arXiv:2310.07177*, 2023.
- [49] S. Massaroli, M. Poli, D. Fu, H. Kumbong, R. Parnichkun, D. Romero, A. Timalsina, Q. McIntyre, B. Chen, A. Rudra, et al. Laughing hyena distillery: Extracting compact recurrences from convolutions. *Advances in Neural Information Processing Systems*, 36, 2024.

- [50] H. Mehta, A. Gupta, A. Cutkosky, and B. Neyshabur. Long Range Language Modeling via Gated State Spaces. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=5MkYIYCbva>.
- [51] Y. Meng, M. Xia, and D. Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.
- [52] J. Mercat, I. Vasiljevic, S. Keh, K. Arora, A. Dave, A. Gaidon, and T. Kollar. Linearizing large language models. *arXiv preprint arXiv:2405.06640*, 2024.
- [53] S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.
- [54] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. *arXiv preprint arXiv:1809.02789*, 2018.
- [55] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, H. Cao, X. Cheng, M. Chung, M. Grella, K. K. GV, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.
- [56] B. Peng, J. Quesnelle, H. Fan, and E. Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.
- [57] M. Poli, S. Massaroli, E. Nguyen, D. Y. Fu, T. Dao, S. Baccus, Y. Bengio, S. Ermon, and C. Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pages 28043–28078. PMLR, 2023.
- [58] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.
- [59] T. R. Ralambomihanta, S. Mohammadzadeh, M. S. N. Islam, W. Jabbour, and L. Liang. Scavenging hyena: Distilling transformers into long convolution models. *arXiv preprint arXiv:2401.17574*, 2024.
- [60] L. Ren, Y. Liu, Y. Lu, Y. Shen, C. Liang, and W. Chen. Samba: Simple hybrid state space models for efficient unlimited context language modeling. *arXiv preprint arXiv:2406.07522*, 2024.
- [61] B. Roziere, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, T. Re-me-z, J. Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.
- [62] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.
- [63] I. Schlag, K. Irie, and J. Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning*, pages 9355–9366. PMLR, 2021.
- [64] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [65] U. Shaham, E. Segal, M. Ivgi, A. Efrat, O. Yoran, A. Haviv, A. Gupta, W. Xiong, M. Geva, J. Berant, et al. Scrolls: Standardized comparison over long language sequences. *arXiv preprint arXiv:2201.03533*, 2022.
- [66] S. Shleifer and A. M. Rush. Pre-trained summarization distillation. *CoRR*, abs/2010.13002, 2020. URL <https://arxiv.org/abs/2010.13002>.
- [67] B. Spector and C. Re. Accelerating llm inference with staged speculative decoding. *arXiv preprint arXiv:2308.04623*, 2023.

- [68] Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, J. Wang, and F. Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.
- [69] 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 *International Conference on Learning Representations*, 2020.
- [70] Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. URL <https://huggingface.co/datasets/teknium/OpenHermes-2.5>.
- [71] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [72] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.
- [73] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [74] R. Waleffe, W. Byeon, D. Riach, B. Norick, V. Korthikanti, T. Dao, A. Gu, A. Hatamizadeh, S. Singh, D. Narayanan, et al. An empirical study of mamba-based language models. *arXiv preprint arXiv:2406.07887*, 2024.
- [75] J. Wang, J. N. Yan, A. Gu, and A. M. Rush. Pretraining without attention. *arXiv preprint arXiv:2212.10544*, 2022.
- [76] H. Xia, T. Ge, P. Wang, S.-Q. Chen, F. Wei, and Z. Sui. Speculative Decoding: Exploiting Speculative Execution for Accelerating Seq2seq Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3909–3925, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.257. URL <https://aclanthology.org/2023.findings-emnlp.257>.
- [77] J. Yang, C. E. Jimenez, A. Wettig, K. Lieret, S. Yao, K. Narasimhan, and O. Press. Swe-agent: Agent computer interfaces enable software engineering language models, 2024.
- [78] N. Yang, T. Ge, L. Wang, B. Jiao, D. Jiang, L. Yang, R. Majumder, and F. Wei. Inference with reference: Lossless acceleration of large language models. *arXiv preprint arXiv:2304.04487*, 2023.
- [79] S. Yang, B. Wang, Y. Shen, R. Panda, and Y. Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.
- [80] S. Yang, B. Wang, Y. Zhang, Y. Shen, and Y. Kim. Parallelizing linear transformers with the delta rule over sequence length. *arXiv preprint arXiv:2406.06484*, 2024.
- [81] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.
- [82] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.
- [83] M. Zhang, K. Bhatia, H. Kumbong, and C. Ré. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. *arXiv preprint arXiv:2402.04347*, 2024.
- [84] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

A Evaluation on Long Context Tasks

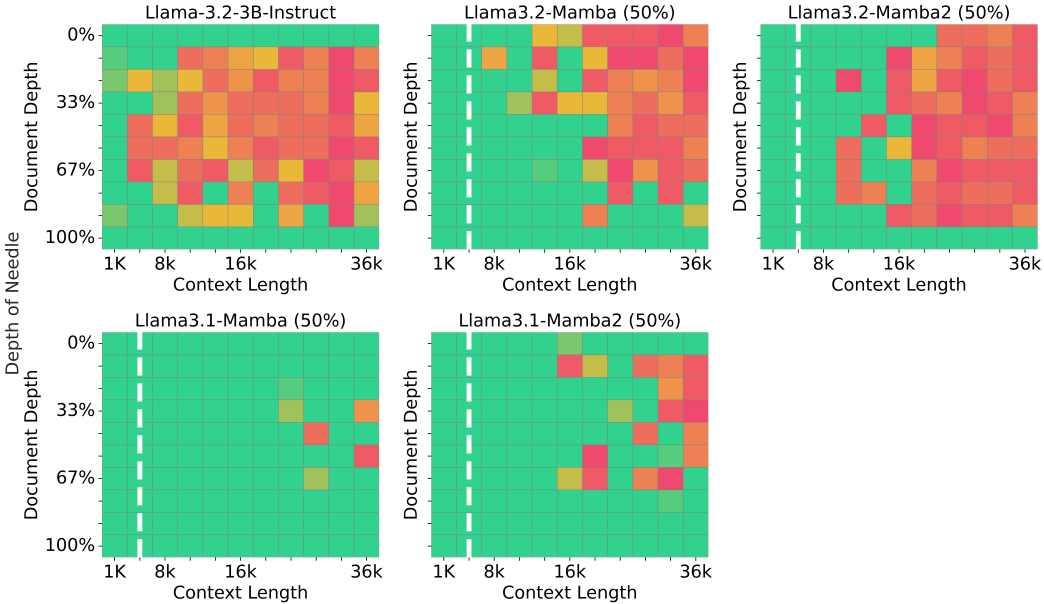


Figure 4: Needle in a Haystack evaluation. Green squares represent a high retrieval success rate, while the white dashed line marks the longest examples encountered during distillation training. The Y-axis indicates the distance to the retrieved target.

Figure 4 illustrates the results of Needle in a Haystack. Although the distillation length is only 2k, our distilled 3B models (Mamba-Llama3.2-3B (50%) and Mamba2-Llama3.2-3B (50%)) achieve perfect accuracy up to 10k, which is better than Llama-3.2-3B-Instruct. Similarly, the distilled 8B models (Mamba-Llama3.1-8B (50%) and Mamba2-Llama3.1-8B (50%)) achieve perfect accuracy up to 16k, with Mamba-Llama3.1-8B demonstrating good results up to 38k.

B Benchmark Evaluation

We also report few-shot evaluations on OpenLLMLeaderboard by conducting 25 shots on ARC-Challenge [14], 10 shots on HellaSwag [82], 5 shots on MMLU [34], and 5 shots on Winogrande [62]. For TruthFulQA, the mc2 metric is reported in this benchmark. For GSM8K [16], we follow the evaluation for instruct tuned model [51], which uses ZeroEval [45], a benchmark designed for chat models. We also include the CRUX [29] from that benchmark, which is designed for evaluating reasoning on code. All models are evaluated with greedy decoding in the ZeroEval.

Table 5 shows that the performance of our distilled hybrid models matches that of the best open-source linear RNN models on the Open LLM Leaderboard, while outperforming their corresponding open-source instruct models in GSM8K and CRUX.

C Analysis

Comparison with other distillation approaches Table 6 (left) compares the perplexity of different model variants. We distill using Ultrachat as seed prompt [20] in one epoch and compare the perplexity. We find that removing more layers gets significantly worse. We also compare our distillation approach with a previous baseline. This approach distills a Transformer model into a Hyena model [57], as proposed in [59]. They use a different distillation approach using progressive knowledge transfer, wherein the student model is trained starting from the first layer and progressively extending to subsequent layers. While

Model (% Att)	ARC	HS	MMLU	WG	TQ	GSM8K	CRUX
Falcon Mamba-7B	62.03	80.82	62.11	73.64	53.42	41.32	8.88
RecurrentGemma-9B	52.00	80.40	60.50	73.60	38.60	38.51	26.25
Mamba-Llama3 (50%)	56.57	78.99	59.26	69.06	58.85	67.85	27.88
Mamba-Llama3 (25%)	55.03	75.66	52.68	62.83	55.03	40.64	15.62
Mamba-Llama3 (12.5%)	52.90	72.46	49.20	59.19	53.00	26.91	11.25
Mamba2-Llama3 (50%)	60.41	77.97	56.67	71.35	66.60	59.36	24.88
Mamba2-Llama3 (25%)	59.22	76.88	53.94	64.88	64.64	38.13	13.25
Mamba2-Llama3 (12.5%)	53.33	72.16	50.85	63.61	61.12	35.03	10.25
Mamba2-Llama3 (0%)	53.51	70.31	44.21	58.91	52.31	-	-

Table 5: Results on the Open LLM Leaderboard and ZeroEval Leaderboard. For GSM8K and CRUX, we chose the zero-shot evaluation using ZeroEval, which is designed for evaluating instruct models. We evaluated the corresponding instruct-tuned models for Falcon Mamba-7b and RecurrentGemma-9B, specifically Falcon Mamba-7b-instruct and RecurrentGemma-9B-it.

Model (% Att)	PPL	Ratio	Model	Hyb Mamba (50% Att)	Hyb Mamba (25% Att)
Teacher: Zephyr (7B)	2.02	1	Dis	5.55	5.01
Mamba-Zephyr (50%)	2.09	1.03	Dis+SFT	5.61	4.97
Mamba-Zephyr (25%)	2.20	1.09	Dis+DPO	5.42	4.84
Mamba-Zephyr (6.25%)	2.46	1.22	Dis+SFT+DPO	6.69	6.10
Mamba-Zephyr (0%)	3.36	1.66			
Teacher: Pythia (70M)	51.4	1			
Distill Hyena	121.2	2.36			

Table 6: (Left) Perplexity comparison between our distillation approach and [59]. (Right) Ablation study of different alignment methods of the Distilled Hybrid Mamba on the MT-benchmark using OpenHermes 2.5 as the SFT dataset.

it is challenging to compare, our distill shows a smaller degradation (1.03 for 50 % attention, 1.09 for 25 % attention, 1.22 for 6.35% attention, and 3.36 for no attention), while the Distill Hyena model is trained in WikiText [53] dataset with a much smaller model and shows large perplexity degrade.

Does distilling from preferences help? In Table 6 (Right), we show the impact of different steps in the alignment process of the distillation. We observe that SFT or DPO alone does not yield much improvement, while SFT + DPO yields the best score. Models are trained using Zephyr as the teacher model and the OpenHermes 2.5 [70] dataset as the SFT dataset, and UltraFeedback [17] as the DPO dataset.

Pseudo Label Distillation Ablations. We consider several different model ablation studies in Table 7. For these experiments we consider training for 5k steps using the pseudo-label approaches on the Ultrachat [20] dataset. Table 7 (Left) presents the results of distillation with various initializations. According to this table, initializing weights from a transformer

Model	Mamba (0% Att)		Hyb Mamba (50% Att)		Model	Hyb Mamba (25% Att)		Hyb Mamba (50% Att)	
	Froz	-Froz	Froz	-Froz		Step	-Step	Step	-Step
+ Attention-Init	3.36	66.7	2.09	9.1	+ Interleave	2.20	2.29	2.09	-
-Attention-Init	18.2	20.3	7.4	11.2	-Interleave	2.89	-	2.41	-

Table 7: (Left) Perplexity comparison with different initialization at first stage. (Right) Perplexity comparison with different Mamba interleaving layers and stepwise distillation at first stage.

is crucial for performance. Without weight initialization from a transformer, perplexity significantly worsens for both pure Mamba models and hybrid models. Also, freezing MLP layers can help the student model focus on learning the interaction of tokens and better mimic attention layers. Table 7 (Right) shows also see smaller benefits from progressive distillation and interleaving the attention layers with Mamba.

Attention Initialization. We compare the default random initialization of Mamba with reusing the linear projection from the attention using the same recipe. Both models are trained using Zephyr as the teacher model and the OpenHermes 2.5 [70] dataset as the SFT dataset, and UltraFeedback [17] as the DPO dataset.

Model	LAMBADA (ppl)	MMLU	ARC-C	TruthfulQA	HellaSwag	MT-Bench (score)	AlpacaEval (LC win %)
+ Attention init	6.20	47.98	49.15	46.67	75.07	6.69	14.11
- Attention init	55.01	26.21	25.26	34.01	27.91	1.04	0.02

Table 8: Performance of Zephyr-Mamba (50% attention) with different initialization.

Table 8 compares the performance of the hybrid model using two different initialization methods: default random initialization and reusing the linear projection from the attention. The model performs significantly better with reusing the linear projection from the attention compared to random initialization, across all evaluated benchmarks. This result confirms that initialization from attention weights is critical.

Model	LAMBADA (ppl)	MMLU	ARC-C	TruthfulQA	HellaSwag	MT-Bench (score)	AlpacaEval (LC win %)
50% Att w Mamba	6.20	47.98	49.15	46.67	75.07	6.69	14.11
50% Att w/o Mamba	151.98	24.46	21.93	32.39	27.91	1.01	0

Table 9: Performance of Hybrid-Mamba with different initialization.

Necessity of Linear RNN. We train a model that removes Mamba blocks from the model entirely using the same recipe to see if the model can adapt. Both models are trained using Zephyr as the teacher model, with the OpenHermes 2.5 [70] dataset as the SFT dataset and UltraFeedback [17] as the DPO dataset. Table 9 compares the performance of the model with and without Mamba blocks. The model with Mamba performs significantly better than the one without it. This confirms that adding Mamba layers is critical and that the improved performance is not solely attributable to the remaining attention mechanism.

D Related Work

Attention-free models. Attention-free models offer improved computational and memory efficiency, making them increasingly popular for various language processing tasks, including autoregressive language modeling. Models like S4 [27] and its subsequent variants [31, 28] have shown promising results in long-range synthetic tasks [69]. Gated SSM architectures, such as GSS [50] and BiGS [75], incorporate a gating mechanism into SSMs for (bidirectional) language modeling. The recently introduced Mamba model [26] argues that the static dynamics of these methods fail to incorporate input-specific context selection within the hidden state, which could be crucial for tasks like language modeling. Mamba has been shown to outperform Transformers across different model sizes and scales. Additionally, several other sub-quadratic model architectures [57, 79, 19, 1, 2, 22, 4, 80] and hybrid architectures [23, 44] have also been proposed.

Distillation from Transformers. There have been relatively few attempts to distill on to linear RNN style models. Laughing Hyena [49] proposes to distill the long convolution into a state space representation, enabling constant time inference in Hyena [57]. Ralambomihanta et al. [59] introduces a progressive knowledge approach to distill small transformer models (70M) into Hyena models. [6]

Speculative Decoding. Speculative decoding [67, 41, 11, 76, 10] has recently emerged as a promising method to accelerate the inference process of large language models, particularly

Transformers. This approach utilizes a smaller draft model to speculatively generate candidate tokens, which the larger target model then verifies. Leviathan et al. [41], Chen et al. [11] proposed a rejection sampling scheme to improve inference quality, while Spector and Re [67] organized candidate tokens into a tree structure to enable more efficient verification. Subsequent work has examined both trained draft models [5, 13, 48] and training-free draft models [32, 78, 24].

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- Delete this instruction block, but keep the section heading "NeurIPS paper checklist",
- Keep the checklist subsection headings, questions/answers and guidelines below.
- Do not modify the questions and only use the provided macros for your answers.

1. Claims

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

Answer: [Yes]

Justification: The claims made in the abstract and introduction are supported by our theoretical analysis in section 2 and experimental results in section 5.

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 discuss the limitation of this paper in Section ??

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 show all the theorems and formulas in Section 2.

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 will release our model and code. Everything will be open-sourced.

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 will definitely release our model and code. Everything will be open-sourced. we will discuss the detailed steps and resources required for reproduction in Section 5.

Guidelines:

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

6. Experimental Setting/Details

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

Answer: [Yes]

Justification: We report the training details, include dataset, hyper-parameters and optimizer and etc in section 5.

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 present the error bars in the evaluation benchmarks in Tables 2 and 5, following the standard for these benchmarks.

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 report the number of GPUs and the training time for GPUs in Section 5. Given the extensive GPU time required for LLM training, we use only a very limited amount, which is conducive to reproduction.

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 reviewed the NeurIPS Code of Ethics and respect the NeurIPS Code of Ethics.

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 discuss Broader Impacts in section ??.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

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

Answer: [\[Yes\]](#)

Justification: We discuss risk of pretrained language models in section ??.

Guidelines:

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

12. Licenses for existing assets

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

Answer: [\[Yes\]](#)

Justification: We cite the original paper that produced the code package or dataset.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

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

Answer: [Yes]

Justification: We communicate the details of the dataset/code/model as part of their submissions via structured templates.

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: The paper does not involve crowdsourcing 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: The paper does not involve crowdsourcing 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.