Distil-xLSTM: Learning Attention Mechanisms through Recurrent Structures

Anonymous ACL submission

Abstract

The current era of Natural Language Processing (NLP) is dominated by Transformer models. However, novel architectures relying on recurrent mechanisms, such as xLSTM and Mamba, have been proposed as alternatives to attention-based models. Although computation is done differently than with the attention mechanism[1], these recurrent models yield good results and sometimes even outperform state-of-the-art attention-based models. In this work, we propose Distil-xLSTM, an xLSTM-based Small Language Model (SLM) trained by distilling knowledge from a Large Language Model (LLM) that shows promising results while being compute and scale efficient. Our Distil-xLSTM focuses on approximating a transformer-based model attention parametrization using its recurrent sequence mixing components and shows good results with minimal training.

1 Introduction

Large Language Models (LLMs) have become central to modern NLP research, with state-of-the-art models such as Mistral (Jiang et al., 2023) and LLaMA (Touvron et al., 2023) demonstrating impressive capabilities across a wide range of tasks. While recent trends show growing interest in training these models at a smaller scale evident in efforts like Phi (Abdin et al., 2024) transformer-based architectures still suffer from a core limitation: the quadratic complexity of their self-attention mechanism (Vaswani et al., 2017). This constraint hinders scalability and deployment in resource-limited settings, despite their success in domains such as code generation (Hui et al., 2024) and multimodal tasks (Wang et al., 2024).

To address these limitations, alternative architectures based on recurrent mechanisms have emerged. Notably, Mamba (Dao and Gu, 2024) and xLSTM (Beck et al., 2024) offer linear time complexity and improved memory efficiency. These models show strong performance across vision, language, and multimodal tasks (Alkin et al., 2024; Anthony et al., 2024; Ren et al., 2024), prompting renewed interest in non-attention-based models. Furthermore, work by (Katharopoulos et al., 2020) showed that causal transformers can be reformulated as recurrent networks, further blurring the lines between attention and recurrence.

In this paper, we explore whether the expressive power of attention-based models can be transferred to a compact recurrent model. Specifically, we introduce **Distil-xLSTM**, a small language model (SLM) built on the xLSTM architecture and trained via knowledge distillation (Hinton et al., 2015) from a transformer-based LLM. xLSTM's enhanced memory mixing and parallel processing capabilities allow it to approximate attention-like behavior without relying on quadratic operations.

Unlike conventional distillation pipelines that transfer knowledge within the same architectural family, our work investigates *cross-architecture distillation* from transformer to recurrent raising the question: can attention dynamics be learned by a recurrent model with a fraction of the complexity?

Our contributions are threefold:

- We propose Distil-xLSTM, the first xLSTM-based SLM distilled from a transformer LLM.
- We introduce a dual-annealing distillation loss that adapts over time, helping the stu-

[1] In this paper, attention refers to self-attention (Vaswani et al., 2017)

dent bridge the architectural and capacity gap.

- We demonstrate that recurrent models can emulate attention behavior efficiently, achieving competitive results with minimal compute.

The remainder of this paper is structured as follows. Section 2 reviews key concepts. Section 4 details our approach, followed by experiments in Section 5. We discuss related work in Section 3 and conclude in Section 6.

2 Background

Transformer-based models rely on self-attention mechanisms (Vaswani et al., 2017), which allow each token to contextualize itself based on all others in the sequence. While highly effective, the quadratic complexity with respect to sequence length makes this approach computationally expensive, particularly for long-context scenarios.

Recurrent models such as LSTMs (Hochreiter and Schmidhuber, 1997) offer an alternative with constant memory and time per step, but struggle with long-term dependencies. The Extended LSTM (xLSTM) architecture (Beck et al., 2024) addresses these limitations by introducing two key innovations. The sLSTM variant introduces scalar memory with a novel memory mixing approach and stabilization mechanisms to improve gradient flow. The mLSTM variant generalizes memory to a matrix form, enabling parallel content-based memory access using learned key, value, and query projections. Together, these enhancements allow xLSTM to scale effectively while retaining recurrent advantages.

To further improve scalability, we consider knowledge distillation (Hinton et al., 2015), a training strategy where a compact student model learns to mimic a larger teacher model. By aligning the student's output distribution with softened teacher outputs often using Kullback-Leibler divergence in combination with standard cross-entropy the student inherits its performance traits with reduced computational cost.

3 Related Work

Knowledge distillation. The Born-Again Multi-task (BAM) framework (Clark et al., 2019) introduced teacher annealing for multi-task learning, where a student model transitions from soft targets to hard labels as training progresses. Annealing-KD (Jafari et al., 2021) extended this idea by reducing temperature over epochs to better align the teacher's signal with the student's limited capacity.

Our proposed δ-distillation builds on these insights by annealing both the soft target weight (α) and temperature (T) over time. This dual-annealing mechanism allows the student to gradually internalize the teacher's dark knowledge, achieving both performance gains and effective compression.

Distillation for architecture simplification. Bick et al. (Bick et al., 2025) proposed MOHAWK, a method for distilling transformers into Mamba-based hybrids through staged parameter reuse and alignment. While their focus is on hybrid architectures, our work targets purely recurrent models specifically xLSTMs.

In contrast to MOHAWK, δ-distillation transfers only the embedding and classification layers, preserving architectural independence while benefiting from teacher guidance. This choice supports deployment in low-resource or transformer-incompatible environments.

Table 1 summarizes key differences.

4 δ-Distillation Process

Contemporary state-of-the-art language models can be conceptualized as comprising three principal components: an **embedding layer**, **attention blocks** (facilitating sequence mixing), and the **classification head** (functioning as the channel mixer) (Bick et al., 2025). The attention blocks constitute the critical architecture underlying these models' efficacy, wherein intricate token relationships are learned, effectively capturing dependencies within the input sequence.

Informed by this framework of sequence and channel mixers, we hypothesize that a recurrent model architecture, specifically one predicated on xLSTM, can approximate the internal representations generated by the attention layers of a transformer. This hypothesis ex-

Table 1: Comparison of δ-distillation with related works.

Method	**Teacher Architecture**	**Student Architecture**	**Goal**
Teacher Annealing (Clark et al., 2019)	Transformer	Transformer	Student performance improvement
Annealing-KD (Jafari et al., 2021)	Transformer	Transformer	Teacher knowledge compression
MOHAWK (Bick et al., 2025)	Transformer	Mamba/Hybrid	Block-wise matrix alignment
δ-distillation	Transformer	xLSTM	Student performance improvement and teacher hidden parametrization approximation

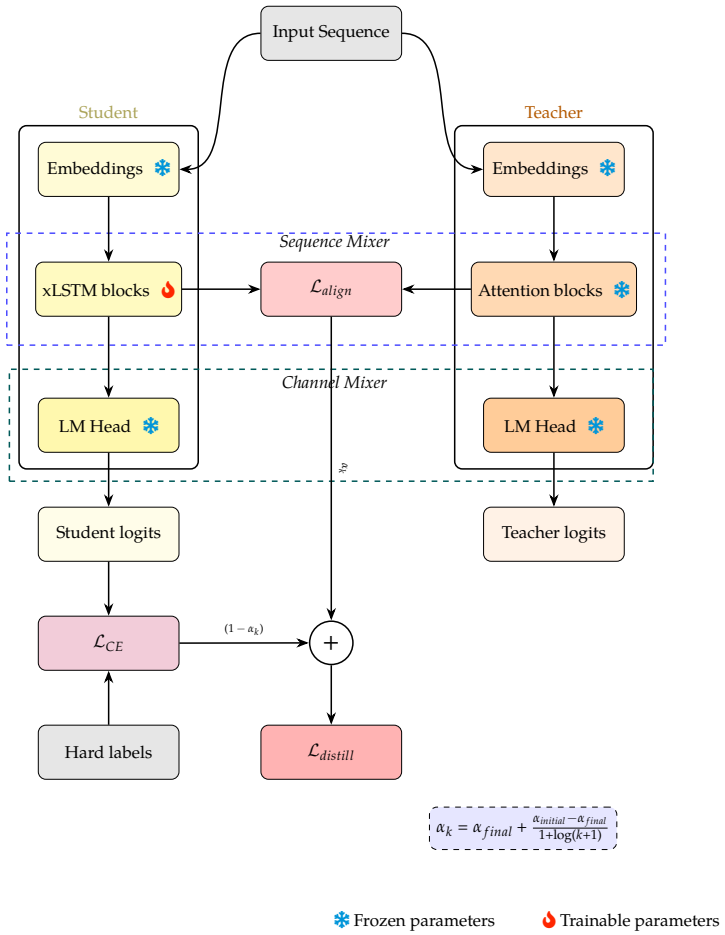

Figure 1: Proposed distillation framework

tends the foundational work of Katharopoulos et al. (Katharopoulos et al., 2020), who demonstrated that transformer layers with causal masking can be reformulated as linear recurrent neural networks, with recurrence considered temporally. Their analysis reframes self-attention operations into row-wise computations, establishing a theoretical basis for modeling attention mechanisms through recurrent architectures. Through the linearization of attention via kernel-based methodologies, they established a framework for approximating attention mechanisms without incurring quadratic computational complexity.

Building upon this theoretical foundation, our distillation framework (illustrated in Fig-

ure 1) adopts a novel methodological approach: utilizing the teacher model's embedding layer and classification head weights to initialize the corresponding components in the student model. This initialization strategy presupposes that the teacher's parameters for these components have achieved optimal or near-optimal configurations. Consequently, our primary investigative focus shifts to approximating the teacher's sequence mixer, specifically its attention blocks exclusively through xLSTM blocks. This architectural design simplifies the distillation process while ensuring that the student model maintains the capacity to replicate the teacher's rich internal representations. Through this recurrent formulation, our framework bridges the gap between transformer-based and recurrent architectures, while simultaneously demonstrating the feasibility of achieving transformer-comparable performance with computationally more efficient recurrent models.

To address the inherent challenges of knowledge transfer from a transformer to an xLSTM model, we introduce a novel framework termed **δ-distillation**. This methodology reconceptualizes the traditional knowledge distillation paradigm by implementing a time-varying loss function, wherein the scaling parameter α undergoes progressive reduction throughout the training process. This gradual modulation encourages the student model to initially leverage the teacher's dark knowledge and subsequently transition its learning focus toward the hard labels provided by the dataset. Furthermore, rather than instructing the student to emulate the teacher's output distribution, our objective is to enable the student to learn an approximation of the teacher's hidden parametrization.

The fundamental principles of δ-Distillation are articulated as follows:

Progressive Annealing. The parameter α is subjected to annealing within each epoch following a logarithmic schedule, ensuring a smooth decay that facilitates stable gradient propagation. Across successive epochs, α is further reduced by a constant factor δ, thereby diminishing the student's dependence on the teacher over the course of training.

Logarithmic Schedule. The parameter α_k at a given global training step k is computed utilizing the following schedule:

$$\alpha_k = \alpha_{\text{final}} + \frac{\alpha_{\text{initial}} - \alpha_{\text{final}}}{1 + \log(k+1)} \tag{1}$$

Epoch-Wise Decay. After each epoch, α undergoes reduction by a constant factor δ (Equation (2)), ensuring systematic diminution over the entire training period:

$$\alpha \leftarrow \max(\alpha - \delta\alpha, \alpha_{\text{final}}) \tag{2}$$

Convergence Analysis. The limit of the schedule function as $k \to +\infty$ is derived as follows:

$$\lim_{k\to+\infty} \alpha_k = \lim_{k\to+\infty} \left(\alpha_{\text{final}} + \frac{\alpha_{\text{initial}} - \alpha_{\text{final}}}{1 + \log(k+1)} \right) \tag{3}$$

$$= \alpha_{\text{final}} + \underbrace{\lim_{k\to+\infty} \frac{\alpha_{\text{initial}} - \alpha_{\text{final}}}{1 + \log(k+1)}}_{=0} \tag{4}$$

$$= \alpha_{\text{final}} \tag{5}$$

This mathematical formulation ensures that α_k asymptotically converges to its final value, enabling the student to continue receiving minimal guidance while predominantly learning from hard labels.

Time-Varying Loss Function. A central component of δ-distillation is its time-varying loss function that evolves dynamically within and across epochs. Our distillation loss comprises a weighted sum of two distinct components:

Alignment Loss ($\mathcal{L}align$): The primary focus of δ-distillation involves hidden representation approximation; consequently, we employ the mean of layer-wise Frobenius norms between the teacher's and student's hidden states (Equation 7). We subsequently scale $\mathcal{L}_{\text{align}}$ by a factor of $1/\sqrt{||h_S||}$ to mitigate the high magnitude of the Frobenius norm, where $||h_S||$ denotes the number of elements in tensor h_S, representing the hidden states produced by the student model.

Task Loss ($\mathcal{L}_{CE}$): Given that we train Distil-xLSTM for next token prediction, our task loss corresponds to the cross-entropy loss function.

The combined loss function is formally defined as:

$$\mathcal{L}\text{distill} = (1 - \alpha_k) \cdot \mathcal{L}_{\text{CE}} + \alpha_k \cdot \frac{\mathcal{L}_{\text{align}}}{\sqrt{||h_S||}} \tag{6}$$

$$\mathcal{L}\text{align} = \frac{1}{L}\sum_{l=1}^{L}||h_T^{(l)} - h_S^{(l)}||_F \tag{7}$$

Where:

- $\alpha_k \in [0, 1]$: Determines the relative weight attributed to the teacher's guidance.
- $1/\sqrt{||h_S||}$: Functions as a normalization term for the alignment loss.
- L: Represents the number of hidden layers comprising both the teacher and student models.

The distillation process is formally specified in Algorithm 1. To enhance stability during the distillation process, we initialize the student model's sequence mixer using the following methodological approach:

1. **Number of Sequence Mixing Layers:** Let L_T denote the number of attention layers in the teacher's sequence mixer. The student's sequence mixer is initialized with $L_S = L_T$ xLSTM blocks, ensuring that the student model possesses comparable expressive capacity to that of the teacher model.
2. **Number of Heads:** Let H_T denote the number of attention heads within each attention layer of the teacher. Each xLSTM block in the student model is initialized with $H_S = \text{roundup}(H_T, 4)$, where $\text{roundup}(x, k)$ rounds x up to the nearest multiple of k. This parameterization ensures that the number of heads in the student's xLSTM blocks achieves both expressive capacity and computational efficiency.

Through this initialization strategy, we address the following critical considerations:

- **Capacity Matching:** By establishing $L_S = L_T$, the student model attains equivalent depth to the teacher. Given the xLSTM's inherently lower parameter count compared to an attention layer, this approach ensures that the student model can acquire comparable expressive capacity without incurring excessive computational costs.
- **Expressive Attention Mechanisms:** Through the parameterization $H_s = \text{roundup}(H_t, 4)$, the student's xLSTM blocks incorporate sufficient computation heads to effectively emulate the teacher's attention mechanisms while maintaining computational efficiency.

The salient advantages of our approach can be summarized as follows:

- **Dynamic Teacher-Student Balance:** Through the gradual transition from teacher-guided knowledge distillation to pseudo-autonomous learning, the combined loss function ensures that the student model effectively assimilates both the teacher's domain expertise and the inherent structural patterns within the dataset.
- **Enhanced Generalization:** The calibrated equilibrium between $\mathcal{L}_{\text{align}}$ (attention approximation) and $\mathcal{L}_{\text{CE}}$ (hard labels) mitigates overfitting to either the teacher's dark knowledge or the dataset's idiosyncrasies, thereby promoting superior generalization performance on unseen data distributions.
- **Gradient-Stable Learning Progression:** The progressive modulation of α facilitates a stable and controlled transition in learning focus, circumventing abrupt alterations that might otherwise destabilize the optimization trajectory.

The δ-distillation framework thus achieves an optimal balance between teacher guidance and independent learning, enabling efficient knowledge transfer into compact recurrent architectures while preserving model performance.

5 Experimental Results

5.1 Experimental Configuration

We conducted comprehensive training of the Distil-xLSTM model utilizing SmolLM2-360M (Allal et al., 2025) as the teacher model. Experimental procedures were executed on an Nvidia T4 GPU employing FP16 mixed precision training methodology (Micikevicius et al., 2018). The training regimen encompassed

Algorithm 1 δ-Distillation Framework

Input: α_{initial}, α_{final}, $\delta\alpha$, n_{epochs}, steps_per_epoch
for epoch = 1 to n_{epochs} **do**
  for step = 1 to steps_per_epoch **do**
    Perform forward pass and compute the distillation loss:

$$\mathcal{L}_{\text{distill}} = (1 - \alpha_k) \cdot \mathcal{L}_{\text{CE}} + \alpha_k \cdot \frac{\mathcal{L}_{\text{align}}}{\sqrt{||h_S||}}$$

    Perform backward pass and update model parameters
    Update α_k using the schedule:

$$\alpha_k \leftarrow \alpha_{\text{final}} + \frac{\alpha - \alpha_{\text{final}}}{1 + \log(\text{step} + 1)}$$

  end for
  Update α for the next epoch: $\alpha \leftarrow \max(\alpha - \delta\alpha, \alpha_{\text{final}})$
end for

processing 512M tokens extracted from the FineWeb dataset (Lozhkov et al., 2024) over a single epoch.

For our experimental investigations, we primarily employed the mLSTM block architecture within our model, selected for its superior recall capacity, thereby ensuring improved training dynamics. Through the reuse of the embedding layer and classification head weights from the teacher model, our Distil-xLSTM architecture incorporates 32 mLSTM blocks. The resultant model comprises 279M parameters, of which only 184M parameters (approximately 65.94% of the total parameter count), corresponding to the sequence mixer's parameters, are actively trained during the distillation process. This architectural configuration significantly reduces computational training requirements while preserving performance characteristics comparable to the teacher model.

5.2 Training Results

Our experimental results demonstrate the effectiveness of the δ-distillation framework. Throughout the training process spanning 10 epochs, we observed consistent convergence of the total loss, indicating effective knowledge transfer from the teacher model (SmolLM2-360M) to our more efficient Distil-xLSTM architecture. [2]

The cross-entropy loss exhibited steady decline over the training period, demonstrating the student model's increasing proficiency in learning from hard labels. Concurrently, we monitored the alignment loss based on the Frobenius norm, which showed initial fluctuations before stabilizing. This behavior aligns with our δ-distillation methodology, where the model progressively shifts emphasis from teacher guidance to independent learning from training data.

Notably, the gradient norm measurements revealed a significant reduction over the course of training, decreasing from initial values around 10 to stabilize below 8. This reduction indicates that our incorporation of the Frobenius norm effectively stabilized the training process, requiring less aggressive parameter updates while maintaining performance comparable to the teacher model.

The perplexity metrics, as presented in Table 2, offer compelling evidence of our approach's efficacy. On C4 benchmark, Distil-xLSTM achieved a perplexity of 566, positioning it between the teacher model's 373 and the baseline xLSTM's 1576. More remarkably, on the LAMBADA benchmark, our Distil-xLSTM substantially outperformed both the teacher model and the baseline xLSTM, recording a perplexity of 3375 compared to 47953 and 12011, respectively.

These results are particularly noteworthy considering that only 65.94% of Distil-xLSTM's

[2] Figures illustrating the training dynamics have been omitted due to space constraints.

Table 2: Perplexity comparison across language models on C4 and LAMBADA benchmarks. Lower perplexity indicates better performance. Best scores are shown in **bold**.

Model	C4	LAMBADA
Pretrained SmolLM2 (Teacher)	**373**	47953
Distil-xLSTM (Student)	566	**3375**
xLSTM (Baseline)	1576	12011

parameters (184M out of 279M) were actively trained during the distillation process. This efficiency, coupled with the model's strong performance metrics, underscores the effectiveness of our δ-distillation framework in transferring knowledge from a transformer-based teacher to a recurrent architecture while maintaining and in some cases exceeding the performance characteristics of the teacher model.

6 Conclusion

In this work, we introduce **Distil-xLSTM**, an xLSTM-based SLM designed to approximate the attention mechanisms of transformer-based models through cross-architecture knowledge distillation. Our principal contributions are as follows:

- **Cross-Architecture Distillation**: We demonstrate effective knowledge transfer from a transformer-based teacher to a purely recurrent student architecture (xLSTM). This methodological approach bridges the fundamental gap between attention-based and recurrent computational paradigms, thereby enabling efficient model deployment in resource-constrained computational environments.

- **Architectural Innovations**: We leverage xLSTM's enhanced capabilities, specifically the mLSTM block architecture, parallel computation methodologies, and stabilizer states to effectively approximate attention mechanisms. The student model implements a reduced yet expressively powerful architecture, initialized with approximately 22% fewer parameters and optimized computation head configurations derived from the teacher model.

- **Alignment via Frobenius Norm**: We introduce a novel hidden state alignment loss term to facilitate compressed and stabilized knowledge transfer. This mathematical formulation aligns the student's latent representations with those of the teacher, thereby mitigating architectural disparities and enhancing training stability throughout the distillation process.

- **Computational Efficiency**: Our framework achieves significant computational efficiency through strategic weight reuse (specifically the embedding layer and classification head from the teacher model) and minimal trainable parameters (65% of the total parameter count), substantially reducing training computational requirements. Experimental evaluations conducted on 512M tokens with a model comprising 279M parameters demonstrate convergence characteristics comparable to transformer baselines, despite the linear scaling properties inherent to recurrent architectural designs.

While Distil-xLSTM establishes a foundation for cross-architecture knowledge distillation from transformers to recurrent models, several promising directions remain. Extending this framework to other modalities particularly vision could test its generality, especially given the dominance of attention in visual modeling.

Adaptive distillation strategies that tailor knowledge transfer to task complexity and model capacity offer another avenue for boosting efficiency. Finally, scaling experiments across a broader spectrum of model sizes both smaller students and larger teachers will help assess the robustness and applicability of our method in diverse deployment scenarios.

Limitations

While the results are promising, they were achieved on a limited scale due to resource constraints. As part of our future work, we aim to scale up our experiments to larger datasets and more complex tasks, which will further test the robustness and generalizability of Distil-xLSTM. We believe this direction holds significant promise for environments requiring efficient yet capable models, particularly in resource-constrained settings.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint*. ArXiv:2404.14219.

Benedikt Alkin, Maximilian Beck, Korbinian Pöppel, Sepp Hochreiter, and Johannes Brandstetter. 2024. Vision-LSTM: xLSTM as Generic Vision Backbone. *arXiv preprint*. ArXiv:2406.04303.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Quentin Anthony, Yury Tokpanov, Paolo Glorioso, and Beren Millidge. 2024. BlackMamba: Mixture of Experts for State-Space Models. *arXiv preprint*. ArXiv:2402.01771 version: 1.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. 2024. xlstm: Extended long short-term memory. In *Thirty-eighth Conference on Neural Information Processing Systems*.

Aviv Bick, Kevin Li, Eric Xing, J Zico Kolter, and Albert Gu. 2025. Transformers to ssms: Distilling quadratic knowledge to subquadratic models. *Advances in Neural Information Processing Systems*, 37:31788–31812.

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D Manning, and Quoc Le. 2019. Bam! born-again multi-task networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5931–5937.

Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint*. ArXiv:1503.02531.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, and 5 others. 2024. Qwen2.5-Coder Technical Report. *arXiv preprint*. ArXiv:2409.12186.

Aref Jafari, Mehdi Rezagholizadeh, Pranav Sharma, and Ali Ghodsi. 2021. Annealing knowledge distillation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2493–2504.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint*. ArXiv:2310.06825.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. *Preprint*, arXiv:1710.03740.

Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. 2024. Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling. *arXiv preprint*. ArXiv:2406.07522 version: 1.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*. ArXiv:2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint*. ArXiv:2409.12191.