LENGTH GENERALIZATION NEEDS PROPER REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Length generalization is the ability of sequential models to perform well on context lengths unseen during training. In this work, we show that the challenge of achieving length generalization is related not only to architectural choices such as positional encoding and the attention mechanism but also to the training procedure itself. We study how regularization affects length generalization and find that weight decay can hinder extrapolation. In contrast, dropout improves extrapolation when its placement within the architecture is reconsidered. In that regard, we show that the standard placement of dropout before layer normalization introduces a systematic distributional mismatch, and that applying dropout just before the linear projection resolves this issue. For example, a modified SmolLM3 with sliding window attention, continually pre-trained with dropout, can extrapolate perfectly to 64$\times$ on Needle-in-a-Haystack and far beyond the pre-training context size on RULER and HELMET. Mamba2 also benefits from dropout, suggesting an architecture-agnostic nature of the problem. We further propose Variance-Preserving Affine Dropout (VPAD), a new dropout strategy that substantially reduces the resulting pre-activation variance mismatch, leading to further extrapolation improvements in transformers. [1]

1 INTRODUCTION

As we demand that large language models (LLMs) operate over increasingly longer contexts, length generalization remains a central challenge for long-context modeling. To avoid the challenges of length generalization, the community relies not on the extrapolation ability of models but on continued training on longer contexts after pre-training on much shorter ones (Fu et al., 2024; Gao et al., 2025; NVIDIA et al., 2025; 2026; Olmo et al., 2026). The phenomenon of poor length extrapolation is ubiquitous across the most widely used architectures, such as transformers and linear RNNs (Pitorro et al., 2024; Buitrago & Gu, 2025; Ben-Kish et al., 2025a; Jelassi et al., 2024). For transformers, key obstacles leading to poor out-of-distribution (OOD) performance are the Positional Encoding (PE) (Press et al., 2022; Kazemnejad et al., 2023; Ruoss et al., 2023; Su et al., 2024) and the attention mechanism itself (Veličković et al., 2025; Barbero et al., 2024; Vasylenko et al., 2026). For linear RNNs, these challenges have been linked to their fixed-size memory mechanism (Jelassi et al., 2024; Arora et al., 2024) and, as recently observed by Buitrago & Gu (2025), to a train-test mismatch in the distribution of recurrent states.

In this work, we show that the problem of length generalization is more than an architectural challenge and can be related to overfitting to the in-distribution (ID) length, i.e., the pre-training context length. Buitrago & Gu (2025) showed that training linear RNNs for longer worsens length extrapolation. We demonstrate that longer training degrades length extrapolation not only in linear RNNs but in transformers as well. Moreover, we show that, despite being a standard regularization technique, weight decay (WD) significantly speeds up the deterioration. This eventually leads to a complete loss of extrapolation. In contrast, regular dropout (Srivastava et al., 2014) offers significant benefits, but its standard placement in common implementations leads to a mismatch of a neuron's pre-activation statistics between training and inference: normalization applied to the masked input attenuates the mean of the signal reaching the linear projection. We show that dropout must instead be placed just before the linear projection, so that normalization acts on a non-masked

[1] Our code and models will be made publicly available upon acceptance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

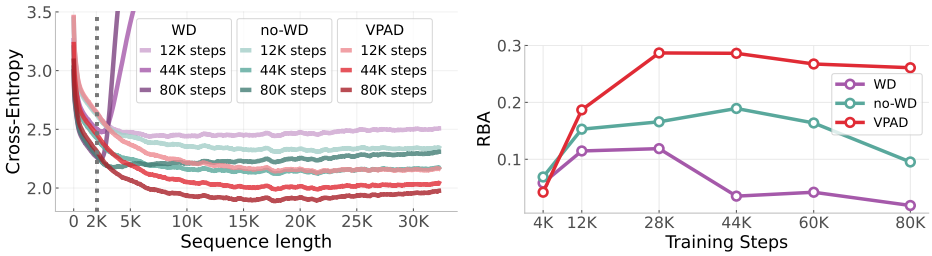

Figure 1: Transformer performance across training steps. **Left**: per-position CE loss (512-window moving average); the dotted line denotes the pre-training context length (2K). **Right**: RULER extrapolation, RBA averaged over 4–8K. Training with WD leads to overfitting to the pre-training context length. In contrast, dropout (VPAD) substantially mitigates the issue.

input. Additionally, because regular dropout still leads to a variance mismatch, we derive Variance-Preserving Affine Dropout (VPAD)—a generalization of Alpha-Dropout (Klambauer et al., 2017) to non-centered activations—which also preserves the variance of the input to the linear projection. We show that both dropout variants substantially improve the extrapolation of transformers. For instance, the SmolLM3 checkpoint pre-trained on 4K context length and continually pre-trained on the same length with dropout extrapolates perfectly to $64\times$ on Needle-in-a-Haystack and far beyond its pre-training context on RULER and HELMET. Similarly, Mamba2 benefits, improving its extrapolation by at least $4\times$ on RULER and HELMET. Overall, we show that *length-based overfitting* affects multiple common architectures, and proper regularization can substantially mitigate it.

2 Longer Training Hinders Length Generalization

We hypothesize that the training duration affects length generalization in transformers, mirroring the phenomenon observed for linear RNNs by Buitrago & Gu (2025): a model trained on a small number of tokens extrapolates better than one trained for longer. Fig. 1 shows that longer training inevitably deteriorates extrapolation. In particular, the prevalent training approach with weight decay leads to the most severe outcome: in both Cross-Entropy (CE) and Recall-Based Accuracy (RBA) on the RULER benchmark (Hsieh et al., 2024), optimal performance is reached at around 12–28K steps, after which the model degrades. Interestingly, training without weight decay (no-WD) substantially postpones this effect. Optimal performance moves to 44K, with a lower and flatter CE loss and a 55% higher RULER score. Eventually, however, longer training degrades the no-WD model as well: RBA drops to almost half of its maximum by the end of training.

This experiment confirms that transformers are also affected by the issue in Buitrago & Gu (2025). We hypothesize that models overfit to the training context length, and this can be mitigated by using methods aimed at improving generalization more broadly. In the following sections, we show that dropout substantially alleviates the problem, allowing models to extrapolate much better.

3 Background

Transformers and Mamba2 Blocks. Transformers (Vaswani et al., 2017) and Mamba2 (Dao & Gu, 2024) both process an input sequence by stacking L identical blocks. In transformers, each block is composed of a self-attention sublayer and a position-wise feed-forward network (FFN) sublayer, both wrapped in residual connections together with a normalization layer. In modern transformers such as GPT-2 (Radford et al., 2019), LLaMA (Touvron et al., 2023), Mistral (Jiang et al., 2023), and Qwen (Yang et al., 2025a), the full block performs the following:

$$\boldsymbol{X}^i = \boldsymbol{R}^{i-1} + \mathrm{Attn}(\mathrm{Norm}(\boldsymbol{X}^{i-1})), \quad \boldsymbol{X}^{i+1} = \boldsymbol{R}^i + \mathrm{FFN}(\mathrm{Norm}(\boldsymbol{X}^i)), \tag{1}$$

where $\boldsymbol{R}^i = \boldsymbol{X}^i$, for easiness of notation, is the residual stream of sublayer $i \in \{1, \dots, 2L\}$; $\mathrm{Norm}(\cdot)$ is normalization applied row-wise to its input. Mamba2 follows the same convention, wrapping each block as $\boldsymbol{X}^i = \boldsymbol{R}^{i-1} + \mathrm{Mamba2}(\mathrm{Norm}(\boldsymbol{X}^{i-1}))$.

In both architectures, each sublayer first applies a linear projection to its input before any nonlinear processing. Given the normalized input $\hat{\boldsymbol{X}} = \mathrm{Norm}(\boldsymbol{X}) \in \mathbb{R}^{n\times d}$ and input projections $\boldsymbol{W}_{\{Q,K,V\}} \in \mathbb{R}^{d\times d}$, the attention sublayer—single-head for brevity[2]—computes

$$\mathrm{Attn}: \ [\boldsymbol{Y}_Q, \boldsymbol{Y}_K, \boldsymbol{Y}_V] = [\hat{\boldsymbol{X}}\boldsymbol{W}_Q, \hat{\boldsymbol{X}}\boldsymbol{W}_K, \hat{\boldsymbol{X}}\boldsymbol{W}_V], \quad \boldsymbol{U}_{\mathrm{Attn}} = \mathrm{softmax}\big(s\boldsymbol{Y}_Q\boldsymbol{Y}_K^\top\big)\,\boldsymbol{Y}_V\,\boldsymbol{W}_P, \quad (2)$$

where $\boldsymbol{W}_P \in \mathbb{R}^{d\times d}$ is the output projection; s is an attention scaling factor which is usually set to $d^{-1/2}$ but can be length-dependent (Chiang & Cholak, 2022; Nakanishi, 2025; Vasylenko et al., 2026). For the FFN, the input projections are $\boldsymbol{W}_{\{G,U\}} \in \mathbb{R}^{d\times d_{\mathrm{ffn}}}$, and the sublayer computes

$$\mathrm{FFN}: \ [\boldsymbol{Y}_{\mathrm{gate}}, \boldsymbol{Y}_{\mathrm{up}}] = [\hat{\boldsymbol{X}}\boldsymbol{W}_G, \hat{\boldsymbol{X}}\boldsymbol{W}_U], \quad \boldsymbol{U}_{\mathrm{FFN}} = \big(\phi(\boldsymbol{Y}_{\mathrm{gate}}) \odot \boldsymbol{Y}_{\mathrm{up}}\big)\,\boldsymbol{W}_D, \quad (3)$$

where ϕ is an elementwise nonlinearity and $\boldsymbol{W}_D \in \mathbb{R}^{d_{\mathrm{ffn}}\times d}$ is the down projection. In Mamba2, an input projection $\boldsymbol{W}_{\mathrm{in}} \in \mathbb{R}^{d\times \tilde{d}}$ produces the gate $\boldsymbol{Z}$, the SSM input $\boldsymbol{X}_{\mathrm{ssm}}$, the state matrices $\boldsymbol{B}$ and $\boldsymbol{C}$, and the step size $\boldsymbol{\Delta}$. Omitting the short convolution, the output normalization, and the output projection for clarity, and using the SSM-attention duality, the block computes

$$\boldsymbol{Y}_{\mathrm{SSD}} = [\boldsymbol{Z}, \boldsymbol{X}_{\mathrm{ssm}}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\Delta}] \ = \ \hat{\boldsymbol{X}}\boldsymbol{W}_{\mathrm{in}}, \ \mathrm{Mamba2}(\boldsymbol{Y}_{\mathrm{SSD}}) \ = \ \big((\boldsymbol{L}\odot\boldsymbol{C}\boldsymbol{B}^\top)\,\boldsymbol{X}_{\mathrm{ssm}}\big)\odot\phi(\boldsymbol{Z}), \quad (4)$$

where $\boldsymbol{L} \in \mathbb{R}^{n\times n}$ is a causal decay mask built from $\boldsymbol{\Delta}$ and a learned decay rate.

RMSNorm. The original choice of $\mathrm{Norm}(\cdot)$ is LayerNorm (Ba et al., 2016), but modern LLMs conventionally replace it with Root Mean Square Normalization (Zhang & Sennrich, 2019):

$$\mathrm{RMSNorm}(\boldsymbol{x}) \ = \ \boldsymbol{\gamma} \odot \frac{\boldsymbol{x}}{\mathrm{RMS}(\boldsymbol{x}) + \epsilon}, \qquad \mathrm{RMS}(\boldsymbol{x}) \ = \ \sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}. \quad (5)$$

Dropout. Dropout (Srivastava et al., 2014) is a regularization method used in many transformer architectures. At training time, dropout replaces each element of a hidden vector $\boldsymbol{u} \in \mathbb{R}^d$ by $\mathrm{Drop}(\boldsymbol{u})_i = \frac{b_i}{1-p}\,u_i$, where $b_1, \ldots, b_d \sim \mathrm{Bernoulli}(1-p)$ are i.i.d. masks and $p \in [0, 1)$ is the drop probability. The $1/(1-p)$ rescaling (known as "inverted dropout") preserves the expectation of pre-activations during training, $\mathbb{E}[\boldsymbol{u}] = \mathbb{E}[\mathrm{Drop}(\boldsymbol{u})]$. At inference time, no masking is applied. Within a transformer block, dropout is conventionally applied to the output of each sublayer just before the residual addition, so any two successive sublayers execute:

$$\tilde{\boldsymbol{X}}^i \ = \ \boldsymbol{R}^{i-1} + \mathrm{Drop}\big(\underbrace{\mathrm{Attn}(\mathrm{Norm}(\tilde{\boldsymbol{X}}^{i-1}))}_{\boldsymbol{U}_{\mathrm{Attn}}}\big), \quad \tilde{\boldsymbol{X}}^{i+1} \ = \ \boldsymbol{R}^i + \mathrm{Drop}\big(\underbrace{\mathrm{FFN}(\mathrm{Norm}(\tilde{\boldsymbol{X}}^i))}_{\boldsymbol{U}_{\mathrm{FFN}}}\big), \quad (6)$$

where $\mathrm{Drop}(\cdot)$ is dropout applied row-wise to its input with an independent mask $\boldsymbol{b}^j$ for each row j. As a result, the effect of dropout on the next linear layer (e.g. from an attention or feedforward sublayer) is mediated through an RMSNorm (Fig. 2, Left). We revisit this placement—Pre-Norm dropout—in Sec. 4 and show that it introduces a systematic mismatch between training-time and inference-time pre-activation statistics. As we will show, dropout can be placed just before linear projections to mitigate the problem (Pre-Linear dropout, Sec. 4.3; see also Fig. 2, Right):

$$\boldsymbol{X}^i \ = \ \boldsymbol{R}^{i-1} + \underbrace{\mathrm{Attn}(\mathrm{Drop}(\mathrm{Norm}(\boldsymbol{X}^{i-1})))}_{\boldsymbol{U}_{\mathrm{Attn}}}, \quad \boldsymbol{X}^{i+1} \ = \ \boldsymbol{R}^i + \underbrace{\mathrm{FFN}(\mathrm{Drop}(\mathrm{Norm}(\boldsymbol{X}^i)))}_{\boldsymbol{U}_{\mathrm{FFN}}}. \quad (7)$$

4 Dropout and Train-Test Mismatch

In common transformer implementations, such as in HuggingFace transformers (Wolf et al., 2020) and Megatron (Shoeybi et al., 2019), dropout is typically applied before adding the residual (Eq. 6), that is, before the normalization layer of the next transformer block. In Sec. 4.2, we show that this placement—*Pre-Norm* dropout—introduces a systematic mismatch between the pre-activation means seen at training and at inference. Sec. 4.3 shows that *Pre-Linear* dropout (Eq. 7), which masks the normalized representation just before the linear layer, does not suffer from the issue. Finally, Sec. 4.4 describes a variance-preserving variant of Pre-Linear dropout that also reduce the variance matches. Full proofs of the propositions below are given in App. B.

[2]Our analysis is not affected by the number of heads.

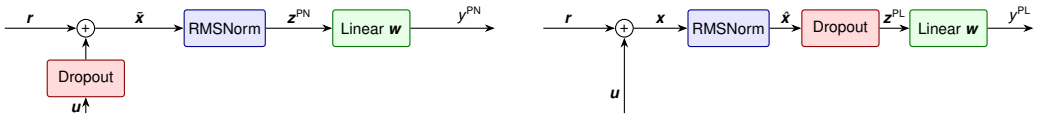

Figure 2: Dropout placements. **Left**: Pre-Norm dropout (Eq. 6). **Right**: Pre-Linear dropout. $\boldsymbol{r}$ — residual stream, $\boldsymbol{u}$ — output of a sublayer (e.g. Attn, FFN, Mamba2), Linear — linear projection.

4.1 Theoretical Setup

We first introduce concepts and assumptions used for our theoretical analysis. We study the pre-activation of a single-neuron linear layer that follows a normalization layer. Let $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ be the input to the normalization layer. We assume that $\boldsymbol{x}$ is a multivariate random variable with components x_i drawn i.i.d. from a distribution with mean $\mu_x := \mathbb{E}[x_i]$ and variance $\sigma_x^2 := \mathrm{Var}[x_i]$. This is a common assumption (He et al., 2015; Gehring et al., 2017; Vankadara et al., 2024) that simplifies the derivations.

For brevity, let $\hat{x}_i := x_i/(\mathrm{RMS}(\boldsymbol{x}) + \epsilon)$ be the output of the RMSNorm transformation with $\boldsymbol{\gamma} = \mathbf{1} \in \mathbb{R}^d$. Let $\boldsymbol{w} \in \mathbb{R}^d$ be the weights of a single output neuron, and let $\boldsymbol{z}$ denote whatever is actually fed into that neuron—which depends on whether and where dropout is applied. For instance, at inference time, no dropout is applied, so $\boldsymbol{z} = \hat{\boldsymbol{x}}$ and the *pre-activation* is simply $y^{\text{test}} = \boldsymbol{w}^\top \boldsymbol{z} = \boldsymbol{w}^\top \hat{\boldsymbol{x}} \in \mathbb{R}$.[3] As $d \to \infty$, by the law of large numbers, $\mathrm{RMS}(\boldsymbol{x}) \to \sqrt{\mu_x^2 + \sigma_x^2}$ (see App. B.1), so the mean $\mu_{\hat{x}} := \mathbb{E}[\hat{x}_i]$ and variance $\sigma_{\hat{x}}^2 := \mathrm{Var}[\hat{x}_i]$ of the normalized coordinate satisfy $\mu_{\hat{x}} \to \frac{\mu_x}{\sqrt{\mu_x^2+\sigma_x^2+\epsilon}}$ and $\sigma_{\hat{x}}^2 \to \frac{\sigma_x^2}{(\sqrt{\mu_x^2+\sigma_x^2+\epsilon})^2}$, and the pre-activation has baseline mean and variance:

$$\mathbb{E}[y^{\text{test}}] \;=\; \mu_{\hat{x}} \sum_{i=1}^{d} w_i, \qquad \mathrm{Var}[y^{\text{test}}] \;=\; \sigma_{\hat{x}}^2 \,\|\boldsymbol{w}\|^2 + C(\hat{\boldsymbol{x}}, \boldsymbol{w}), \tag{8}$$

where $C(\hat{\boldsymbol{x}}, \boldsymbol{w}) = \sum_{i \neq j} w_i w_j \,\mathrm{Cov}[\hat{x}_i, \hat{x}_j]$ and $\mathrm{Cov}[\hat{x}_i, \hat{x}_j]$ is the cross-coordinate covariance induced by the shared normalization denominator (the $\hat{x}_i$ are no longer i.i.d. but exchangeable). Eq. (8) is the inference baseline against which Pre-Norm, Pre-Linear, and VPAD are compared.

4.2 Pre-Norm dropout

Under Pre-Norm dropout—the conventional placement, Eq. 6—the masked sublayer output is added to the residual stream, and the sum is what the next normalization layer reads (Fig. 2, Left). We therefore split the normalization input into $\boldsymbol{x} = \boldsymbol{r} + \boldsymbol{u}$, where $\boldsymbol{u}$ (mean μ_u, variance σ_u^2) is the current sublayer output—the component the dropout mask acts on—and $\boldsymbol{r}$ is the residual accumulated by earlier sublayers. At training time, dropout replaces $\boldsymbol{u}$ by $\frac{\boldsymbol{b}}{1-p} \odot \boldsymbol{u}$ before the residual addition, so the normalized vector and the quantity fed to the linear layer are

$$\tilde{x}_i \;=\; r_i + \frac{b_i}{1-p}\, u_i, \qquad z_i^{\text{PN}} \;=\; \frac{\tilde{x}_i}{\mathrm{RMS}(\tilde{\boldsymbol{x}}) + \epsilon}, \qquad y^{\text{PN}} = \boldsymbol{w}^\top \boldsymbol{z}^{\text{PN}}. \tag{9}$$

Proposition 1 (Pre-Norm dropout)**.** *Assume $\mu_x \neq 0$. As $d \to \infty$,*

$$\mathbb{E}[y^{\text{PN}}] \;\longrightarrow\; c_\epsilon\, \mathbb{E}[y^{\text{test}}], \qquad c_\epsilon \;=\; \left(1 + \tfrac{p}{1-p}\,\kappa\right)^{-1/2} + O(\epsilon), \tag{10}$$

where $\kappa \;=\; \mathbb{E}[u_i^2]/\mathbb{E}[x_i^2]$.

Whenever $\mu_x \neq 0$ and $p > 0$, Pre-Norm dropout attenuates the pre-activation mean, creating a train-test mismatch in the inputs to the linear projection layer. Since the mean shifts, this can result in activation of completely different neurons, disrupting the subsequent layers, which in turn can propagate error to the next tokens by attention.

[3] We omit the bias term as it does not affect our analysis: the training- and inference-time expectations are shifted by the same constant, and the variance is translation-invariant.

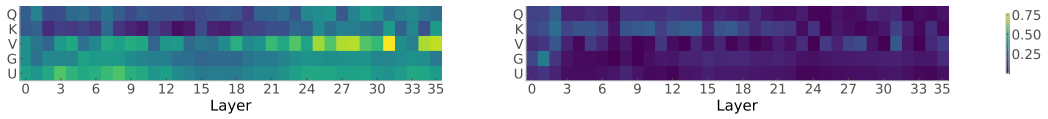

Figure 3: Comparison of the relative error $|\mathrm{Var}[y^{\mathrm{Dropout}}] - \mathrm{Var}[y^{\mathrm{test}}]|/\mathrm{Var}[y^{\mathrm{test}}]$ of the pre-activation variances across layers of SmolLM3 fine-tuned with **Left**: Pre-Linear Dropout; **Right**: VPAD. VPAD substantially reduces the error. Moreover, the error decreases with depth for VPAD but grows for PL-D (first/last 18 layers, averaged over matrices and layers: 0.09/0.06 vs. 0.29/0.37).

4.3 Pre-Linear dropout

The alternative placement applies RMSNorm to the residual stream first, and only then standard inverted dropout, immediately before the linear layer (Fig. 2, Right): $z_i^{\mathrm{PL}} = \frac{b_i}{1-p}\,\hat{x}_i$. Because the mask now acts on the already-normalized representation, the result is independent of how $\boldsymbol{x}$ decomposes into residual and sublayer contributions. The pre-activation moments are

$$\mathbb{E}[y^{\mathrm{PL}}] = \mathbb{E}[y^{\mathrm{test}}], \qquad \mathrm{Var}[y^{\mathrm{PL}}] = \mathrm{Var}[y^{\mathrm{test}}] + \frac{p}{1-p}\,\mathbb{E}[\hat{x}_i^2]\,\|\boldsymbol{w}\|^2. \tag{11}$$

That is, Pre-Linear dropout preserves the mean exactly but inflates the variance, irrespective of the input. This is the familiar behavior of inverted dropout: the mask preserves expectations, but the extra randomness it introduces is not removed by the $1/(1-p)$ rescaling (Hendrycks & Gimpel, 2017; Li et al., 2019); see the proof in App. B.3 (Proposition 2).

4.4 Variance-Preserving Affine Dropout

Since Pre-Linear dropout inflates the variance for any $p > 0$, we hypothesize that, similar to the train-test mismatch of the mean, the variance mismatch can affect the subsequent layers as well. We therefore derive *Variance-Preserving Affine Dropout* (VPAD): a deterministic affine correction of the masked activation so that *both* moments are preserved. With mean $\mu_{\hat{x}}$ and variance $\sigma_{\hat{x}}^2$ of the normalized input $\hat{x}_i$ (see Sec. 4.1), VPAD replaces the masked activation $b_i\hat{x}_i$ by

$$z_i^{\mathrm{VPAD}} = \alpha\, b_i\, \hat{x}_i + \beta, \qquad \alpha = \sigma_{\hat{x}}\big((1-p)\,(\sigma_{\hat{x}}^2 + p\,\mu_{\hat{x}}^2)\big)^{-1/2}, \qquad \beta = \mu_{\hat{x}}\big(1-\alpha(1-p)\big). \tag{12}$$

The coefficients α and β are chosen so $\mathbb{E}[z_i^{\mathrm{VPAD}}] = \mu_{\hat{x}}$ and $\mathrm{Var}[z_i^{\mathrm{VPAD}}] = \sigma_{\hat{x}}^2$. Consequently the pre-activation matches the inference baseline exactly in the mean, and in the variance up to an attenuated covariance term:

$$\mathbb{E}[y^{\mathrm{VPAD}}] = \mathbb{E}[y^{\mathrm{test}}], \qquad \mathrm{Var}[y^{\mathrm{VPAD}}] - \mathrm{Var}[y^{\mathrm{test}}] = \big(\alpha^2(1-p)^2 - 1\big)\,C(\hat{\boldsymbol{x}}, \boldsymbol{w}) \tag{13}$$

(Proposition 3, App. B.4). The masks attenuate each pairwise covariance by the factor $\alpha^2(1-p)^2 = \frac{\sigma_{\hat{x}}^2(1-p)}{\sigma_{\hat{x}}^2 + p\,\mu_{\hat{x}}^2} \in (0,1)$, so VPAD matches the per-coordinate variances exactly and shrinks the covariance term; hence, the residual variance mismatch never exceeds $|C(\hat{\boldsymbol{x}}, \boldsymbol{w})|$. In contrast, Pre-Linear dropout has fixed $\frac{p}{1-p}\|\boldsymbol{w}\|^2$ inflation. Fig. 3 compares the relative errors for two SmolLM3 models fine-tuned with either Pre-Linear dropout or VPAD: VPAD substantially reduces the error, and later layers exhibit smaller errors than the earlier ones.

In practice, we replace $\mu_{\hat{x}}$ and $\sigma_{\hat{x}}^2$ by the empirical estimates $\frac{1}{d}\sum_i \hat{x}_i$ and $\frac{1}{d}\sum_i(\hat{x}_i - \mu_{\hat{x}})^2$, which concentrate around their population values for large d. For $\mu_{\hat{x}} = 0$ and $\sigma_{\hat{x}}^2 = 1$, the coefficients (12) recover the affine correction of Alpha-Dropout (Klambauer et al., 2017) with the dropped value set to zero ($\alpha' = 0$): VPAD is thus its extension to non-centered activations—exactly the situation created by RMSNorm. Another distinction from Alpha-Dropout is that α and β are computed per token while Alpha-Dropout does not depend on the input at all.

5 Experiments

5.1 Experimental Setup

We conduct experiments in two settings: training from scratch and continual pre-training (CPT). In all experiments, we apply weight decay only to the models labeled "WD". Note that we train

the different setups with different amounts of data. The results must therefore be read within each experimental setting and model architecture. We report all experimental details in App. C.

Training from scratch For experiments with transformers, we use a Llama-3-like model but replace RoPE with NAPE (Vasylenko et al., 2026), which has shown its effectiveness at extrapolation; we use no attention scaling or any other intervention related to PE. We train 400M-parameter models with a 2K context length on 40B tokens in five settings: weight decay (WD), no weight decay (no-WD), Pre-Norm Dropout (PN-D), Pre-Linear Dropout (PL-D), and VPAD. For Mamba2 (Dao & Gu, 2024), we follow Buitrago & Gu (2025) and train 85M-parameter models with a 1K context length on 25B tokens. For evaluation, we measure CE loss on 2K-long documents and Recall-Based Accuracy (RBA) on Needle-in-a-Haystack (NIAH-1 from RULER) and overall RULER.

Continual Pre-training To test dropout at scale, we employ continual pre-training of larger models. For transformers, we use the SmolLM3-3B (Bakouch et al., 2025) checkpoint (stage 3) pre-trained with a 4K context. We convert its RoPE layers to sliding-window (SW) attention with a window size of 4K while leaving the NoPE layers untouched. Models are then trained with the same 4K context size on 1B tokens. To test the complementarity of dropout and attention scaling, we train models—SmolLM3-SW-8K—with an 8K context and a 4K sliding window, as well as a variant with the input- and length-dependent attention scaling of Vasylenko et al. (2026), denoted S$\times$. We compare the SmolLM3-SW-8K models with the original SmolLM3-Base, which was trained on an additional 100B tokens after stage 3, and with SmolLM3-Base extended with YaRN RoPE. For Mamba2, we apply CPT to a 1.3B-parameter model for 0.5B tokens. We evaluate all models on NIAH-1, RULER (Hsieh et al., 2024), and a subset of HELMET (Yen et al., 2025) tasks: In-Context Learning (ICL), Retrieval-Augmented Generation (RAG), and Long-document QA (LongQA).

5.2 Regularization Can Mitigate Length-Based Overfitting

Sec. 2 showed that, as training progresses, the extrapolation of transformers trained with or without WD inevitably worsens. Here, we show that other regularization methods mitigate this effect.

Fig. 1 shows that dropout, specifically VPAD, mitigates length overfitting substantially. Training for 80K steps does not bend the CE loss and RBA curves as much as it does for WD or no-WD. Moreover, the CE loss stays below both baselines at all training steps. On RULER, the dropout model already surpasses all baselines at 12K steps and, at 44K, outperforms the second-best model—no-WD—by 60%. At the final training step (Fig. 4), all dropout variants improve the CE loss and allow the models to extrapolate much further on NIAH-1 and RULER, while the WD model collapses almost immediately at $2\times$ the ID length. CPT of SmolLM3 (Table 1) gives a similar picture. Neither the original model nor its SW variant extrapolates even on the simplest NIAH-1 task, and fine-tuning with WD or no-WD does not change this. In contrast, all dropout variants significantly mitigate the problem, allowing for non-zero performance even at $8\times$ the ID length.

Dropout is not the only regularization that improves extrapolation: we observe moderate gains (App. D) for checkpoint averaging and attention dropout, which drops attention weights. This is reminiscent of classical *overfitting*, but with respect to context length: a model fits its training distribution while its OOD performance degrades, and proper regularization mitigates it. Importantly, the effect is not specific to PE, as improvements hold for both NAPE and RoPE-SW models.

5.3 Effect of Dropout Placement

Sec. 4 predicts that the position of the dropout mask relative to normalization is decisive: Pre-Norm dropout attenuates the pre-activation mean (Proposition 1). We hypothesize that the train-test mismatch is detrimental for extrapolation and that placing dropout just before the Linear layer (see Fig. 2, Pre-Linear) is a better approach. Here, we test this hypothesis empirically.

Fig. 4 shows the CE loss and the performance on NIAH-1 and on RULER for models trained from scratch. For CE and NIAH-1, Pre-Norm Dropout lags behind even the no-WD model, which justifies its disappearance from pre-training pipelines; in contrast, Pre-Linear Dropout is significantly better than WD, no-WD, and PN-D. The same ordering holds at the 3B scale under continual pre-training: in Table 1 (full NIAH-1 in Table 12), PN-D extrapolates only partially to $2\times$ the ID length on NIAH-1 and collapses beyond it, whereas PL-D extrapolates perfectly, retaining 100.0 RBA up to

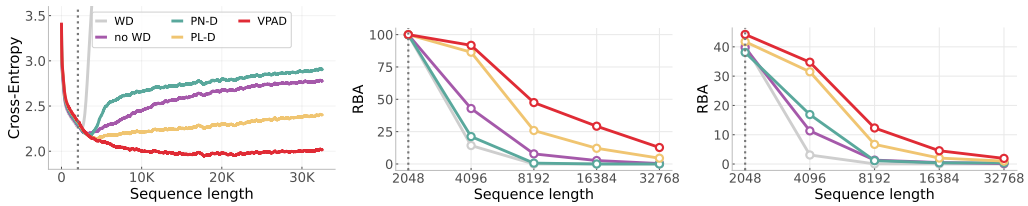

Figure 4: Results for transformers trained from scratch, evaluated on **Left:** ArXiv papers; **Middle:** NIAH-1; **Right:** RULER. The dotted lines represent the pre-trained context length.

64× (256K). The gap on overall RULER is equally consistent: PL-D more than doubles PN-D at 2× extrapolation and remains operational at 4–8×, where PN-D is near zero. In addition, dropout placement also affects sparse transformers (see Table 11 in App. D.4).

5.4 Effect of Variance Preservation (VPAD)

With the placement fixed to Pre-Linear (Sec. 4.3), we ask whether removing the remaining variance mismatch helps further by replacing standard inverted dropout with VPAD. When training from scratch, VPAD outperforms all other methods on all three evaluations of Fig. 4: its CE curve is the lowest and flattest at OOD lengths, and its RBA is the highest on both NIAH-1 and RULER at every context size—at 8K (4× extrapolation), it scores around 50 on NIAH-1 while the second-best model, PL-D, stays under 30. Although VPAD's CE is slightly higher in the 0–2K range, its ID performance does not suffer: at 2K, its RBA is the best among the five settings.

Under continual pre-training, VPAD nearly matches the perfect extrapolation of PL-D on NIAH-1, retaining 100.0 RBA up to 32× and 99.2 at 64× the ID length, and achieves the best RULER results at 2–8× extrapolation. However, although OOD performance improves as the dropout probability increases, ID performance drops (Table 1, CPT: VPAD); this also holds for PL-D and PN-D (Table 8 in App. D.2), with VPAD degrading the least. The effect is milder on standard short-context benchmarks (Table 7 in App. D.1): on-par or slightly better performance for $p = 0.1$ and 0.2, with minimal degradation for $p = 0.3$. In both cases, we presume the model unlearns some behaviors as it adapts to the new training dynamics, particularly for stronger dropout. On HELMET at 2–8× the ID length (Table 2), the SmolLM3-SW and WD baselines collapse to near-zero performance, while both dropout variants extrapolate: PL-D is better at 8K in ICL and LongQA, but VPAD is consistently better at 4–8×. Moreover, VPAD keeps improving on ICL and LongQA as the context grows, suggesting that it is more effective on much longer sequences. This is also evident from Table 3, where its score increases until 32× the pre-training length, whereas PL-D peaks at 4×.

Table 1: SmolLM3 performance on NIAH-1 and on RULER. Grey — the ID length. Bold — best.

	NIAH-1					RULER			
Model	4K	8K	16K	32-128K	256K	4K	8K	16K	32K
Orig.	**100.0**	0.0	0.0	— 0.0 —	0.0	**83.5**	0.1	0.0	0.0
SW	**100.0**	0.0	0.0	— 0.0 —	0.0	83.4	0.5	0.0	0.0
CPT									
SW, WD	**100.0**	0.0	0.0	— 0.0 —	0.0	**83.5**	1.1	0.0	0.0
SW, no-WD	**100.0**	0.0	0.0	— 0.0 —	0.0	**83.5**	1.1	0.0	0.0
SW, PN-D $p = 0.3$	**100.0**	82.2	6.2	— 0.0 —	0.0	76.1	22.5	2.1	0.9
SW, PL-D $p = 0.3$	**100.0**	**100.0**	**100.0**	— **100.0** —	**100.0**	74.1	50.6	40.2	25.5
CPT: VPAD									
SW, VPAD $p = 0.1$	**100.0**	**100.0**	57.2	30.4 → 0.0	0.0	80.4	45.2	5.9	2.0
SW, VPAD $p = 0.2$	**100.0**	**100.0**	**100.0**	96.2 → 86.0	49.4	79.9	**55.0**	38.8	19.4
SW, VPAD $p = 0.3$	**100.0**	**100.0**	**100.0**	— **100.0** —	99.2	77.8	54.3	**44.1**	**27.4**

5.5 Dropout and Attention Scaling Are Complementary

To improve extrapolation, we can apply length-dependent attention scaling, which has been shown to mitigate softmax-related issues (Chiang & Cholak, 2022; Nakanishi, 2025; Peng et al., 2024). In this section, we demonstrate that dropout and attention scaling are complementary. For evaluation, we use the ICL task from HELMET. The prominent distinction of the ICL task is that the model

Table 2: OOD performance of SmolLM3 and Mamba2 on ICL, RAG and LongQA from HELMET. For SmolLM3, $p = 0.3$ and, for Mamba2, $p = 0.2$. Bold — best.

	ICL			RAG			LongQA		
Model	8K	16K	32K	8K	16K	32K	8K	16K	32K
Original Models									
SmolLM3-SW	0.0	0.0	0.0	0.2	0.0	0.0	4.8	4.3	5.7
Mamba2	8.4	4.0	4.0	5.4	2.0	1.6	7.3	4.5	3.2
CPT: SmolLM3									
SW, WD	0.0	0.0	0.0	0.3	0.0	0.0	6.3	6.7	6.7
SW, PL-D	**61.8**	59.8	58.9	39.4	34.3	28.8	**17.7**	15.2	16.8
SW, VPAD	57.8	**61.5**	**61.8**	**40.2**	**35.3**	**31.0**	14.9	**16.3**	**18.6**
CPT: Mamba2									
WD	5.2	1.5	1.6	6.6	2.3	1.5	**12.6**	9.2	7.5
PL-D	**20.5**	**17.2**	**9.8**	**20.1**	**17.1**	11.2	10.8	**11.6**	**12.1**
VPAD	18.4	15.1	9.6	18.5	15.3	**12.1**	9.3	8.4	9.8

Table 3: Performance of SmolLM3-Base and SmolLM3-SW-8K (S×) on ICL. Bud. — pre-training budget following stage 3. Grey — the ID length of the SW models. Red — the ID length of Base and YaRN. $(.)^*$ — YaRN factor 4 and $(.)^{**}$ — YaRN factor 8.

	SmolLM3-Base		CPT: SmolLM3-SW-8K (Ours)			
Len.	Base	YaRN	PL-D	PL-D S×	VPAD	VPAD S×
8K	54.5	48.1	60.5	61.9	58.3	**62.5**
16K	65.6	61.9	61.8	64.4	61.7	**66.1**
32K	73.2	70.2	63.6	66.3	65.9	**71.4**
64K	78.2	76.9	62.3	67.4	66.1	**72.8**
128K	4.6	83.3	62.1	68.8	66.1	**74.1**
256K	0.0	77.2*	62.8	72.0	68.4	**76.6**
512K	0.0	57.4**	63.2	69.2	68.0	**76.2**
Bud.	100B	100B	1.5B	1.5B	1.5B	1.5B

has to learn from many examples (many-shot), and many LLMs benefit from it (Yen et al., 2025). In this setting, we assume an extrapolation improvement when the model's score either increases or remains relatively stable across context lengths. We show our results in Table 3.

We first observe that Base is not capable of extrapolating $2\times$, which usually happens with RoPE models without RoPE extension. Even though YaRN gains the maximum score overall, it is not able to extrapolate far beyond $2\times$ the ID length. In contrast, all SW-VPAD models improve even at 256K ($32\times$ extrapolation), and at 512K their degradation is negligible. VPAD S× significantly surpasses other SW models at all sequence lengths and is superior at 512K overall. Moreover, the gap between VPAD and VPAD S× at 128–512K is significant—about 8 points. PL-D S× performance is less pronounced, but the same pattern holds: the scaling improves scores substantially. This indicates that dropout and attention scaling are remedies to different extrapolation-related issues.

5.6 MAMBA2

We next test whether our findings transfer to Mamba2—an architecture different from transformers.

Synthetic Copying Following Buitrago & Gu (2025), we conduct experiments on the copying task (Jelassi et al., 2024). We report the mean Exact Match Accuracy at a sequence length of 150 (EMA-150) (EMA is used by Jelassi et al. (2024)) and Character-Level Accuracy at 300 (CLA-300) (as in Buitrago & Gu (2025)) in Table 4; standard deviations over five seeds are given in Table 13 (App. D). Table 4 shows that on average the results are consistent with our observation for transformers: WD worsens the length generalization ability and dropout improves it.

Training from scratch Fig. 5 shows the cross-entropy loss of the trained models. As with transformers, the performance of the WD model begins to degrade almost immediately beyond the pre-training length, indicated by the dotted line. The model trained without WD (no-WD) shows a better CE curve, but it still lags behind the dropout methods, which substantially flatten their curves, indicating improved extrapolation. In contrast to transformers, PL-D and VPAD perform similarly in terms of CE. Furthermore, no-WD, PL-D, and VPAD improve performance on NIAH-1 (Table 4), with VPAD substantially outperforming all other baselines at 2–$4\times$ extrapolation.

CPT Table 5 shows results on NIAH-1 and RULER (see Table 15 for more results). As with transformers, extrapolation improves when Mamba2 is fine-tuned with dropout. On NIAH-1, PL-D and VPAD allow Mamba2 to perform perfectly at 16K and achieve non-zero performance at 32K, extending the extrapolation range $4\times$ over the original and non-dropout models. For RULER, all dropout variants improve length generalization substantially starting from the ID length. At $4\times$ the ID length, PL-D and VPAD double the score, and at 8–$16\times$ they maintain non-zero performance while the original and non-dropout models fail completely. PN-D mirrors the placement effect we observed for transformers (Sec. 5.3): it improves performance on RULER at $2\times$, but its gains at longer lengths are much less pronounced. Further evidence for our regularization hypothesis comes from the HELMET results in Table 2, where we show $\geq 4\times$ extrapolation performance. The dropout variants substantially boost the results of ICL and RAG, giving a two- to ten-fold improvement

over the original model. Notably, VPAD interacts differently with Mamba2 than with transformers: VPAD is better on NIAH-1 and at the ID length of RULER, but it either lags behind or performs similarly to PL-D in extrapolation on both RULER and HELMET.

Table 4: Mamba2-45M and Mamba2-85M performance on Copy and NIAH-1, respectively.

	Copy		NIAH-1		
Model	EMA-150	CLA-300	1K	2K	4K
WD	33.1	42.8	72.3	30.5	1.1
no-WD	42.7	45.7	**91.2**	50.6	1.1
PL-D	61.5	50.9	90.3	43.4	10.5
VPAD	**68.9**	**58.0**	88.9	**76.4**	**30.2**

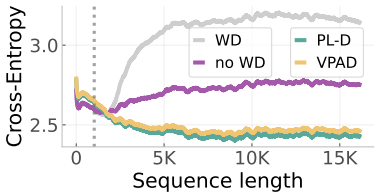

Figure 5: CE loss of Mamba2-85M models on The Pile.

Table 5: Mamba2-1.3B performance on NIAH-1 and on RULER. Bold — best.

	NIAH-1					RULER				
Model	2K	4K	8K	16K	32K	2K	4K	8K	16K	32K
Orig.	**100.0**	**100.0**	**100.0**	0.0	0.0	38.3	14.3	8.5	0.4	0.0
CPT										
WD	**100.0**	**100.0**	**100.0**	0.0	0.0	42.0	16.2	9.6	0.7	0.0
no-WD	**100.0**	**100.0**	**100.0**	0.0	0.0	42.2	16.0	9.6	0.0	0.0
PN-D, $p = 0.2$	**100.0**	**100.0**	**100.0**	11.8	0.0	50.7	22.5	11.5	2.6	0.8
PL-D, $p = 0.2$	**100.0**	**100.0**	**100.0**	**100.0**	22.8	48.9	**28.8**	18.0	**13.7**	5.3
VPAD, $p = 0.2$	**100.0**	**100.0**	**100.0**	**100.0**	**45.0**	**50.9**	26.3	**18.5**	13.3	**5.8**

6 Related Work

Extrapolation failures in transformers are commonly attributed to PE (Press et al., 2022; Sun et al., 2023; Kazemnejad et al., 2023; Chen et al., 2023; Peng et al., 2024; Li et al., 2024) and to softmax attention (Chiang & Cholak, 2022; Veličković et al., 2025; Barbero et al., 2024; Nakanishi, 2025). Buitrago & Gu (2025) show that linear RNNs overfit to length as training progresses. We demonstrate that the same overfitting happens in transformers, is distinct from PE- and softmax- related issues, and can be mitigated by proper regularization. Du et al. (2025) aligns output distributions across sequence lengths with an additional regularization term; we instead study standard regularization. Zhou et al. (2024) observe that a substantial dropout rate (0.2) severely impairs extrapolation on integer addition, so their models use no dropout. We attribute this to the train-test mismatch analyzed here and reach the opposite conclusion: correctly placed dropout substantially improves length generalization. Wu et al. (2021) and Sankararaman et al. (2022) explore placing dropout just before linear projections. Both target in-distribution performance and, importantly, do not analyze the mask's interaction with normalization. In contrast, we show that the placement relative to the normalization layer determines the train-test mismatch of pre-activation statistics, which affects length generalization. Inverted dropout removes the train-test gap of pre-activation means (Srivastava et al., 2014) but not of their variances (Hendrycks & Gimpel, 2017; Wager et al., 2013). Li et al. (2019) identify a related dropout–batch-normalization mismatch in CNNs. We show the analogous mean mismatch in transformers combining dropout with layer normalization, and that extrapolation benefits from fixing the mismatch and from matching the variance. See more related work in App. A.

7 Conclusion

We showed that transformers can progressively lose extrapolation ability during training, suggesting a form of length-based overfitting that proper regularization can substantially mitigate. In that regard, we showed that standard Pre-Norm dropout introduces a train-test mismatch in pre-activation statistics, Pre-Linear dropout removes the mean mismatch, and VPAD additionally preserves the marginal variance. These changes substantially improve extrapolation, including perfect NIAH recall up to $64\times$ the training context in our continual-pretraining setting. Our results therefore suggest that length generalization is affected not only by the architecture and positional encoding but also by the choice of regularization. The right choice of dropout can mitigate length-based overfitting without introducing new train-test distribution shifts.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing the proofs and fix notation issues. We also used generative AI tools for summarizing and analyzing existing literature, searching for information, discover new literature besides the manual investigation, and edit the writing and polishing the paper for readability. For the CPT experiments, we used generative AI tools to write and edit software code. All uses of AI tools were accompanied by a careful review process. In particular, we checked the proofs and made them more accessible, and LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several steps to make our results reproducible. All assumptions of the theoretical claims are stated in Sec. 4.1, and complete proofs are given in App. B. The experimental setup is fully specified in App. C: model architectures and all training hyperparameters (Table 6), the dropout implementation for both transformers and Mamba2, and the evaluation protocol, including the metrics and datasets used for each benchmark. All base checkpoints and training corpora are publicly available. Our code and models will be made publicly available upon acceptance.

REFERENCES

Loubna Ben allal, Anton Lozhkov, Elie Bakouch, Gabriel Martin Blazquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Agustín Piqueres Lajarín, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan Son NGUYEN, Ben Burtenshaw, Clémentine Fourrier, Haojun Zhao, Hugo Larcher, Mathieu Morlon, Cyril Zakka, Colin Raffel, Leandro Von Werra, and Thomas Wolf. SmolLM2: When smol goes big — data-centric training of a fully open small language model. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=3JiCl2A14H.

Chenxin An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, and Lingpeng Kong. Training-free long-context scaling of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1493–1510. PMLR, 2024. URL https://proceedings.mlr.press/v235/an24b.html.

Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. Why does the effective context length of LLMs fall short? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/884baf65392170763b27c914087bde01-Abstract-Conference.html.

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38546–38556, 2022. doi: 10.52202/068431-2793. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fb7451e43f9c1c35b774bcfad7a5714b-Abstract-Conference.html.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/448fc91f669c15d10364ee01d512cc10-Abstract-Conference.html.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. URL https://arxiv.org/abs/1607.06450.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. Hugging Face Blog, 2025. URL https://huggingface.co/blog/smollm3.

Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João G.M. Araújo, Alex Vitvitskyi, Razvan Pascanu, and Petar Veličković. Transformers need glasses! information over-squashing in language tasks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 98111–98142. Curran Associates, Inc., 2024. doi: 10.52202/079017-3114. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b1d35561c4a4a0e0b6012b2af531e149-Abstract-Conference.html.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150.

Assaf Ben-Kish, Itamar Zimerman, Shady Abu-Hussein, Nadav Cohen, Amir Globerson, Lior Wolf, and Raja Giryes. Decimamba: Exploring the length extrapolation potential of mamba. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=iWSl5Zyjjw.

Assaf Ben-Kish, Itamar Zimerman, Muhammad Jehanzeb Mirza, Lior Wolf, James R. Glass, Leonid Karlinsky, and Raja Giryes. Overflow prevention enhances long-context recurrent LLMs. In *Second Conference on Language Modeling*, 2025b. URL https://openreview.net/forum?id=h99hJlU99U.

Ricardo Buitrago and Albert Gu. Understanding and improving length generalization in recurrent models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5856–5874. PMLR, 2025. URL https://proceedings.mlr.press/v267/buitrago25a.html.

Guanzheng Chen, Xin Li, Zaiqiao Meng, Shangsong Liang, and Lidong Bing. CLEX: Continuous length extrapolation for large language models. In *The Twelfth International Conference on Learning Representations*, pp. 14352–14369, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3df38ca67befaed9c03b95ffee07d9f8-Abstract-Conference.html.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023. URL https://arxiv.org/abs/2306.15595.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. LongloRA: Efficient fine-tuning of long-context large language models. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=6PmJoRfdaK.

Ta-Chung Chi, Ting-Han Fan, Peter Ramadge, and Alexander Rudnicky. KERPLE: Kernelized relative positional embedding for length extrapolation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0610. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/37a413841a614b5414b333585e7613b8-Abstract-Conference.html.

Ta-Chung Chi, Ting-Han Fan, Alexander Rudnicky, and Peter Ramadge. Dissecting transformer length extrapolation via the lens of receptive field analysis. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13522–13537, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.756. URL https://aclanthology.org/2023.acl-long.756/.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7654–7664, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.527. URL https://aclanthology.org/2022.acl-long.527/.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019. URL https://arxiv.org/abs/1904.10509.

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

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. LongRoPE: Extending LLM context window beyond 2 million tokens. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11091–11104. PMLR, 2024. URL https://proceedings.mlr.press/v235/ding24i.html.

Tianqi Du, Haotian Huang, Yifei Wang, and Yisen Wang. Long-short alignment for effective long-context modeling in LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14515–14535. PMLR, 2025. URL https://proceedings.mlr.press/v267/du25a.html.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14125–14134. PMLR, 2024. URL https://proceedings.mlr.press/v235/fu24d.html.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. URL https://arxiv.org/abs/2101.00027.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively). In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7376–7399, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.366. URL https://aclanthology.org/2025.acl-long.366/.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1243–1252. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/gehring17a.html.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Contextual position encoding: Learning to count what's important, 2024. URL https://arxiv.org/abs/2405.18719.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. LM-Infinite: Zero-shot extreme length generalization for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3991–4008. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.222. URL https://aclanthology.org/2024.naacl-long.222/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/He_Delving_Deep_into_ICCV_2015_paper.html.

Zhenyu He, Guhao Feng, Shengjie Luo, Kai Yang, Liwei Wang, Jingjing Xu, Zhi Zhang, Hongxia Yang, and Di He. Two stones hit one bird: Bilevel positional encoding for better length extrapolation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17858–17876. PMLR, 2024. URL https://proceedings.mlr.press/v235/he24c.html.

Dan Hendrycks and Kevin Gimpel. Adjusting for dropout variance in batch normalization and weight initialization. In *International Conference on Learning Representations, Workshop Track*, 2017. URL https://openreview.net/forum?id=S1AtgaPug.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Ermo Hua, Che Jiang, Xingtai Lv, Kaiyan Zhang, Youbang Sun, Yuchen Fan, Xuekai Zhu, Biqing Qi, Ning Ding, and Bowen Zhou. Fourier position embedding: Enhancing attention's periodic extension for length generalization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24932–24949. PMLR, 2025. URL https://proceedings.mlr.press/v267/hua25b.html.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence*, pp. 876–885, 2018. URL https://auai.org/uai2018/proceedings/papers/313.pdf.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21502–21521. PMLR, 2024. URL https://proceedings.mlr.press/v235/jelassi24a.html.

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. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. LLM maybe LongLM: SelfExtend LLM context window without tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22099–22114. PMLR, 2024. URL https://proceedings.mlr.press/v235/jin24b.html.

Nikhil Kandpal, Brian Lester, Colin A. Raffel, Sebastian Majstorovic, Stella Biderman, Baber Abbasi, Luca Soldaini, Enrico Shippole, A. Feder Cooper, Aviya Skowron, Shayne Longpre, Lintang Sutawika, Alon Albalak, Zhenlin Xu, Guilherme Penedo, Loubna Ben Allal, Elie Bakouch, John Pressman, Honglu Fan, Dashiell Stander, Guangyu Song, Aaron Gokaslan, John Kirchenbauer, Tom Goldstein, Brian Bartoldson, Bhavya Kailkhura, and Tyler Murray. The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text. In *Advances in Neural Information Processing Systems*, volume 38, pp. 63792–63848, 2025. doi: 10.52202/085713-1918. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/52acc050138d6f40dad6f12f91a4ce22-Abstract-Datasets_and_Benchmarks_Track.html.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1082. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4e85362c02172c0c6567ce593122d31c-Abstract-Conference.html.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 971–980, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/5d44ee6f2c3f71b73125876103c8f6c4-Abstract.html.

Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. Functional interpolation for relative positions improves long context transformers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2f55a8b7b1c2c6312eb86557bb9a2bd5-Abstract-Conference.html.

Xiang Li, Shuo Chen, Xiaolin Hu, and Jian Yang. Understanding the disharmony between dropout and batch normalization by variance shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2682–2690, 2019. doi: 10.1109/CVPR.2019.00279. URL https://openaccess.thecvf.com/content_

CVPR_2019/html/Li_Understanding_the_Disharmony_Between_Dropout_and_Batch_Normalization_by_Variance_CVPR_2019_paper.html.

Wanchao Liang, Tianyu Liu, Less Wright, Will Constable, Andrew Gu, Chien-Chin Huang, Iris Zhang, Wei Feng, Howard Huang, Junjie Wang, Sanket Purandare, Gokul Nadathur, and Stratos Idreos. Torchtitan: One-stop pytorch native solution for production ready LLM pretraining. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SFN6Wm7YBI.

Xiaoran Liu, Hang Yan, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of RoPE-based extrapolation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d33b741600f100b31256c70a46f66ec9-Abstract-Conference.html.

Peng Lu, Jerry Huang, Qiuhao Zeng, Xinyu Wang, Boxing Chen, Philippe Langlais, and Yufei Cui. Mamba modulation: On the length generalization of Mamba models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0653. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1bfc9f74afa91b9b8add5a97a97001a1-Abstract-Conference.html.

Xin Ma, Yang Liu, Jingjing Liu, and Xiaoxu Ma. Mesa-extrapolation: A weave position encoding method for enhanced extrapolation in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2587. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9446c291a8744a125a0bda5b18f4d5a1-Abstract-Conference.html.

Eran Malach, Omid Saremi, Sinead Williamson, Arwen Bradley, Aryo Lotfi, Emmanuel Abbe, Joshua Susskind, and Etai Littwin. To infinity and beyond: Tool-use unlocks length generalization in state space models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4a6884ee51a981046ecf5b449f776dff-Abstract-Conference.html.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=2M9CUnYnBA.

Ken M. Nakanishi. Scalable-softmax is superior for attention. *arXiv preprint arXiv:2501.19399*, 2025. URL https://arxiv.org/abs/2501.19399.

NVIDIA, Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduchintala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, Aleksander Ficek, Alex Kondratenko, Alex Shaposhnikov, Alexander Bukharin, Ali Taghibakhshi, Amelia Barton, Ameya Sunil Mahabaleshwarkar, Amy Shen, Andrew Tao, Ann Guan, Anna Shors, Anubhav Mandarwal, Arham Mehta, Arun Venkatesan, Ashton Sharabiani, Ashwath Aithal, Ashwin Poojary, Ayush Dattagupta, Balaram Buddharaju, Banghua Zhu, Barnaby Simkin, Bilal Kartal, Bita Darvish Rouhani, Bobby Chen, Boris Ginsburg, Brandon Norick, Brian Yu, Bryan Catanzaro, Charles Wang, Charlie Truong, Chetan Mungekar, Chintan Patel, Chris Alexiuk, Christian Munley, Christopher Parisien, Dan Su, Daniel Afrimi, Daniel Korzekwa, Daniel Rohrer, Daria Gitman, David Mosallanezhad, Deepak Narayanan, Dima Rekesh, Dina Yared, Dmytro Pykhtar, Dong Ahn, Duncan Riach, Eileen Long, Elliott Ning, Eric Chung, Erick Galinkin, Evelina Bakhturina, Gargi Prasad, Gerald Shen, Haifeng Qian, Haim Elisha, Harsh Sharma, Hayley Ross, Helen Ngo, Herman Sahota, Hexin Wang, Hoo Chang Shin, Hua Huang, Iain Cunningham, Igor Gitman, Ivan Moshkov, Jaehun Jung, Jan Kautz, Jane Polak Scowcroft, Jared Casper, Jian Zhang, Jiaqi Zeng, Jimmy Zhang, Jinze Xue, Jocelyn Huang, Joey Conway, John Kamalu, Jonathan Cohen, Joseph Jennings, Julien Veron Vialard, Junkeun Yi, Jupinder Parmar, Kari Briski, Katherine Cheung, Katherine Luna, Keith Wyss, Keshav Santhanam, Kezhi Kong, Krzysztof Pawelec, Kumar Anik, Kunlun Li, Kushan Ahmadian, Lawrence McAfee, Laya Sleiman, Leon Derczynski, Luis Vega, Maer Rodrigues de Melo, Makesh Narsimhan Sreedhar, Marcin Chochowski, Mark Cai, Markus Kliegl, Marta Stepniewska-Dziubinska, Matvei Novikov, Mehrzad Samadi, Meredith Price, Meriem Boubdir, Michael Boone, Michael Evans, Michal Bien, Michal Zawalski, Miguel Martinez, Mike Chrzanowski, Mohammad Shoeybi, Mostofa Patwary, Namit

Dhameja, Nave Assaf, Negar Habibi, Nidhi Bhatia, Nikki Pope, Nima Tajbakhsh, Nirmal Kumar Juluru, Oleg Rybakov, Oleksii Hrinchuk, Oleksii Kuchaiev, Oluwatobi Olabiyi, Pablo Ribalta, Padmavathy Subramanian, Parth Chadha, Pavlo Molchanov, Peter Dykas, Peter Jin, Piotr Bialecki, Piotr Januszewski, Pradeep Thalasta, Prashant Gaikwad, Prasoon Varshney, Pritam Gundecha, Przemek Tredak, Rabeeh Karimi Mahabadi, Rajen Patel, Ran El-Yaniv, Ranjit Rajan, Ria Cheruvu, Rima Shahbazyan, Ritika Borkar, Ritu Gala, Roger Waleffe, Ruoxi Zhang, Russell J. Hewett, Ryan Prenger, Sahil Jain, Samuel Kriman, Sanjeev Satheesh, Saori Kaji, Sarah Yurick, Saurav Muralidharan, Sean Narenthiran, Seonmyeong Bak, Sepehr Sameni, Seungju Han, Shanmugam Ramasamy, Shaona Ghosh, Sharath Turuvekere Sreenivas, Shelby Thomas, Shizhe Diao, Shreya Gopal, Shrimai Prabhumoye, Shubham Toshniwal, Shuoyang Ding, Siddharth Singh, Siddhartha Jain, Somshubra Majumdar, Soumye Singhal, Stefania Alborghetti, Syeda Nahida Akter, Terry Kong, Tim Moon, Tomasz Hliwiak, Tomer Asida, Tony Wang, Tugrul Konuk, Twinkle Vashishth, Tyler Poon, Udi Karpas, Vahid Noroozi, Venkat Srinivasan, Vijay Korthikanti, Vikram Fugro, Vineeth Kalluru, Vitaly Kurin, Vitaly Lavrukhin, Wasi Uddin Ahmad, Wei Du, Wonmin Byeon, Ximing Lu, Xin Dong, Yashaswi Karnati, Yejin Choi, Yian Zhang, Ying Lin, Yonggan Fu, Yoshi Suhara, Zhen Dong, Zhiyu Li, Zhongbo Zhu, and Zijia Chen. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model, 2025. URL https://arxiv.org/abs/2508.14444.

NVIDIA, Aaron Blakeman, Aaron Thomas, Aastha Jhunjhunwala, Abhibha Gupta, Abhinav Khattar, Adam Rajfer, Adi Renduchintala, Adil Asif, Aditya Vavre, Adriana Flores Miranda, Ahmad Bilal, Aileen Zaman, Ajay Hotchandani, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, Alex Gronskiy, Alex Kondratenko, Alex Steiner, Alex Ye, Alexander Bukharin, Alexandre Milesi, Ali Taghibakhshi, Alice Gatti, Alisa Liu, Alok Kumar, Amar Phanishayee, Ameya Sunil Mahabaleshwarkar, Amir Klein, Amit Zuker, Amnon Geifman, Anahita Bhiwandiwalla, Ananth Subramaniam, Andrea Santilli, Andrew Fulks, Andrew McHarg, Andrew Tao, Andrii Skliar, Anjulie Agrusa, Ankur Srivastava, Ankur Verma, Anna Shors, Anna Warno, Antoni-Joan Solergibert I Llaquet, Arham Mehta, Arkadiusz Nowaczynski, Arti Jain, Ashwath Aithal, Ashwin Poojary, Asif Ahamed, Asit Mishra, Asma Kuriparambil Thekkumpate, Atefeh Sohrabizadeh, Avinash Kaur, Avinash Vem, Ayush Dattagupta, Barath Subramaniam Anandan, Bardiya Sadeghi, Ben Lanir, Benedikt Schifferer, Besmira Nushi, Bilal Kartal, Bill Thiede, Bita Darvish Rouhani, Bo Deng, Bob Schatz, Boris Ginsburg, Boxin Wang, Brad Nemire, Brandon Norick, Brian Dang, Brian Westphal, Brian Yu, Brucek Khailany, Bryan Catanzaro, Carlo del Mundo, Caryln Aarish, Chankyu Lee, Chantal Hwang, Charbel Sakr, Charles Wang, Charlie Truong, Chen Cui, Cheng Cheng, Cheng-Ping Hsieh, Chenghao Zhang, Chenhui Deng, Chintan Patel, Chris Alexiuk, Christian Cosgrove, Christian Munley, Christine Harvey, Christopher Parisien, Chunyang Shen, Coco Li, Collin Neale, Cynthia Gao, Cyril Meurillon, Dan Gil, Dan Su, Dan Zhao, Dane Corneil, Daniel Afrimi, Daniel Egert, Daniel Korzekwa, Daniel Lo, Daniel Machlab, Daniel Serebrenik, Daniil Sorokin, Daria Gitman, Daria Levy, Darko Stosic, David Mosallanezhad, David Yu, Davit Karamyan, Deena Donia, Deep Debroy, Deepak Narayanan, Devin O'Kelly, Dheeraj Peri, Dhruv Nathawani, Di, Wu, Dima Rekesh, Divyanshu Kakwani, Donald Plummer, Dong Anh, Dongfeng Yu, Dongfu Jiang, Donnie Kim, Dorrin Poorkay, Duncan Riach, Dusan Stosic, Dustin VanStee, Eavan Meng, Edgar Minasyan, Edward Lin, Eileen Margaret Peters Long, Elad Sarafin, Elad Segal, Elena Lantz, Ellie Evans, Elliott Ning, Eric Chung, Eric Harper, Eric Pham-Hung, Eric Tramel, Eric Yang, Erick Galinkin, Erik Pounds, Erika Goncalves Goncalves, Evan Briones, Evan Wu, Evelina Bakhturina, Evgeny Tsykunov, Ewa Dobrowolska, Faisal Ladhak, Farzan Memarian, Fay Wang, Fei Jia, Felipe Soares, Felipe Vieira Frujeri, Feng Chen, Fengguang Lin, Ferenc Galko, Frank Sun, Frankie Siino, Frida Hou, Gal Hubara Agam, Gal Kaplun, Gantavya Bhatt, Gargi Prasad, Garvit Kulshreshtha, George Armstrong, Gerald Shen, Giulio Borghesi, Gordana Neskovic, Gorkem Batmaz, Grace Lam, Greg Mason, Greg Pauloski, Grigor Nalbandyan, Grzegorz Chlebus, Grzegorz Karch, Guan-Ting Liu, Guoming Zhang, Guyue Huang, Haggai Maron, Haifeng Qian, Haim Elisha, Haoxing Ren, Haran Kumar Shiv Kumar, Haribhau Hud, Harris Nover, Harrison Saturley Hall, Hayate Iso, Helen Ngo, Herbert Hum, Herman Sahota, Hexin Wang, Himanshu Soni, Hovhannes Tamoyan, Hua Li, Huanhuan Chen, Hui Li, Hui Wang, Huy Nguyen, Ian Chiles, Ido Galil, Ido Shahaf, Igor Gitman, Igor Shovkun, Ilya Loshchilov, Ingo Guehring, Itamar Schen, Itay Levy, Itay Neeman, Ivan Moshkov, Izik Golan, Izzy Putterman, Jaemin Choi, Jakub Slowikowski, Jan Kautz, Jane Polak Scowcroft, Jared Casper, Jatin Mitra, Jeffrey Glick, Jenny Chen, Jesse Oliver, Jiacheng Xu, Jiafan Zhu, Jialin Song, Jian Zhang, Jiantao Jiao, Jiaqi Zeng, Jie Lou, Jim King, Jimmy Zhang, Jingquan Wang, Jinhang Choi, Jinju Chu,

Joey Conway, Joey Guman, Johan Jatko, Johannes Rausch, John Kamalu, John Roberts, Johnny Greco, Johnny Mensel, Jonah Alben, Jonas Yang, Jonathan Cohen, Jonathan Raiman, Joseph Jennings, Joshua Mabry, Joshua Pierce, Joyjit Daw, Julien Veron Vialard, Junkeun Yi, Jupinder Parmar, Kajal Jain, Kan Zhu, Kari Briski, Katherine Cheung, Katherine Luna, Keith Willowhawk, Keith Wyss, Keshav Santhanam, Kevin Shih, Kezhi Kong, Khanh Nguyen, Khushi Bhardwaj, Kirthi Shankar Sivamani, Konstantinos Krommydas, Krishna C. Puvvada, Krzysztof Pawelec, Kumar Anik, Kyle Keprios, Kylie Day, Lawrence McAfee, Leo Du, Leon Derczynski, Li Ding, Linda Liu, Lingjie Wu, Lior Kadoch, Lizzie Wei, Luis Vega, Luke Robison, Lun Su, Maarten Van Segbroeck, Maciej Jakub Mikulski, Maer Rodrigues de Melo, Magda Sypula, Mahan Fathi, Makesh Narsimhan Sreedhar, Makesh Tarun Chandran, Manoj Kilaru, Maor Ashkenazi, Marc Cuevas, Marc Romeijn, Marcin Chochowski, Mark Cai, Mark Mozolewski, Markus Kliegl, Marta Stepniewska-Dziubinska, Martyna Patelka, Mattei Machczynski, Matvei Novikov, Mauricio Ferrato, Maximilian Golub, Mehrzad Samadi, Melissa Corpuz, Mengru Wang, Mengxi Wu, Meredith Price, Meriem Boubdir, Micah Schaffer, Michael Andersch, Michael Boone, Michael Gschwind, Michael Lightstone, Michael Loh, Michal Bien, Michal Zawalski, Michelle Gill, Miguel Martinez, Mikail Khona, Mike Chrzanowski, Mike Houston, Mingyuan Ma, Minseok Lee, Mohamed Fawzy, Mohammad Dabbah, Mohammad Shoeybi, Mostofa Patwary, Nabin Mulepati, Najeeb Nabwani, Namit Dhameja, Narimane Hennouni, Natalie Hereth, Nathaniel Pinckney, Nave Algarici, Nave Assaf, Netanel Haber, Nicholas Knight, Nick Reamaroon, Nickson Quak, Nidhi Bhatia, Nikhil Desai, Nikolai Ludwig, Nima Tajbakhsh, Ning Xu, Nir Ailon, Nirmal Juluru, Nitin Nitin, Ofri Masad, Oleg Rybakov, Oleksii Hrinchuk, Oleksii Kuchaiev, Olivia Viessmann, Olivier Delalleau, Oluwatobi Olabiyi, Omer Ullman Argov, Omri Puny, Oren Tropp, Pablo Ribalta, Pallab Bhattacharya, Panos Lampropoulos, Parth Mannan, Pasha Shamis, Patrick Legresley, Paul Gibbons, Pavlo Molchanov, Pawel Morkisz, Peter Dykas, Peter Jin, Pierre-Yves Aquilanti, Pinky Xu, Piotr Januszewski, Piotr Laskiewicz, Pooya Jannaty, Prakash Gurumurthy, Pranav Prashant Thombre, Prasoon Varshney, Pritam Gundecha, Przemek Tredak, Puhui Meng, Qiyu Wan, Rabeeh Karimi Mahabadi, Rachel Oberman, Rachit Garg, Radha Sri-Tharan, Rahul Kandu, Rakshit Sanadhya, Ran El-Yaniv, Ran Zilberstein, Rasoul Shafipour, Ray Macalisang, Rayen Tian, Reka Kovacs, Renjie Pi, Rick Izzo, Rima Shahbazyan, Rishabh Garg, Rishi Puri, Rita Fernandes Neves, Ritchie Zhao, Ritika Borkar, Ritu Gala, Riyad Islam, Robert Clark, Robert Hesse, Robert Kirby, Roger Waleffe, Rohit Watve, Roi Koren, Ron Banner, Ruoxi Zhang, Russell J. Hewett, Ryan Prenger, Ryan Stewart, Ryota Egashira, Sadegh Mahdavi, Saee Paliwal, Sagar Singh, Sahil Modi, Salika Dave, Samantha Shinagawa, Samuel Kriman, Sandip Bhaskar, Sangkug Lym, Sanjay Kariyappa, Sanjeev Satheesh, Saran Vikas Murari, Satish Pasumarthi, Saurabh Mishra, Saurav Muralidharan, Scott Hara, Sean Narentharen, Selvaraj Anandaraj, Seonjin Na, Seonmeyong Bak, Seonmyeong Bak, Sepehr Sameni, Seph Mard, Serge Panev, Seth Henneman, Seth Poulos, Shahar Mor, Shantanu Acharya, Shaona Ghosh, Sharath Turuvekere Sreenivas, Sharon Mendelson, Shaun Kotek, Shawn Wang, Shay Aharon, Shaya Gharghabi, Sheng-Chieh Lin, Shi Chen, Shiqing Fan, Shirish Baskaran, Shreya Gopa, Shrimai Prabhumoye, Shubham Pachori, Shubham Toshniwal, Shuoyang Ding, Shwetha Krishnamurthy, Siddharth Singh, Simeng Sun, Sirshak Das, Sivakumar Arayandi Thottakara, Smita Ithape, Somshubra Majumdar, Soumye Singhal, Sri Harsha Singudasu, Sridhar Bhuvanapalli, Srimukh Veccham, Stas Sergienko, Stefania Alborghetti, Stephen Ge, Su Rong, Sugam Dipak Devare, Sukrit Rao, Sumeet Kumar Barua, Sungsoo Ha, Sunny Gai, Suriya Gunasekar, Suseella Panguluri, Suyog Gupta, Sviataslau Hinzburh, Sweta Priyadarshi, Syeda Nahida Akter, Talor Abramovich, Tan Bui, Tanay Varshney, Tatevik Ter-Hovhannisyan, Teodor-Dumitru Ene, Terry Kong, Thanh Do, Tianhe Zhang, Tiffany Moore, Tijmen Blankevoort, Tim Moon, Tiyasa Mitra, Tom Balough, Tomasz Grzegorzek, Tomasz Hliwiak, Tomer Asida, Tomer Bar Natan, Tomer Keren, Tomer Ronen, Tony Salim, Tony Wang, Traian Rebedea, Tugrul Konuk, Twinkle Vashishth, Udi Karpas, Ushnish De, Vahid Noorozi, Venkat Srinivasan, Venmugil Elango, Vibhor Agrawal, Victor Cui, Vijay Korthikanti, Vikas Mehta, Vinay Rao, Virginia Wu, Vitaly Kurin, Vitaly Lavrukhin, Vladimir Anisimov, Vu Pham, Wanli Jiang, Wasi Uddin Ahmad, Wataru Ishihara, Wei Du, Wei Ping, Weiheng Chai, Wenliang Dai, Wesley Helmholz, Will Jennings, Will Zhu, Wojciech Prazuch, Xiaowei Ren, Xiwen Yu, Yan Breek, Yang Chen, Yang Yu, Yangyi Chen, Yaniv Galron, Yashaswi Karnati, Yejin Choi, Yev Meyer, Yi-Fu Wu, Yian Zhang, Ying Lin, Yonatan Geifman, Yonggan Fu, Youngeun Kwon, Yu Yao, Yugi Guvvla, Yuki Huang, Yunsheng Liu, Zach Moshe, Zachary Newell, Zhilin Wang, Zhiyu Li, Zhongbo Zhu, Zhuolin Yang, Zihan Liu, Zijie Yan, and Zsolt-Alon Wertheimer. Nemotron 3 ultra: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning, 2026. URL https://arxiv.org/abs/2606.15007.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2026. URL https://arxiv.org/abs/2512.13961.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wHBfxhZu1u.

Hugo Pitorro, Pavlo Vasylenko, Marcos Treviso, and André F. T. Martins. How effective are state space models for machine translation? In *Proceedings of the Ninth Conference on Machine Translation*, pp. 1107–1124. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.wmt-1.111. URL https://aclanthology.org/2024.wmt-1.111/.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=R8sQPpGCv0.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://www.jmlr.org/papers/v21/20-074.html.

Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. Randomized positional encodings boost length generalization of transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1889–1903. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-short.161. URL https://aclanthology.org/2023.acl-short.161/.

Karthik Abinav Sankararaman, Sinong Wang, and Han Fang. Bayesformer: Transformer with uncertainty estimation, 2022. URL https://arxiv.org/abs/2206.00826.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2074. URL https://aclanthology.org/N18-2074/.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019. URL https://arxiv.org/abs/1909.08053.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL https://www.jmlr.org/papers/v15/srivastava14a.html.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14590–14604. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.816. URL https://aclanthology.org/2023.acl-long.816/.

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. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

Leena Chennuru Vankadara, Jin Xu, Moritz Haas, and Volkan Cevher. On feature learning in structured state space models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2736. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9c7eeda2dc98e61baa9a5884afd231bc-Abstract-Conference.html.

Dusan Varis and Ondřej Bojar. Sequence length is a domain: Length-based overfitting in transformer models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8246–8257, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.650. URL https://aclanthology.org/2021.emnlp-main.650/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Pavlo Vasylenko, Hugo Pitorro, André F. T. Martins, and Marcos V. Treviso. Long-context generalization with sparse attention. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/19c9708f31ec44b5b1cbd67f91d05d95-Abstract-Conference.html.

Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. Softmax is not enough (for sharp size generalisation). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 61190–61211. PMLR, 2025. URL https://proceedings.mlr.press/v267/velickovic25a.html.

Stefan Wager, Sida Wang, and Percy Liang. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems*, volume 26, pp. 351–359, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/38db3aed920cf82ab059bfccbd02be6a-Abstract.html.

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

Suyuchen Wang, Ivan Kobyzev, Peng Lu, Mehdi Rezagholizadeh, and Bang Liu. Resonance RoPE: Improving context length generalization of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 586–598. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.32. URL https://aclanthology.org/2024.findings-acl.32/.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Zhen Wu, Lijun Wu, Qi Meng, Yingce Xia, Shufang Xie, Tao Qin, Xinyu Dai, and Tie-Yan Liu. UniDrop: A simple yet effective technique to improve transformer without extra cost. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3865–3878, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.302. URL https://aclanthology.org/2021.naacl-main.302/.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5e5fd18f863cbe6d8ae392a93fd271c9-Abstract-Conference.html.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4643–4663. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.260. URL https://aclanthology.org/2024.naacl-long.260/.

Mingyu Xu, Xin Men, Bingning Wang, Qingyu Zhang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Base of RoPE bounds context length. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2773. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9f12dd32d552f3ad9eaa0e9dfec291be-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a. URL https://arxiv.org/abs/2505.09388.

Songlin Yang, Yikang Shen, Kaiyue Wen, Shawn Tan, Mayank Mishra, Liliang Ren, Rameswar Panda, and Yoon Kim. PaTH attention: Position encoding via accumulating Householder transformations. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-2080. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/59c27bf8d56d3d50c7aeaf7535dee975-Abstract-Conference.html.

Zhifan Ye, Kejing Xia, Yonggan Fu, Xin Dong, Jihoon Hong, Xiangchi Yuan, Shizhe Diao, Jan Kautz, Pavlo Molchanov, and Yingyan Celine Lin. LongMamba: Enhancing Mamba's long-context capabilities via training-free receptive field enlargement. In *The Thirteenth International Conference on Learning Representations*, 2025.

URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ab5d50d269e52f8eed497062311ff173-Abstract-Conference.html.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context models effectively and thoroughly. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f5332c8273d02729730a9c24dec2135e-Abstract-Conference.html.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12360–12371, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html.

Chuanyang Zheng, Yihang Gao, Han Shi, Minbin Huang, Jingyao Li, Jing Xiong, Xiaozhe Ren, Michael Ng, Xin Jiang, Zhenguo Li, and Yu Li. DAPE: Data-adaptive positional encoding for length extrapolation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0838. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f050fa9f0d898e3f265d515f50ae8f9-Abstract-Conference.html.

Yongchao Zhou, Uri Alon, Xinyun Chen, Xuezhi Wang, Rishabh Agarwal, and Denny Zhou. Transformers can achieve length generalization but not robustly. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models (ME-FoMo)*, 2024. URL https://openreview.net/forum?id=DWkWIh3vFJ.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. PoSE: Efficient context window extension of LLMs via positional skip-wise training. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/524ef58c2bd075775861234266e5e020-Abstract-Conference.html.

A Extended Related Work

Length generalization in transformers. Varis & Bojar (2021) identify sequence length as a domain to which transformers overfit, and Anil et al. (2022) show that even large models fail to extrapolate. Extrapolation failures are commonly attributed to the positional encoding: relative and bias-based encodings improve extrapolation—relative position representations (Shaw et al., 2018; Raffel et al., 2020), ALiBi (Press et al., 2022), KERPLE (Chi et al., 2022), Sandwich (Chi et al., 2023), and xPos (Sun et al., 2023)—while later encodings learn, adapt, or redesign the positional map (Li et al., 2024; Zheng et al., 2024; Golovneva et al., 2024; He et al., 2024; Hua et al., 2025; Wang et al., 2024; Yang et al., 2025b). Kazemnejad et al. (2023) show that decoder-only models can extrapolate better with no positional encoding at all, and Ruoss et al. (2023) randomize training positions to cover longer test lengths. The second commonly named cause is softmax attention, which disperses as the input grows: Chiang & Cholak (2022) show that this dilution breaks extrapolation and counteract it with a log-length scaling of the logits, Veličković et al. (2025) argue that dispersion is fundamental to softmax, and Nakanishi (2025) propose Scalable-Softmax as a replacement. Barbero et al. (2024) analyze over-squashing and representational collapse on long sequences. Our experiments with models trained from scratch use NAPE (Vasylenko et al., 2026), a NoPE + ALiBi strategy, and the CPT ones use RoPE inside sliding-window attention layers. Moreover, we employ input- and length-dependent attention scaling. We show that regularization provides further gains on top of a proper PE strategy (Sec. 5.2) and attention scaling (Sec. 5.5), indicating that PE and attention are not the whole story.

Extending the context window of pre-trained models. Many works adapt pre-trained models to longer contexts: position interpolation (Chen et al., 2023), frequency-aware RoPE rescaling (Peng et al., 2024), scaling laws and theoretical bounds for RoPE extrapolation (Liu et al., 2024; Xu et al., 2024), LongRoPE (Ding et al., 2024), CLEX (Chen et al., 2024a), PoSE (Zhu et al., 2024), and long-context training recipes (Xiong et al., 2024; Fu et al., 2024; Gao et al., 2025; Chen et al., 2024b). Training-free methods modify positional information or attention computation at inference (Jin et al., 2024; An et al., 2024; Ma et al., 2024; An et al., 2025) or use sliding-window attention with retained initial tokens (Xiao et al., 2024; Han et al., 2024), following earlier sparse/local-attention architectures (Child et al., 2019; Beltagy et al., 2020; Jiang et al., 2023). Our CPT experiments adopt sliding-window attention in the RoPE layers and rely on regularization rather than positional remapping. Importantly, we do not fine-tune on longer sequences.

Regularization for length generalization. Du et al. (2025) study length generalization from the perspective of model output distributions. They introduce Long-Short Misalignment, which measures the discrepancy between output distributions across sequences of different lengths, and improve length generalization by explicitly regularizing this discrepancy during training. Zhou et al. (2024) investigate length generalization on integer addition and show that it is highly sensitive to training conditions. In particular, they observe that a dropout rate of 0.2 severely impairs length generalization, while 0.1 provides no significant improvement, and consequently train their models without dropout. In contrast, we study how standard regularization choices themselves affect length-based overfitting in language models. We show that dropout can substantially improve length generalization when its placement is reconsidered, and identify the train-test mismatch induced by conventional dropout placement as an important factor. Unlike Du et al. (2025), our approach does not explicitly enforce consistency across different sequence lengths and does not require two forward passes.

Length generalization in recurrent and state-space models. Fixed-state models face their own extrapolation limits. Buitrago & Gu (2025) attribute failures of linear RNNs to overfitting to the recurrent states explored during training, Ye et al. (2025) to global channels failing to adaptively extend their receptive fields beyond the training length, and Jelassi et al. (2024) show that copying is harder without attention. Empirically, such models lag transformers on tasks requiring in-context recall (Waleffe et al., 2024; Arora et al., 2024). Proposed remedies include token decimation (Ben-Kish et al., 2025a), receptive-field enlargement (Ye et al., 2025), spectrum scaling of the state-transition dynamics (Lu et al., 2025), chunked inference that prevents state overflow (Ben-Kish et al., 2025b), and tool use (Malach et al., 2026). Our Mamba2 experiments show that proper regularization can also improve extrapolation.

Relation to Alpha-Dropout. The affine correction underlying VPAD coincides with that of Alpha-Dropout (Klambauer et al., 2017), which likewise restores both the mean and the variance of a masked activation through a deterministic scale and shift. On one hand, VPAD is the special case in which dropped units are set to zero rather than to the SELU saturation value α' (i.e. $\alpha' = 0$). On the other hand, unlike Alpha-Dropout, VPAD makes no zero-mean and unit-variance assumption on the activation. VPAD is the general, non-centered case that arises under RMSNorm.

B PROOFS

Throughout this appendix we use the notation of Sec. 4.1: $\boldsymbol{x} \in \mathbb{R}^d$ has i.i.d. entries with mean μ_x and variance σ_x^2; the masks $b_i \sim \mathrm{Bernoulli}(1-p)$ are independent of $\boldsymbol{x}$, where $p \in [0, 1)$ is the drop probability; and the normalized coordinate $\hat{x}_i = x_i/(\mathrm{RMS}(\boldsymbol{x}) + \epsilon)$. Expectations are taken over all sources of randomness—the input activations and the dropout masks, which are independent of each other.

B.1 MOMENTS OF THE NORMALIZED COORDINATE AND INFERENCE BASELINE

We work in the large-d regime, where the relevant sample averages are concentrated around their expectations. Write $y^{\mathrm{test}} = \sum_i w_i \hat{x}_i$. By the strong law of large numbers, $\mathrm{RMS}(\boldsymbol{x})^2 = \frac{1}{d}\sum_j x_j^2 \to \mu_x^2 + \sigma_x^2$ almost surely. Since $r \mapsto 1/(\sqrt{r} + \epsilon)$ is continuous on $[0, \infty)$, the continuous mapping theorem gives $\hat{x}_i = x_i/(\mathrm{RMS}(\boldsymbol{x}) + \epsilon) \to x_i/(\sqrt{\mu_x^2 + \sigma_x^2} + \epsilon)$ almost surely. Next, since $|\hat{x}_i| \le |x_i|/\epsilon$ and $\hat{x}_i^2 \le x_i^2/\epsilon^2$ are integrable envelopes that do not depend on d (as $\mathrm{RMS}(\boldsymbol{x}) + \epsilon \ge \epsilon$), dominated convergence ($\lim_{d\to\infty} \mathbb{E}[\hat{x}_i] = \mathbb{E}[\lim_{d\to\infty} \hat{x}_i]$, and likewise for $\hat{x}_i^2$) gives

$$\mu_{\hat{x}} \to \frac{\mu_x}{\sqrt{\mu_x^2 + \sigma_x^2} + \epsilon}, \qquad \mathbb{E}[\hat{x}_i^2] \to \frac{\mu_x^2 + \sigma_x^2}{(\sqrt{\mu_x^2 + \sigma_x^2} + \epsilon)^2}, \qquad \sigma_{\hat{x}}^2 = \mathbb{E}[\hat{x}_i^2] - \mu_{\hat{x}}^2 \to \frac{\sigma_x^2}{(\sqrt{\mu_x^2 + \sigma_x^2} + \epsilon)^2}. \tag{14}$$

This proves the per-coordinate limits, and linearity of expectation gives the mean claim of Eq. (8): $\mathbb{E}[y^{\mathrm{test}}] = \sum_i w_i \mathbb{E}[\hat{x}_i] = \mu_{\hat{x}} \sum_i w_i$. By exchangeability, every coordinate has the same variance $\sigma_{\hat{x}}^2$, hence:

$$\mathrm{Var}[y^{\mathrm{test}}] \;=\; \sum_i w_i^2 \,\mathrm{Var}[\hat{x}_i] \;+\; \sum_{i \ne j} w_i w_j \,\mathrm{Cov}[\hat{x}_i, \hat{x}_j] \;=\; \sigma_{\hat{x}}^2 \,\|\boldsymbol{w}\|^2 \;+\; C(\hat{\boldsymbol{x}}, \boldsymbol{w}),$$

where $C(\hat{\boldsymbol{x}}, \boldsymbol{w}) = \sum_{i \ne j} w_i w_j \,\mathrm{Cov}[\hat{x}_i, \hat{x}_j]$ and $\mathrm{Cov}[\hat{x}_i, \hat{x}_j]$ is the cross-coordinate covariance induced by the shared normalization denominator. Note that Eq. 8 also holds for non-i.i.d. $\boldsymbol{x}$ and non-exchangeable $\hat{\boldsymbol{x}}$. It only requires that the $\hat{x}_i$ share the same mean and variance. Thus, the observations in Sec. 4.3 and 4.4 also hold in that case.

B.2 PROOF OF PROPOSITION 1 (PRE-NORM DROPOUT)

Recall from Sec. 4.2 that under Pre-Norm dropout the normalization input decomposes as $\boldsymbol{x} = \boldsymbol{r} + \boldsymbol{u}$, with $\boldsymbol{u}$ (mean μ_u, variance σ_u^2) the sublayer output and $\boldsymbol{r}$ the residual; the masked input and normalized coordinate are $\tilde{x}_i = r_i + \frac{b_i}{1-p} u_i$ and $z_i^{\mathrm{PN}} = \tilde{x}_i/(\mathrm{RMS}(\tilde{\boldsymbol{x}}) + \epsilon)$, Eq. (9). We take the pairs (r_i, u_i) i.i.d. across coordinates and independent of the masks $\{b_j\}$, with no independence assumed between r_i and u_i; throughout, $\mathbb{E}[b_i] = \mathbb{E}[b_i^2] = 1 - p$. Write $\kappa = \mathbb{E}[u_i^2]/\mathbb{E}[x_i^2] = \mathbb{E}[u_i^2]/(\mu_x^2 + \sigma_x^2)$ as in Sec. 4.2.

Moments of the masked input. Inverted dropout preserves the mean,

$$\mathbb{E}[\tilde{x}_i] = \mathbb{E}[r_i] + \frac{\mathbb{E}[b_i]}{1-p}\mathbb{E}[u_i] = \mathbb{E}[r_i] + \mathbb{E}[u_i] = \mu_x,$$

and the masked second moment is

$$\begin{aligned}\mathbb{E}[\tilde{x}_i^2] &= \mathbb{E}[r_i^2] + 2\,\frac{\mathbb{E}[b_i]}{1-p}\,\mathbb{E}[r_i u_i] + \frac{\mathbb{E}[b_i^2]}{(1-p)^2}\,\mathbb{E}[u_i^2] \\ &= \underbrace{\mathbb{E}[r_i^2] + 2\,\mathbb{E}[r_i u_i] + \mathbb{E}[u_i^2]}_{=\,\mathbb{E}[(r_i+u_i)^2] \,=\, \mu_x^2 + \sigma_x^2} + \Big(\tfrac{1}{1-p} - 1\Big)\mathbb{E}[u_i^2] \;=\; (\mu_x^2 + \sigma_x^2)\Big(1 + \tfrac{p}{1-p}\,\kappa\Big),\end{aligned}$$

using $\frac{1}{1-p} - 1 = \frac{p}{1-p}$ and $\mathbb{E}[u_i^2] = \kappa\,(\mu_x^2 + \sigma_x^2)$. The cross term $2\,\mathbb{E}[r_i u_i]$ is identical at training and inference—the inverted-dropout factor restores the mask expectation to one—so the correlation between r_i and u_i cancels and only the diagonal term $\mathbb{E}[u_i^2]$ is inflated. Since the mean is preserved, $\mathrm{Var}[\tilde{x}_i] = \mathbb{E}[\tilde{x}_i^2] - \mu_x^2 = \sigma_x^2 + \frac{p}{1-p}\,\kappa\,(\mu_x^2 + \sigma_x^2)$.

Normalized coordinate. By the strong law of large numbers, $\mathrm{RMS}(\tilde{\boldsymbol{x}})^2 = \frac{1}{d}\sum_j \tilde{x}_j^2 \to (\mu_x^2 + \sigma_x^2)\big(1 + \frac{p}{1-p}\,\kappa\big)$ almost surely as $d \to \infty$, while at inference $\mathrm{RMS}(\boldsymbol{x})^2 \to \mu_x^2 + \sigma_x^2$. Hence $z_i^{\mathrm{PN}} = \tilde{x}_i/(\mathrm{RMS}(\tilde{\boldsymbol{x}}) + \epsilon) \to \tilde{x}_i/\big(\sqrt{(\mu_x^2 + \sigma_x^2)(1 + \frac{p}{1-p}\,\kappa)} + \epsilon\big)$ almost surely, and since $|z_i^{\mathrm{PN}}| \le |\tilde{x}_i|/\epsilon$ is an integrable dominating envelope that does not depend on d, dominated convergence yields

$$\mathbb{E}[z_i^{\mathrm{PN}}] \to \frac{\mu_x}{\sqrt{(\mu_x^2 + \sigma_x^2)\big(1 + \frac{p}{1-p}\,\kappa\big)} + \epsilon}, \qquad \text{and, at inference,} \qquad \mathbb{E}[\hat{x}_i] \to \frac{\mu_x}{\sqrt{\mu_x^2 + \sigma_x^2} + \epsilon}.$$

Aggregation through the linear layer. The mean claim needs only linearity: $\mathbb{E}[y^{\mathrm{PN}}] = \mathbb{E}[z_i^{\mathrm{PN}}]\sum_i w_i$ and $\mathbb{E}[y^{\mathrm{test}}] = \mathbb{E}[\hat{x}_i]\sum_i w_i$, so for $\mu_x \neq 0$ and any weights with $\sum_i w_i \neq 0$ the ratio is free of $\boldsymbol{w}$ at every finite d and

$$\frac{\mathbb{E}[y^{\mathrm{PN}}]}{\mathbb{E}[y^{\mathrm{test}}]} \;=\; \frac{\mathbb{E}[z_i^{\mathrm{PN}}]}{\mathbb{E}[\hat{x}_i]} \;\to\; \frac{\sqrt{\mu_x^2 + \sigma_x^2} + \epsilon}{\sqrt{(\mu_x^2 + \sigma_x^2)\big(1 + \frac{p}{1-p}\,\kappa\big)} + \epsilon} \;=\; c_\epsilon \;=\; \Big(1 + \tfrac{p}{1-p}\kappa\Big)^{-1/2} + O(\epsilon)$$

(when $\mu_x = 0$, both $\mathbb{E}[z_i^{\mathrm{PN}}]$ and $\mathbb{E}[\hat{x}_i]$ tend to 0 and there is no mean mismatch to speak of). □

B.3 Pre-Linear dropout

The mismatch of Pre-Linear dropout, Eq. (11), is a classical property of inverted dropout, so we state and prove it here rather than in the main text.

Proposition 2 (Pre-Linear dropout)**.** *The pre-activation moments satisfy*

$$\mathbb{E}[y^{\mathrm{PL}}] \;=\; \mathbb{E}[y^{\mathrm{test}}], \qquad \mathrm{Var}[y^{\mathrm{PL}}] \;=\; \mathrm{Var}[y^{\mathrm{test}}] \;+\; \frac{p}{1-p}\,\mathbb{E}[\hat{x}_i^2]\,\|\boldsymbol{w}\|^2. \tag{15}$$

Proof. By independence of b_i and $\hat{x}_i$, and using $\mathbb{E}[b_i] = \mathbb{E}[b_i^2] = 1 - p$,

$$\mathbb{E}[z_i^{\mathrm{PL}}] = \frac{1}{1-p}\,\mathbb{E}[b_i]\,\mathbb{E}[\hat{x}_i] = \mu_{\hat{x}}, \qquad \mathbb{E}[(z_i^{\mathrm{PL}})^2] = \frac{1}{(1-p)^2}\,\mathbb{E}[b_i^2]\,\mathbb{E}[\hat{x}_i^2] = \frac{\mathbb{E}[\hat{x}_i^2]}{1-p}.$$

Hence, since $\mathbb{E}[\hat{x}_i^2] - \mu_{\hat{x}}^2 = \sigma_{\hat{x}}^2$,

$$\mathrm{Var}[z_i^{\mathrm{PL}}] \;=\; \frac{\mathbb{E}[\hat{x}_i^2]}{1-p} - \mu_{\hat{x}}^2 \;=\; \sigma_{\hat{x}}^2 \;+\; \frac{p}{1-p}\,\mathbb{E}[\hat{x}_i^2].$$

Moreover, the cross-coordinate covariances are preserved exactly: for $i \neq j$, since (b_i, b_j) are independent of $(\hat{x}_i, \hat{x}_j)$ and of each other,

$$\mathrm{Cov}[z_i^{\mathrm{PL}}, z_j^{\mathrm{PL}}] \;=\; \frac{\mathbb{E}[b_i]\,\mathbb{E}[b_j]}{(1-p)^2}\,\mathbb{E}[\hat{x}_i\hat{x}_j] - \mathbb{E}[\hat{x}_i]\,\mathbb{E}[\hat{x}_j] \;=\; \mathrm{Cov}[\hat{x}_i, \hat{x}_j]\,:$$

the $1/(1-p)$ rescaling exactly undoes the mask's attenuation of the covariances, so the covariance terms of $\mathrm{Var}[y^{\mathrm{PL}}]$ and $\mathrm{Var}[y^{\mathrm{test}}]$ are both equal to $C(\hat{\boldsymbol{x}}, \boldsymbol{w})$ and cancel in the difference. Aggregation through the linear layer therefore gives

$$\mathbb{E}[y^{\mathrm{PL}}] = \mathbb{E}[y^{\mathrm{test}}], \qquad \mathrm{Var}[y^{\mathrm{PL}}] = \mathrm{Var}[y^{\mathrm{test}}] + \frac{p}{1-p}\,\mathbb{E}[\hat{x}_i^2]\,\|\boldsymbol{w}\|^2. \quad \square$$

B.4 Variance-Preserving Affine Dropout

The scale and shift α, β of (12) are the affine correction of Alpha-Dropout (Klambauer et al., 2017) with dropped value $\alpha' = 0$, generalized to a non-centered activation; we state the moment-matching property and prove it here rather than in the main text.

Table 6: Hyperparameters of the trained models. Ls — number of layers, Dim — model's dimension, FFN-D — hidden dimension of FFN sublayers, H — number of heads, Len — context length, TT — training tokens, BT — tokens per batch, LR — learning rate, LRS — LR scheduler, Ann-R — annealing ratio, min-LR — annealing min LR. LR schedulers: Stable with Warm-Up and Annealing (SWUA), Stable with Warm-Up (SWU).

Model	p	Ls	Dim	FFN-D	H	Len	Data	TT	BT	LR	LRS	Warm-up	Ann-R	min-LR
Trained from scratch														
Transformers, NAPE	0.1	24	768	3072	12	2048	DCLM-Edu	40B	512K	$6 \cdot 10^{-4}$	SWUA	2.5%	20%	$6 \cdot 10^{-5}$
Mamba2-85M, State=128	0.1	12	768	-	24	1024	The Pile	25B	512K	$1 \cdot 10^{-3}$	Cosine	10%	-	$1 \cdot 10^{-4}$
Mamba2-45M, State=128	0.1	12	768	-	24	50-100	Copy	0.3B	19K	$1 \cdot 10^{-3}$	Cosine	5%	-	0
CPT														
SmolLM3, SW	0.1-0.3	36	2048	11008	16	4096	DCLM-Edu	1B	1M	$2 \cdot 10^{-5}$	SWU	10%	-	$2 \cdot 10^{-5}$
SmolLM3-8K, SW	0.3	36	2048	11008	16	8192	DCLM-Edu	1.5B	1M	$2 \cdot 10^{-5}$	SWU	10%	-	$2 \cdot 10^{-5}$
Mamba2-1.3B, State=128	0.1-0.3	48	2048	-	64	2048	DCLM-Edu	0.5B	256K	$1 \cdot 10^{-4}$	Cosine	10%	-	$1 \cdot 10^{-5}$

Proposition 3 (Variance-Preserving Affine Dropout)**.** *With the scale α and shift β of* (12)*, the corrected activation* (12) *preserves both moments of the normalized activation, $\mathbb{E}[z_i^{\mathrm{VPAD}}] = \mu_{\hat{x}}$ and $\mathrm{Var}[z_i^{\mathrm{VPAD}}] = \sigma_{\hat{x}}^2$; consequently*

$$\mathbb{E}[y^{\mathrm{VPAD}}] \;=\; \mathbb{E}[y^{\mathrm{test}}], \qquad \mathrm{Var}[y^{\mathrm{VPAD}}] \;=\; \mathrm{Var}[y^{\mathrm{test}}] \;+\; \big(\alpha^2(1-p)^2 - 1\big)\, C(\hat{\boldsymbol{x}}, \boldsymbol{w}). \tag{16}$$

Proof. Throughout, $\hat{x}_i$ is independent of b_i, with $\mathbb{E}[b_i] = \mathbb{E}[b_i^2] = 1-p$, and $\mathbb{E}[\hat{x}_i] = \mu_{\hat{x}}$, $\mathrm{Var}[\hat{x}_i] = \sigma_{\hat{x}}^2$, hence $\mathbb{E}[\hat{x}_i^2] = \sigma_{\hat{x}}^2 + \mu_{\hat{x}}^2$.

Mean. Since β is deterministic and $\hat{x}_i \perp b_i$,

$$\mathbb{E}[z_i^{\mathrm{VPAD}}] \;=\; \alpha\, \mathbb{E}[b_i]\, \mathbb{E}[\hat{x}_i] + \beta \;=\; \alpha\,(1-p)\,\mu_{\hat{x}} + \beta.$$

Substituting $\beta = \mu_{\hat{x}}\,\big(1 - \alpha\,(1-p)\big)$ yields $\mathbb{E}[z_i^{\mathrm{VPAD}}] = \mu_{\hat{x}}$.

Variance. Since β is constant,

$$\begin{aligned}\mathrm{Var}[\hat{x}_i\, b_i] &= \mathbb{E}[b_i^2]\, \mathbb{E}[\hat{x}_i^2] - \big(\mathbb{E}[b_i]\, \mathbb{E}[\hat{x}_i]\big)^2 \\ &= (1-p)(\sigma_{\hat{x}}^2 + \mu_{\hat{x}}^2) - (1-p)^2 \mu_{\hat{x}}^2 \;=\; (1-p)\,(\sigma_{\hat{x}}^2 + p\,\mu_{\hat{x}}^2),\end{aligned}$$

using $1 - (1-p) = p$. Therefore $\mathrm{Var}[z_i^{\mathrm{VPAD}}] = \alpha^2\,(1-p)\,(\sigma_{\hat{x}}^2 + p\,\mu_{\hat{x}}^2)$; setting this equal to $\sigma_{\hat{x}}^2$ gives $\alpha^2 = \sigma_{\hat{x}}^2 / \big((1-p)(\sigma_{\hat{x}}^2 + p\,\mu_{\hat{x}}^2)\big)$, well-defined and positive since $\sigma_{\hat{x}}^2 + p\,\mu_{\hat{x}}^2 > 0$ for any non-degenerate activation.

Aggregation, with the covariance term. For the pre-activation mean, linearity gives $\mathbb{E}[y^{\mathrm{VPAD}}] = \mu_{\hat{x}} \sum_i w_i = \mathbb{E}[y^{\mathrm{test}}]$. For the variance, the covariance term must be tracked: for $i \neq j$, since (b_i, b_j) are independent of $(\hat{x}_i, \hat{x}_j)$ and of each other,

$$\mathrm{Cov}[z_i^{\mathrm{VPAD}}, z_j^{\mathrm{VPAD}}] \;=\; \alpha^2\, \mathbb{E}[b_i]\, \mathbb{E}[b_j]\, \big(\mathbb{E}[\hat{x}_i \hat{x}_j] - \mathbb{E}[\hat{x}_i]\, \mathbb{E}[\hat{x}_j]\big) \;=\; \alpha^2 (1-p)^2\, \mathrm{Cov}[\hat{x}_i, \hat{x}_j]\,:$$

every pairwise covariance is attenuated by the fixed factor $\alpha^2(1-p)^2 = \frac{\sigma_{\hat{x}}^2 (1-p)}{\sigma_{\hat{x}}^2 + p\,\mu_{\hat{x}}^2} \in (0,1)$, while at inference the covariances enter at full strength. □

C Experimental Details

C.1 Experimental Setup

Dropout implementation. As discussed in Sec. 4, Pre-Norm dropout is applied before RMSNorm and thus introduces one source of randomness per sublayer. Pre-Linear and VPAD, however, can have either one or several sources of randomness: a single mask after RMSNorm output or an independent mask per linear projection. For a fair comparison, we apply dropout once after RMSNorm, so that Pre-Linear and VPAD also have one source of randomness per sublayer. Future work could explore applying the mask per linear layer to introduce more randomness.

Training from scratch. For experiments with transformers, we use a Llama-3-like model but replace RoPE with NAPE (Vasylenko et al., 2026), which has shown its effectiveness at extrapolation without any post-training intervention; we use no attention scaling or any other intervention related to positional encoding. We use the DCLM-Edu (allal et al., 2025) dataset for training. We train three 400M-parameter models with different initializations for each of five settings: weight decay (WD), no weight decay (no-WD), Pre-Norm Dropout (PN-D), Pre-Linear Dropout (PL-D), and VPAD. Fig. 1 reports results for the best-performing model, and Fig. 4 reports the average over 3 seeds. The best model is selected based on the average RULER performance at 4–8K. We show the best-performing model to demonstrate that even the best WD model overfits badly when trained longer. The overall picture can be concluded from Fig. 4, which shows the average performance: for instance, no-WD overfits more on average.

Continual Pre-training. In order to test dropout at scale, we resort to continual pre-training of larger models. For transformers, we use the SmolLM3-3B checkpoint (stage 3, step 4720000), pre-trained with a context length of 4096 tokens. We convert its RoPE layers to sliding-window (SW) attention with a window size of 4096 while leaving the NoPE layers untouched (i.e. they use global attention). The model is then trained with the same 4096 context size on about 1B tokens. We evaluate the models on NIAH-1, RULER (Hsieh et al., 2024) and a subset of HELMET (Yen et al., 2025) tasks: In-Context Learning (ICL), Retrieval-Augmented Generation (RAG), and Long-document QA (LongQA).

CPT: complementarity of dropout and attention scaling. We train a SmolLM3 model with an 8K context size and a 4K sliding window on 1.5B tokens. Training on an 8K context allows the model to learn the sliding-window behavior better. It also allows us to report ID performance, since the HELMET evaluation starts at a sequence length of 8K. The probability of dropout is set to $p = 0.3$. We call this model SmolLM3-SW-8K and the scaled variant — SmolLM3-SW-8K S$\times$. For the scaled model, we use an input-and-length-dependent scaling function introduced in Vasylenko et al. (2026). We compare the performance of our models with the original SmolLM3-Base (Bakouch et al., 2025) and SmolLM3-Base with the YaRN RoPE extension. The released SmolLM3-Base checkpoint was trained on around 100B additional tokens after the stage 3 checkpoint (the checkpoint that we use in all our experiments with SmolLM3). We evaluate these models on the ICL tasks from HELMET and generate additional 256K-length and 512K-length sets (only possible for ICL tasks) with 100 samples per task (so 500 in total per set).

For all experiments, we use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and gradient clipping of 1.0. All experiments labeled "WD" use weight decay 0.1 and 0 otherwise. All SW models are trained with a sliding window of 4096. We train models from scratch with TorchTitan (Liang et al., 2025) and train the CPT models with the Hugging Face Transformers library (Wolf et al., 2020). Additional hyperparameters are given in Table 6.

C.2 MAMBA2

Dropout placement Mamba2 blocks have two linear projections: the input projection, $\boldsymbol{W}_{\text{in}}$, and the output projection. To match the transformer experiments as closely as possible, we work with $\boldsymbol{W}_{\text{in}}$: like $W_{\{Q,K,V,G,U\}}$ in the attention and FFN sublayers, it is applied directly to the normalized input. Since the order of RMSNorm and the input projection is the same as in transformers, we can test all dropout variants. The implementation is also identical: dropout is patched into the RMSNorm module rather than attached to $\boldsymbol{W}_{\text{in}}$.

Synthetic Copying Following Buitrago & Gu (2025), we test our findings on the copying task (Jelassi et al., 2024). The source sequence length during training ranges from 50 to 100 tokens, and the number of training samples is 5M. We train five models with different random seeds for each setting. We report the mean Exact Match Accuracy at a sequence length of 150 (EMA-150) and Character-Level Accuracy at 300 (CLA-300) (as in Buitrago & Gu (2025)) in Table 4; standard deviations over five seeds are given in Table 13 (App. D). We chose sequences of length 150 because the Mamba2 models barely generalize to this length, while at length 200, all models achieve zero performance. This is consistent with the results reported by Jelassi et al. (2024).

Table 7: SmolLM3-3B results on short benchmarks. HS - HellaSwag, WG - Winogrande, LMBD - Lambada, OPQA - OpenBookQA

Model	p	ARC-E	ARC-C	HS	PIQA	WG	OPQA	BoolQ	LMBD	Avg.
Original	-	77.8	53.2	75.3	79.0	69.6	44.2	80.4	70.8	68.8
CPT										
WD	-	77.7	53.2	75.5	79.2	68.9	43.4	79.9	70.5	68.5
no-WD	-	77.6	53.3	75.4	79.1	69.6	43.6	80.0	70.6	68.7
PN-D	0.1	78.3	53.2	74.7	79.2	69.4	45.6	78.3	69.7	68.5
PL-D	0.1	79.0	54.3	74.9	79.4	69.1	44.6	81.0	70.4	69.1
VPAD	0.1	79.3	54.9	75.6	79.5	69.8	44.4	80.3	69.9	69.2
PN-D	0.2	78.3	53.0	73.4	79.2	67.7	45.0	78.5	69.8	68.1
PL-D	0.2	78.4	51.7	72.8	78.9	68.7	44.2	80.2	69.6	68.1
VPAD	0.2	79.4	54.4	74.1	79.9	68.3	44.2	80.2	69.2	68.7
PN-D	0.3	77.5	51.5	71.4	77.6	66.5	44.0	78.3	70.0	67.1
PL-D	0.3	76.6	47.7	68.1	77.1	67.2	43.0	77.7	67.8	65.7
VPAD	0.3	77.9	52.6	71.7	79.1	66.7	43.6	78.5	67.6	67.2

Training from scratch In the case of Mamba2, we follow Buitrago & Gu (2025) and train Mamba2 models with 85M parameters, also with three random initializations per setting. The training budget of the Mamba2 models is 25B tokens, as the degradation of length generalization is already noticeable at this stage.

CPT We apply CPT to a 1.3B-parameter Mamba2 model, training for 0.5B tokens.

C.3 Evaluation

We evaluate models' performance via CE loss and Recall-Based Accuracy (RBA) on the Needle-in-a-Haystack task (NIAH-1 of RULER, also known as Passkey Retrieval) and on overall RULER. To evaluate transformers' CE loss, we use 2000 long documents from ArXiv papers—a subset of The Common Pile dataset (Kandpal et al., 2025)—and, for Mamba2, we follow Buitrago & Gu (2025) by using The Pile (Gao et al., 2020). We use the LM Evaluation Harness (Gao et al., 2024) for the short-context benchmarks and RULER. For HELMET (Yen et al., 2025), we use its official implementation.

D Additional empirical results

D.1 Results on Short Benchmarks

We report performance on standard short-context benchmarks in Table 7.

D.2 Extended SmolLM3 Results on RULER

Table 8 shows extended NIAH-1 and RULER results for various dropout probability values.

D.3 Checkpoint Averaging and Attention Dropout

Checkpoint averaging (CA) builds a single model by averaging the weights of several checkpoints collected along one training trajectory, $\bar{\boldsymbol{\theta}} = \frac{1}{K}\sum_{t=1}^{K}\boldsymbol{\theta}_t$, where $\boldsymbol{\theta}_t$ are the parameters at the saved checkpoints. Because it reuses checkpoints already produced during training, CA improves generalization at essentially no additional cost, and it has long been a standard practice. For example, it is used in the original Transformer (Vaswani et al., 2017). To some extent, CA can be related to stochastic weight averaging (Izmailov et al., 2018) and, thus, can be considered as implicit regularization (Morales-Brotons et al., 2024). We believe that, by investigating regularization methods such as CA and dropout, we show that models suffer from *length-based overfitting* that is not related to architectural components but rather is a consequence of the training process.

Table 8: SmolLM3 performance on NIAH-1 and on RULER for all dropout probabilities. Notation as in Table 1. "— V —" means that all entries are equal to V in the range; $V_1 \to V_2$ means that the left bound of the range has value V_1 and the right one V_2. Grey columns — ID length.

	NIAH-1					RULER			
Model	4K	8K	16K	32-128K	256K	4K	8K	16K	32K
Orig.	100.0	0.0	0.0	— 0.0 —	0.0	83.5	0.1	0.0	0.0
SW	100.0	0.0	0.0	— 0.0 —	0.0	83.4	0.5	0.0	0.0
CPT									
SW, WD	100.0	0.0	0.0	— 0.0 —	0.0	83.5	1.1	0.0	0.0
SW, no-WD	100.0	0.0	0.0	— 0.0 —	0.0	83.5	1.1	0.0	0.0
CPT: PN-D									
SW, PN-D $p = 0.1$	100.0	0.0	0.0	— 0.0 —	0.0	80.8	2.6	0.7	0.2
SW, PN-D $p = 0.2$	100.0	0.6	0.0	— 0.0 —	0.0	78.4	3.7	1.9	1.5
SW, PN-D $p = 0.3$	100.0	82.2	6.2	— 0.0 —	0.0	76.1	22.5	2.1	0.9
CPT: PL-D									
SW, PL-D $p = 0.1$	100.0	100.0	75.4	$18.4 \to 0.2$	0.0	79.7	47.8	13.7	3.4
SW, PL-D $p = 0.2$	100.0	100.0	100.0	$100.0 \to 94.4$	36.6	78.9	50.9	38.2	21.7
SW, PL-D $p = 0.3$	100.0	100.0	100.0	— 100.0 —	100.0	74.1	50.6	40.2	25.5
CPT: VPAD									
SW, VPAD $p = 0.1$	100.0	100.0	57.2	$30.4 \to 0.0$	0.0	80.4	45.2	5.9	2.0
SW, VPAD $p = 0.2$	100.0	100.0	100.0	$96.2 \to 86.0$	49.4	79.9	55.0	38.8	19.4
SW, VPAD $p = 0.3$	100.0	100.0	100.0	— 100.0 —	99.2	77.8	54.3	44.1	27.4

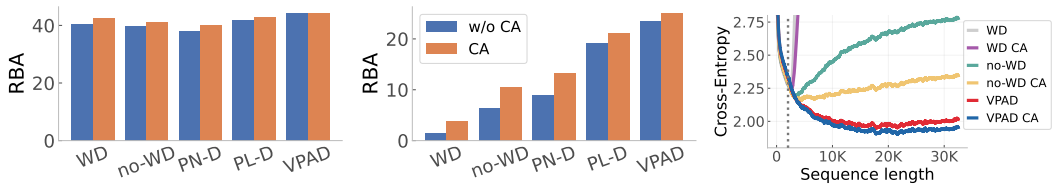

Figure 6: Performance of models trained from scratch before and after applying checkpoint averaging. **Left:** RULER: in-distribution length (2K); **Middle:** RULER: out-of-distribution—average of performance at 4K and 8K; **Right:** Cross-Entropy loss on ArXiv papers from The Common Pile.

For the experiments with checkpoint averaging, we average 10 checkpoints of transformers, which constitute 25% of the training span. We compare the models trained from scratch before and after applying checkpoint averaging in Fig. 6 (see App. C for details). It is evident that the averaging improves the CE loss, especially for the no-WD model. Moreover, this improvement in CE loss transfers to RULER. We observe an improvement in the OOD performance of all models, including those with dropout (though with a diminishing effect), but the models trained without dropout benefit the most. The ID performance is slightly improved as well, except for VPAD, where the performance is unchanged. In the case of SmolLM3, we observe that, even without CPT, CA is able to improve the extrapolation of the SW variant on NIAH-1 and on RULER overall (see Tables 9 and 10). Note that the model was never exposed to a context length larger than 4096 and was never trained with a sliding window. However, it is capable of operating in an SW setting without any additional training.

Additionally, we observe extrapolation improvements when SmolLM3 is fine-tuned with attention dropout, which drops weights in the attention matrix (see Table 10). However, the behavior differs on NIAH-1, where attention dropout brings no improvement at all (Table 9).

D.4 ASEntmax

Vasylenko et al. (2026) showed that adaptive-sparsity transformers length-generalize better than dense (softmax-based) transformers. We observe a similar pattern, and dropout intensifies the effect (Table 11).

Table 9: Performance on NIAH-1 of SmolLM3 fine-tuned with attention dropout

Model	4K	8K	16K
Orig.	100.0	0.0	0.0
Checkpoint Avg.			
SW, $K = 10$	100.0	41.2	0.8
SW, $K = 20$	97.2	66.4	11.4
SW, $K = 30$	80.4	68.2	2.2
CPT: Attention Dropout			
SW, no-WD, $p = 0.1$	100.0	0.0	0.0
SW, no-WD, $p = 0.2$	100.0	0.0	0.0
SW, no-WD, $p = 0.3$	100.0	0.0	0.0

Table 10: Performance on RULER of SmolLM3 fine-tuned with attention dropout

Model	4K	8K	16K
Orig.	83.5	0.1	0.0
Checkpoint Avg.			
SW, $K = 10$	79.4	11.4	0.7
SW, $K = 20$	61.6	13.8	1.6
SW, $K = 30$	43.2	10.7	1.7
CPT: Attention Dropout			
SW, no-WD, $p = 0.1$	80.7	4.3	0.1
SW, no-WD, $p = 0.2$	76.2	13.4	0.9
SW, no-WD, $p = 0.3$	74.3	7.7	1.3

D.5 SmolLM3: NIAH-1 and Checkpoint Averaging

Table 12 shows extended NIAH-1 and checkpoint-averaging results.

D.6 Mamba2

Tables 13 and 14 show the results of the small models trained from scratch on synthetic copying and NIAH-1. Table 15 shows extended results on NIAH-1 and RULER.

Table 11: Performance of ASEntmax models on RULER

Model	2K	4K	8K	16K	32K
Softmax WD	40.39	3.09	0.06	0.04	0.04
Softmax VPAD	44.24	34.76	12.30	4.54	1.93
ASEntmax WD	43.96	20.94	5.44	2.97	2.44
ASEntmax PN-D	42.76	26.85	7.72	2.36	1.59
ASEntmax PL-D	47.30	36.80	20.11	11.39	6.73
ASEntmax VPAD	48.09	37.32	21.64	13.51	6.98

Table 12: Full SmolLM3 results on NIAH-1.

Model	4K	8K	16K	32K	64K	128K	256K
SmolLM3	100.0	0.0	0.0	0.0	0.0	0.0	0.0
SmolLM3-SW	100.0	0.0	0.0	0.0	0.0	0.0	0.0
Checkpoint Averaging							
SmolLM3-SW, $K = 10$	100.0	41.2	0.8	0.0	0.0	0.0	0.0
SmolLM3-SW, $K = 20$	97.20	66.4	11.4	0.0	0.0	0.0	0.0
SmolLM3-SW, $K = 30$	80.40	68.2	2.2	0.0	0.0	0.0	0.0
Continued Pretraining							
SmolLM3-SW, WD	100.0	0.0	0.0	0.0	0.0	0.0	0.0
SmolLM3-SW, no-WD	100.0	0.0	0.0	0.0	0.0	0.0	0.0
SmolLM3-SW, PN-D $p = 0.1$	100.0	0.0	0.0	0.0	0.0	0.0	0.0
SmolLM3-SW, PN-D $p = 0.2$	100.0	0.6	0.0	0.0	0.0	0.0	0.0
SmolLM3-SW, PN-D $p = 0.3$	100.0	82.2	6.2	0.2	0.0	0.0	0.0
SmolLM3-SW, PL-D $p = 0.1$	100.0	100.0	75.4	18.4	2.8	0.2	0.0
SmolLM3-SW, PL-D $p = 0.2$	100.0	100.0	100.0	100.0	100.0	94.4	36.6
SmolLM3-SW, PL-D $p = 0.3$	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SmolLM3-SW, VPAD $p = 0.1$	100.0	100.0	57.2	30.4	5.6	0.0	0.0
SmolLM3-SW, VPAD $p = 0.2$	100.0	100.0	100.0	96.2	91.6	86.0	49.4
SmolLM3-SW, VPAD $p = 0.3$	100.0	100.0	100.0	100.0	100.0	100.0	99.20

Table 13: Extrapolation results of Mamba2 models on the synthetic copying task (mean ± standard deviation over 5 seeds)

Model	EMA-150	CLA-300
WD	33.1 ± 21.9	42.8 ± 12.7
no-WD	42.7 ± 13.7	45.7 ± 14.5
PL-D	61.5 ± 19.6	50.9 ± 8.9
VPAD	68.9 ± 13.6	58.0 ± 10.8

Table 14: Mamba2-85M performance on NIAH-1 (mean ± standard deviation over 3 seeds)

Model	1K	2K	4K
WD	72.3 ± 17.8	30.5 ± 16.1	1.1 ± 1.1
no-WD	91.2 ± 11.1	50.6 ± 31.2	1.1 ± 0.8
PL-D	90.3 ± 4.4	43.4 ± 19.5	10.5 ± 11.1
VPAD	88.9 ± 13.7	76.4 ± 20.3	30.2 ± 22.6

Table 15: Mamba2-1.3B performance on NIAH-1 and on RULER.

	NIAH-1					RULER				
Model	2K	4K	8K	16K	32K	2K	4K	8K	16K	32K
Orig.	100.0	100.0	100.0	0.0	0.0	38.3	14.3	8.5	0.4	0.0
CPT										
WD	100.0	100.0	100.0	0.0	0.0	42.0	16.2	9.6	0.7	0.0
no-WD	100.0	100.0	100.0	0.0	0.0	42.2	16.0	9.6	0.0	0.0
PN-D $p = 0.2$	100.0	100.0	100.0	11.8	0.0	50.7	22.5	11.5	2.6	0.8
PN-D $p = 0.3$	100.0	100.0	99.8	39.0	0.4	50.9	24.8	12.8	5.1	1.7
PL-D $p = 0.1$	99.8	100.0	100.0	99.8	7.0	48.9	23.0	14.7	10.5	1.5
PL-D $p = 0.2$	100.0	100.0	100.0	100.0	22.8	48.9	28.8	18.0	13.7	5.3
PL-D $p = 0.3$	100.0	100.0	100.0	97.2	12.0	46.5	31.4	20.0	13.9	4.8
VPAD $p = 0.1$	100.0	100.0	100.0	99.8	13.0	50.6	23.1	14.8	10.4	2.0
VPAD $p = 0.2$	100.0	100.0	100.0	100.0	45.0	50.9	26.3	18.5	13.3	5.8
VPAD $p = 0.3$	100.0	100.0	100.0	99.8	23.8	46.6	24.8	18.9	15.7	5.5