Scale-invariant attention

Ben Anson
School of Mathematics
University of Bristol
ben.anson@bristol.ac.uk

Xi Wang
Department of Computer Science
Johns Hopkins University
xidulu@gmail.com

Laurence Aitchison
School of Computer Science
University of Bristol
laurence.aitchison@bristol.ac.uk

Abstract

One persistent challenge in LLM research is the development of attention mechanisms that are able to generalise from training on shorter contexts to inference on longer contexts. We propose two conditions that we expect all effective long-context attention mechanisms to have: scale-invariant total attention, and scale-invariant attention sparsity. Under a Gaussian assumption, we show that a simple position-dependent transformation of the attention logits is sufficient for these conditions to hold. Experimentally we find that the resulting scale-invariant attention scheme gives considerable benefits in terms of validation loss when zero-shot generalising from training on short contexts to validation on longer contexts, and is effective at long-context retrieval.

1 Introduction

One key challenge in modern LLMs is scaling up context length at inference time, while maintaining model performance. We approach this question of length generalisation by considering scale invariance. In particular, we are inspired by the "scale-invariant" statistics of natural images (Van der Schaaf & van Hateren, 1996). Scale invariance, for images, is actually highly intuitive, and means that there is structure at all spatial scales. For example, in an image there might be big features that are 100–1000 pixels across, and some small features that are only 1–10 pixels across. In natural images, features at both spatial scales are important: you cannot remove features at either scale without radically altering the image (Van der Schaaf & van Hateren, 1996).

For attention over text, instead of pixels, we considered *token ranges* at different scales:

- 1–10 tokens in the past (e.g. those in the same sentence),
- 10–100 tokens in the past (e.g. those in the same paragraph),
- 100–1,000 tokens in the past (e.g. those in the same section),
- 1,000–10,000 tokens in the past (e.g. those in the same document), and so on.

With this in mind, we define scale-invariant attention as any attention scheme that satisfies two desiderata: scale-invariant total attention and scale-invariant attention sparsity.

Scale-invariant total attention is the property that the sum of attention weights in each of the above ranges is roughly similar. Intuitively, that means that the model attends to both the local context (e.g. 10–100 tokens ago) while at the same time taking account of information from the global context (e.g. 1,000–10,000 tokens ago). Scale-invariant total attention addresses a key issue with attention mechanisms: as the context gets longer, models tend to pay more attention to distant tokens at the

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

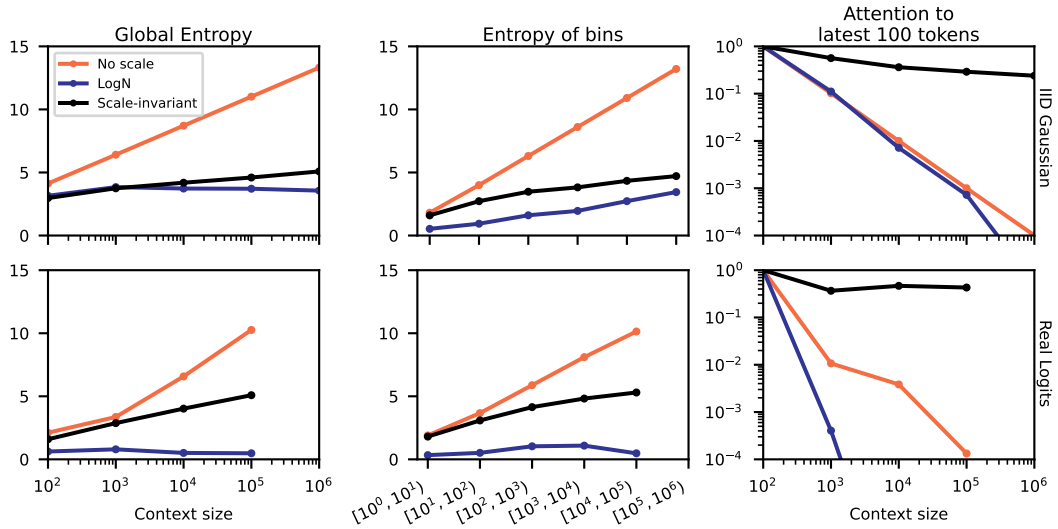


Figure 1: **Scale-invariant attention controls the entropy without sacrificing attention over the local context.** We consider three metrics for attention schemes: (*left*) the global attention entropy, (*middle*) entropy within particular ranges of tokens (e.g. 10–100), and (*right*) total attention to the previous 100 tokens. The top row uses IID Gaussian logits, following our theoretical approach in Sec. 3.2. For LogN, the IID logits are multiplied by $s \log N$, where N is the sequence length and $s = 0.4$. The bottom row uses attention logits sampled from models trained with p-RoPE and 'No scale', LogN, and our scale-invariant transform. With no logit scaling, the attention becomes increasingly diffuse as the context grows (i.e. the distribution over logits has high entropy). LogN scaling reduces the entropy and thus ensures that attention remains sparse even at longer contexts. However, LogN still forfeits the ability to attend to the local context (e.g. 100 most recent) tokens. Scale-invariant attention strikes a balance between low entropy and attending over the local context.

expense of the local context (e.g. Fig. 1, right column; the blue and orange lines decay quickly to zero). While scale-invariant total attention doesn't entirely eliminate that issue, it does ensure that the attention paid to the local context shrinks only very slowly as the context length grows (e.g. Fig. 1, right column; the black lines decay to zero much more slowly).

Total attention tells us the amount of attention a certain region receives. However, scale invariance of the total attention does not tell us about how the attention will be distributed among tokens in the region, i.e. whether the attention is spread out among many tokens or concentrated onto only a few tokens. As the region gets wider (e.g. from 10–100 tokens ago to 10,000–100,000 tokens ago), we might expect attention to spread out over more tokens simply due to the increased number of tokens. This "spreading out" of attention is possibly suboptimal for large contexts (Nakanishi, 2025), and that instead, we should focus attention on only a few of the most relevant tokens. To measure these effects, we use the entropy, which roughly captures the logarithm of the number of tokens attended to.

Scale-invariant attention sparsity captures this notion. In particular, we define two kinds of scale-invariant attention sparsity. **Strong** scale-invariant attention sparsity implies that the number of tokens attended to in each region is constant. For example, if the model attends to 8 tokens in the 10–100 token range, strong scale-invariant attention sparsity says it will also attend to approximately 8 tokens in the 1,000–10,000 token range. Strong scale-invariant attention sparsity implies an extreme increase in sparsity as the context gets longer that may be difficult to achieve with practical attention mechanisms. We therefore also introduce **weak** scale-invariant attention sparsity, which simply states that the sparsity increases as the context gets longer (i.e. attention is relatively dense in the region from 10–100 tokens, and much sparser in the region from 1,000–10,000, but you still attend to more tokens in the 1,000–10,000 region than the 10–100 region).

Our main contributions are:

- We introduce the concepts of scale-invariant total attention, and weak and strong scale-invariant attention sparsity as desirable properties when attending over long contexts.

- We derive a simple, position-dependent transformation of attention logits that provably satisfies scale-invariant total attention, and empirically satisfies weak scale-invariant attention sparsity for Gaussian logits.
- We implement scale-invariant attention in conjunction with p-RoPE (Barbero et al., 2024b). We show that our method, 'scale-invariant p-RoPE', exhibits improvements in validation loss both when doing long-context training, and when zero-shot generalising to longer contexts. Our method also matches the performance of the best alternatives in an out-of-distribution long context 'needle-in-a-haystack' task.

2 Related work

Handling long sequences effectively in Transformer-based models remains a significant challenge and is of high interest to the deep learning community (Bai et al., 2024; Ye et al., 2024; Jin et al., 2024; Beltagy et al., 2020; Ding et al., 2023; Munkhdalai et al., 2024; Bulatov et al., 2024; Liu et al., 2023a; Barbero et al., 2024a; Hu et al., 2024). Below, we discuss several strategies in this literature relating to our work.

Long contexts via the positional encoding: Methods like ALiBi (Press et al., 2021) introduce a static, causal bias directly into the attention logits. ALiBi endows the model with an inductive bias towards recent tokens (Kazemnejad et al., 2023), thus helping with longer contexts. Rotary Position Embeddings (RoPE) (Su et al., 2024) have emerged as the most popular position encoding scheme, encoding relative positions. RoPE does not generalise to longer sequences out of the box (Peng et al., 2023), leading to subsequent research improving RoPE's length extrapolation capabilities (Zhu et al., 2023; Wang et al., 2024). Many achieve this by modifying RoPE's frequency spectrum; for example, Positional Interpolation (PI) (Chen et al., 2023), NTK-aware scaling (bloc97, 2023), YaRN (Peng et al., 2023), and by simply increasing the RoPE base θ parameter (Grattafiori et al., 2024; Gemma Team, 2025; Barbero et al., 2024b; Liu et al., 2023b).

Entropy Control: Beyond positional information, the properties of the attention distribution itself are crucial, especially in long contexts where attention scores can "smear"/"spread out" across many tokens. Scalable-Softmax (SSMax), also known as the 'LogN trick' (Nakanishi, 2025; Chiang & Cholak, 2022; Jianlin, 2021; Bai et al., 2023; Llama 4 Team, 2025), addresses this by multiplying the attention logits by $s \log N$, where N is the context length and s is a learned scale parameter. This multiplier has the effect of sharpening/focusing the attention distribution. Li et al. (2025) adopt a similar approach: controlling the entropy of the attention distribution for better length generalization.

These prior works are perhaps the most similar to our approach. However, the key issue with such approaches is that they treat the local context (e.g. previous 10–100 tokens) in the same way as the global context (e.g. 10,000 tokens ago). As the context gets larger, the attention paid to the 100 most recent tokens drops rapidly for the prior approaches, but stays markedly more consistent with scale-invariant attention (e.g. for LogN see Fig. 1). In contrast, we started off by carefully specifying how we wanted attention to behave in the local and global contexts (Sec. 3.1) by giving the scale-invariant total attention and scale-invariant attention sparsity desiderata. This means that e.g. LogN has a position-independent multiplicative bias, whereas our approach has position-dependent multiplicative and additive biases for the logits.

Efficient Long-Context Training and Inference: A key observation of Xiong et al. (2023) is that continual pretraining on long contexts, after initial pretraining on shorter sequences, is often sufficient and much more computationally efficient. Thus, a simple strategy for improving long-context performance, given sufficient computational budget, is continual pretraining on longer contexts. This has been adopted widely (Grattafiori et al., 2024; Gao et al., 2024; Lieber et al., 2024; Yang et al., 2024; Cohere Team, 2025; Llama 4 Team, 2025; Liu et al., 2024a). Of course, this strategy considerably increases the complexity and memory cost of the pretraining pipeline, making approaches like ours that can zero-shot generalise very valuable. Furthermore, one might expect long-context training to be far easier (e.g. requiring fewer long-context training steps for optimal performance) for approaches that already have good long-context performance due to zero-shot generalisation.

For inference on sequences exceeding the trained context length, alternative strategies bypass applying dense attention over the entire context, allowing for 'infinite attention' (Munkhdalai et al., 2024; Liu et al., 2024b; Martins et al., 2021; Chen et al., 2025; Ding et al., 2023). These include maintaining a fixed-size attention window (Beltagy et al., 2020), retrieving relevant context tokens

using heuristics like top-K proximity (Han et al., 2023), or vector similarity search akin to episodic memory systems (Fountas et al., 2024; Xiao et al., 2024). While effective, these approaches operate at a higher level, managing context rather than modifying the core attention mechanism's ability to process it directly, which is the focus of our work.

3 Methods

Following FlexAttention (Dong et al., 2024), we define the attention "score" as the dot product of a query q and keys K,

$$S_t = \frac{1}{\sqrt{d}} \sum_{\lambda=1}^{d} q_\lambda K_{t\lambda}. \tag{1}$$

Here, d is the head dimension and $\lambda \in \{1, \ldots, d\}$ indexes the feature, and t indexes the position. We are using an unusual form for the sequence indices, but this simplifies notation later. Specifically, we consider a single, fixed query token q. Then, $t \geq 0$ into the keys/values from previous tokens. Critically, t counts backwards from the query token, e.g. $t = 1$ indicates the previous token. Logits are computed by applying an attention modifier function to the score,

$$L_t = L(S_t; t). \tag{2}$$

Here, L_t is the actual value of the attention logits for the token t steps back from the query, while $L(S_t; t)$ is the function used to compute those logits. Thus, $L(S_t; t) = S_t$ recovers standard attention.

For use later, we define the unnormalised attention weights $\tilde{A}_t$, (normalised) attention weights A_t, and normalisers Z, for a sequence of length T, as,

$$\tilde{A}_t = \exp(L_t) \qquad A_t = \frac{\tilde{A}_t}{Z} \qquad Z = \sum_{t=1}^{T} \tilde{A}_t. \tag{3}$$

3.1 Formal definitions of scale-invariant total attention and attention sparsity

Scale-invariant total attention. In our examples we have considered ranges of tokens, 10–100, 100–1,000, 1,000–10,000, 10,000–100,000, etc.. Scale-invariant total attention is the property that the total attention in each of these ranges is somewhat similar, such that the attention allotted to any one range does not dominate the others. We give a formal definition in Def. 3.1, where using $\Delta = 10$ gives the ranges from the examples.

Definition 3.1 (Scale-invariant total attention)**.** *Consider a set of random variables, $\{L_t\}$, representing attention logits. The total unnormalised attention in range t_1 to t_2 is,*

$$Z_{t_1}^{t_2} = \sum_{t=t_1}^{t_2-1} \tilde{A}_t = \sum_{t=t_1}^{t_2-1} \exp(L_t). \tag{4}$$

We say that the total attention is scale-invariant if, for any integer $\Delta > 1$, the expected total unnormalised attention in the range $\{t, t+1, \ldots, t\Delta - 1\}$ is asymptotically constant. That is,

$$\mathrm{E}\left[Z_t^{t\Delta}\right] = \Theta(1) \quad \textit{as} \quad t \to \infty. \tag{5}$$

The "$\Theta(1)$" is big-Θ notation, and means there exist constants $c_1, c_2 > 0$ and t_0 such that for all $t > t_0$, $c_1 \leq \mathrm{E}\left[Z_t^{t\Delta}\right] \leq c_2$.

Scale-invariant unnormalised attention sparsity. Remember that scale-invariant attention sparsity means that attention should be denser for the local context (e.g. 10–100 tokens ago) and sparser for the global context (e.g. 1,000–10,000) tokens ago. To measure attention sparsity, we consider the number of tokens attended to in a region. To evaluate the number of tokens attended to, one approach is to use the entropy over tokens. We measure the sparsity in the region from t_1 to t_2, using the entropy for the distribution over tokens in this region,

$$H_{t_1}^{t_2} = -\sum_{t=t_1}^{t_2-1} \frac{\tilde{A}_t}{Z_{t_1}^{t_2}} \log\left(\frac{\tilde{A}_t}{Z_{t_1}^{t_2}}\right) = -\frac{\sum_{t=t_1}^{t_2-1} \tilde{A}_t \log \tilde{A}_t}{Z_{t_1}^{t_2}} + \frac{\log Z_{t_1}^{t_2} \sum_{t=t_1}^{t_2-1} \tilde{A}_t}{Z_{t_1}^{t_2}}, \tag{6}$$

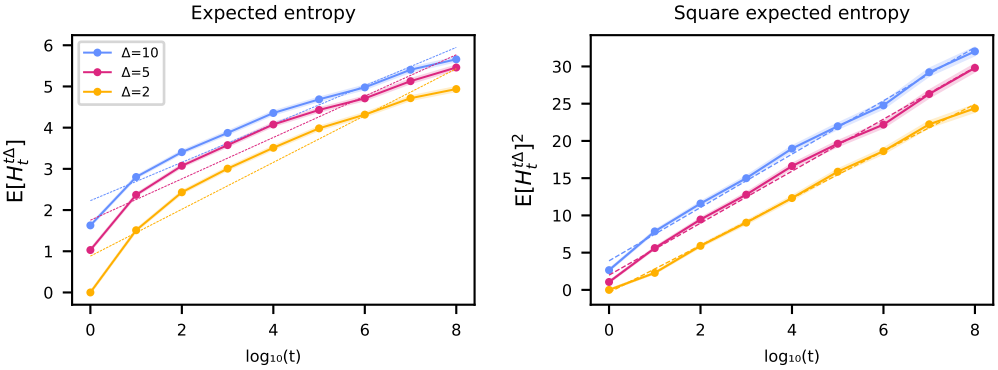


Figure 2: **Expected entropy of scale-invariant attention at different scales is sub-logarithmic**. Here, we sample sequences of independent standard Gaussian logits, and apply the scale-invariant attention transformation. We estimate the expected entropy in ranges $[t, t\Delta)$, where the size of the range is controlled by t (x-axis) and Δ (line color). We see that this expected entropy measure scales sub-logarithmically (left), and with the right plot suggesting a $\sim \sqrt{\log(t)}$ scaling. The dashed lines show a best linear fit.

where A_t is defined in Eq. (3). Defining the unnormalised negentropy as,

$$\tilde{N}_{t_1}^{t_2} = \sum_{t=t_1}^{t_2-1} \tilde{A}_t \log \tilde{A}_t, \tag{7}$$

and remembering the definition $Z_{t_1}^{t_2}$ (Eq. 4), we can then write the entropy in range t_1 to t_2 as,

$$H_{t_1}^{t_2} = -\frac{\tilde{N}_{t_1}^{t_2}}{Z_{t_1}^{t_2}} - \log Z_{t_1}^{t_2}. \tag{8}$$

Thus, it seems reasonable to consider the behavior of the unnormalised negentropy (Eq. 7), because if $\tilde{N}_t^{t\Delta}$ and $Z_t^{t\Delta}$ are asymptotically constant as $t \to \infty$, then by Eq. (8) we expect $H_{t_1}^{t_2}$ to also be asymptotically constant. Thus we define:

Definition 3.2 (Scale-invariant unnormalised attention sparsity)**.** *Consider a set of random variables, $\{L_t\}$, representing attention logits. We say that we have scale-invariant unnormalised attention sparsity if, for any integer $\Delta > 0$,*

$$\mathrm{E}\left[\tilde{N}_t^{t\Delta}\right] = \Theta(1) \quad \textit{as} \quad t \to \infty. \tag{9}$$

where $\tilde{N}_t^{t\Delta}$ is the unnormalised negentropy (Eq. 7).

Weak and strong scale-invariant attention sparsity. While the argument above suggests that scale-invariant unnormalised attention sparsity is an important property, ultimately we are interested in giving formal definitions of weak and strong attention sparsity.

We define weak scale-invariant attention sparsity such that as input lengths increase — for example, from 10–100 to 100–1,000 tokens — the number of attended tokens grows sublinearly. In contrast, standard attention with unscaled logits yields linear growth. Since entropy is roughly the log of the number of attended tokens (a uniform distribution over k tokens has entropy equal to $\log k$), weak sparsity requires sublinear growth in entropy with respect to $\log(t)$ (see Definition 3.3).

Definition 3.3 (Weak scale-invariant attention sparsity)**.** *Consider a set of random variables, $\{L_t\}$, representing attention logits. We say that the attention sparsity is weakly scale-invariant if, for any integer $\Delta > 1$,*

$$\mathrm{E}\left[H_t^{t\Delta}\right] = o(\log t) \quad \textit{as} \quad t \to \infty. \tag{10}$$

Remember $o(\log(t))$ is 'little-o' notation which means that $\mathrm{E}\left[H_t^{t\Delta}\right]$ scales strictly slower than $\log(t)$. Strong scale-invariant attention sparsity implies that the number of tokens attended to is asymptotically constant as we go from e.g. the past 10–100 tokens to the past 1,000–10,000.

Definition 3.4 (Strong scale-invariant attention sparsity)**.** *Consider a set of random variables, $\{L_t\}$, representing attention logits. We say that the attention sparsity is strongly scale-invariant if, for any integer $\Delta > 1$,*

$$\mathrm{E}\left[H_t^{t\Delta}\right] = \Theta(1) \quad as \quad t \to \infty, \tag{11}$$

i.e. we expect $H_t^{t\Delta}$ to be asymptotically constant as $t \to \infty$.

3.2 What characteristics of the logits are required for scale-invariant attention?

Next, we ask what properties would be sufficient for scale-invariant total attention and for some form of scale-invariant attention sparsity. Lemma 1 gives scale-invariant total attention, and Lemma 2 gives scale-invariant unnormalised attention (see Appendix D for the proofs).

Lemma 1. *Consider a set of random variables, $\{L_t\}$, representing attention logits. Let $\tau > 0$ be a lengthscale parameter, and $\alpha > 0$ a multiplicative constant. If the attention logits satisfy,*

$$\mathrm{E}\left[\tilde{A}_t\right] = \frac{\alpha}{t/\tau + 1}, \tag{12}$$

then we have scale-invariant total attention (Def. 3.1).

Lemma 2. *Consider a set of random variables, $\{L_t\}$, representing attention logits. Let $\tau > 0$ be a lengthscale parameter, and $\beta > 0$ a multiplicative constant. If the attention logits satisfy,*

$$\mathrm{E}\left[\tilde{A}_t \log \tilde{A}_t\right] = \frac{\beta}{t/\tau + 1}, \tag{13}$$

then we have scale-invariant unnormalised attention sparsity (Def. 3.2).

To construct an attention mechanism that satisfies Eq. (12) and Eq. (13), we consider a simplified setting in which the logits are marginally Gaussian and arise from taking Gaussian "base logits", $\bar{L}_t \sim \mathcal{N}(0, 1)$, and transforming them by multiplying by a_t and adding a bias, m_t,

$$L_t = a_t \bar{L}_t + m_t \sim \mathcal{N}(m_t, a_t^2). \tag{14}$$

Our goal is to find m_t and a_t^2 such that scale-invariant total attention and scale-invariant unnormalised attention sparsity hold. In particular that requires,

$$\frac{\alpha}{\frac{t}{\tau} + 1} = \mathrm{E}\left[\tilde{A}_t\right] = e^{m_t + a_t^2/2}, \tag{15a}$$

$$\frac{\beta}{\frac{t}{\tau} + 1} = \mathrm{E}\left[\tilde{A}_t \log \tilde{A}_t\right] = (m_t + a_t^2) e^{m_t + a_t^2/2}, \tag{15b}$$

where $\tilde{A}_t = \exp(L_t)$. Solving for m_t and a_t, we have (see Appendix C for details),

$$a_t = \sqrt{2\left[\log(t/\tau + 1) - \log\alpha + \beta/\alpha\right]}, \tag{16a}$$

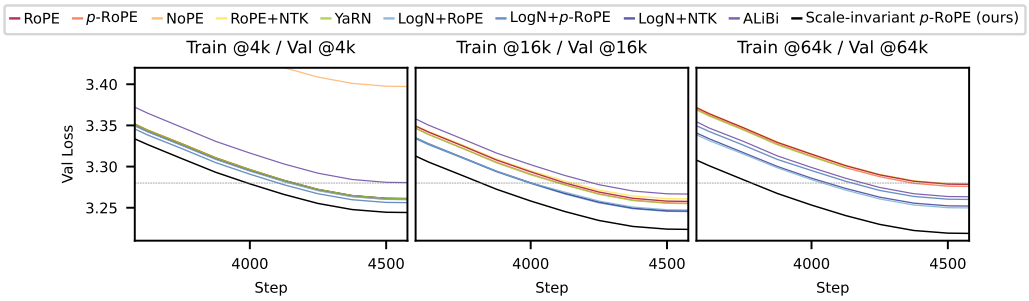
$$m_t = -a_t^2 + \beta/\alpha. \tag{16b}$$

For this solution to be valid, we only require $\beta \geq \alpha \log \alpha$ since $\log(t/\tau + 1) \geq 0$ when $t \geq 0$. We formally summarise the above results in Theorem 1 (proof in Appendix F), which tells us that this approach does indeed give scale-invariant total attention and scale-invariant unnormalised attention sparsity.

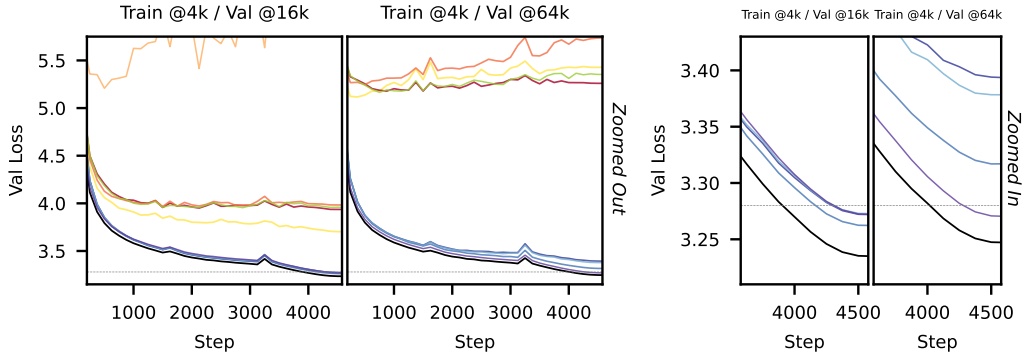
Theorem 1. *Suppose attention logits $\{L_t\}$ are marginally Gaussian with mean m_t and standard deviation a_t defined by Eq. (16). Assuming $\alpha, \beta, \tau > 0$, $\beta \geq \alpha \log \alpha$, then we have scale-invariant total attention and scale-invariant unnormalised attention sparsity.*

We therefore propose to scale the logits in real attention using a_t and m_t defined in Eq. (16). As t increases, the variance, a_t^2, *increases* as the logarithm of t, while the mean *decreases* as the logarithm of t.

Finally, note that while we have proven that we have scale-invariant unnormalised attention with this choice of a_t and m_t, we have not proven that we have strong or weak scale-invariant attention. We therefore checked empirically whether using IID Gaussian logits, scaled by a_t and m_t, gave weak or



(a) Scale-invariant attention improves language modelling over a range of training context lengths.



(b) Scale-invariant attention enables zero-shot long-context generalization of at least 16x.

Figure 3: Validation losses throughout training of a 162M parameter GPT-2-like model with different attention mechanisms (our scale-invariant scheme shown in black). NoPE is omitted in all but top-left and bottom-left panels to avoid excessive zooming out (due to high loss). The gray dashed line shows the baseline of 3.28. The validation loss for many methods increases in unison at steps ∼1500 and ∼3250 (see (b), left) despite aggregating over seeds; this is due to a fixed training and validation data ordering.

strong scale-invariant attention entropy (Fig. 2). We find that $H_t^{t\Delta}$ appears to scale with $\sqrt{\log(t)}$, and hence seems to satisfy weak, but not strong scale-invariant attention sparsity.

Hyperparameters. Introducing a_t and m_t (Eq. 16), appears to have introduced three additional hyperparameters to tune. We reduce this to one additional hyperparameter, the lengthscale τ, by specifying a boundary condition — $a_0^2 = 1$ and $m_0 = 0$. This boundary condition corresponds to not changing the scale of the local tokens. Substituting into Eq. (16b) gives $0 = -1 + \beta/\alpha$. Similarly Eq. (16a) requires that $1 = -2\log\alpha + 2\beta/\alpha$. Putting these together, we obtain $\alpha = \beta = \mathrm{e}^{0.5}$.

That leaves us with the lengthscale as the only hyperparameter. Note that when t is small relative to τ, neither a_t^2 nor m_t change much. Therefore, the timescale sets the size of a local region in which attention is approximately unscaled. Intuitively, it therefore makes sense to choose τ somewhere in the region of 1–100 tokens. In Appendix H we tried $\tau \in \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$ and find that $\tau = 10$ performs best in practice.

4 Experiments

In this section, we compare scale-invariant attention with other dense attention methods including Dynamic NTK interpolation (RoPE+NTK) (bloc97, 2023), LogN scaling/SSMax (Nakanishi, 2025), p-RoPE (Barbero et al., 2024b), and ALiBi (Press et al., 2021). Our results show that our method, scale-invariant p-RoPE, has uniformly lower validation loss at a variety of training lengths (4k, 16k, 64k). Additionally, scale-invariant p-RoPE demonstrates stronger zero-shot long-context generalization (e.g. in 'Train @4k/Val @16k' and 'Train @4k/Val @64k' settings) versus all other methods. Finally, scale-invariant p-RoPE, along with LogN, saturated"needle-in-a-haystack" task.

We pretrained GPT-2-style models (Radford et al., 2019) (with QK-norm, ReLU^2 activations, etc. (Jordan et al., 2024a)) from scratch on the FineWeb dataset (Penedo et al., 2024), using a fixed training data ordering and a 10M token validation set. We trained linear layers with Muon (Jordan et al., 2024b), and remaining parameters with Adam. We implemented scale-invariant attention using FlexAttention (Dong et al., 2024).

We trained two models: one with 162M parameters and another with 304M. The 162M one were trained for 4578 steps on 2.4B tokens over a range of context lengths (4k, 16k, 64k). The 304M parameter models were trained only on the best length-generalising methods, for 10.9k steps on 10B tokens, at 4k context length. We give further experimental details in Appendix G. For the 162M parameter model, we targeted a validation loss of 3.28, following Karpathy's GPT-2 reproduction (Karpathy, 2024; Jordan et al., 2024a), shown in the figures as a horizontal grey dashed line.

Long-context performance. We examined in-distribution, long-context performance of the different attention methods by looking at the validation loss for the same context length used for training. Even in this in-distribution setting, scale-invariant p-RoPE shows strong improvements in validation loss at all training lengths, 4k, 16k, and 64k (Fig. 3a).

Length Generalization. We evaluate length generalization by measuring validation loss on long sequences (16k and 64k) when training on 4k tokens. The 'Train @4k/Val @64k' setting in particular represents a considerable jump of 16× between train and validation. Table 1 and Fig. 3b report results for the 162M model. In the left panel of the Figure, ALiBi, LogN, and our method substantially outperform other approaches in generalizing to longer sequences. The right panel zooms in and shows that scale-invariant p-RoPE achieves the strongest generalization overall.

In preliminary experiments we tried other scale-invariant methods. We found that the most obvious method, scale-invariant RoPE, did not generalise well to long contexts (see Appendix I.3). The p-RoPE method is similar to RoPE but excludes low-frequency/high-wavelength components in the position embedding, possibly suggesting that low-frequency components in RoPE interfere with the scale-invariant transformation $L_t \mapsto a_t L_t + m_t$. LogN also demonstrates stronger performance when paired with p-RoPE rather than RoPE. We were surprised that RoPE+NTK struggled to generalise in the 'Train @4k / Val @64k' setting, but we believe this can be explained by the training context size: 'Train @16k / Val @64k' is much better for RoPE+NTK (see Fig. 7 in the Appendix).

Fig. 4 presents pretraining losses for ALiBi, LogN+p-RoPE, and our method on a larger 304M model — selected due to their performance on the 162M 'Train @4k/Val @64k' task. Scale-invariant p-RoPE maintains its advantage at this larger scale.

Needle in a Haystack. The key benefit of scale-invariant attention is that it balances local and sparse global attention. As such, we might worry that long-context retrieval performance might suffer versus other approaches that do not have specific mechanisms to ensure that attention to the local context does not vanish.

To assess whether long-context information retrieval capabilities suffered, we fine-tuned models on a needle-in-a-haystack task (note that fine-tuning on this task is unusual, but we found that prompting alone was not sufficient to perform this task, as bigger models/more pretraining would be required). Needle-in-a-haystack (Kamradt, 2023) measures a model's ability to precisely retrieve specific details (needles) from a large body of text (the haystack). Our tasks constructs prompts by concatenating text samples from the C4 dataset (Roberts et al., 2019) and embedding "needles" of the form '`The special magic <city> number is <7_digit_number>`'. We insert three (rather than one) needles uniformly, at random, into each context for more signal per example, with each needle contributing separately to the overall accuracy. A successful retrieval requires the model to output both the city and the associated number correctly. We trained models on sequences of length 4k, and tested at 4k, 16k, and 64k.

Table 2 show that scale-invariant p-RoPE and LogN+p-RoPE perform well, while almost all other methods fail almost completely at 64k context length. Thus, despite focusing more on local context, our method does not seem to have suffered in retrieval performance.

Table 1: Final mean validation losses (± 1 standard error across 3 seeds) for different methods when training with 4k context length on a 162M parameter GPT-2-style model. The error bars are small due to consistent training data ordering, and a fixed validation set.

Method	Val @ 4k	Val @ 16k	Val @ 64k
RoPE	3.261 ± 0.001	3.936 ± 0.010	5.260 ± 0.014
p-RoPE	3.260 ± 0.001	3.984 ± 0.008	5.735 ± 0.085
NoPE	3.397 ± 0.000	6.430 ± 0.059	8.125 ± 0.062
RoPE+NTK	3.261 ± 0.001	3.703 ± 0.007	5.430 ± 0.026
YaRN	3.261 ± 0.000	3.958 ± 0.018	5.353 ± 0.065
LogN+RoPE	3.260 ± 0.001	3.273 ± 0.004	3.378 ± 0.011
LogN+p-RoPE	3.256 ± 0.001	3.262 ± 0.002	3.317 ± 0.005
LogN+NTK	3.261 ± 0.001	3.272 ± 0.002	3.394 ± 0.025
ALiBi	3.281 ± 0.001	3.272 ± 0.001	3.270 ± 0.000
Scale-invariant p-RoPE (ours)	$\mathbf{3.244} \pm 0.001$	$\mathbf{3.235} \pm 0.001$	$\mathbf{3.247} \pm 0.001$

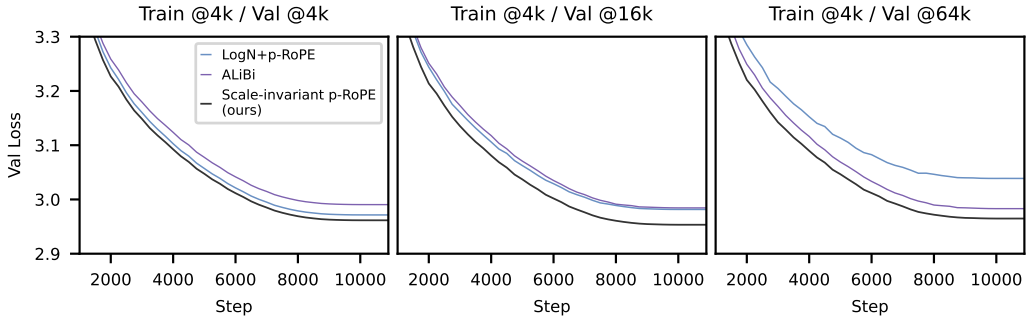


Figure 4: Validation losses throughout training for a 304M parameter model.

Table 2: Mean validation accuracies on the needle-in-a-haystack task, ± 1 standard error. Metrics were calculated over 3 seeds, after 300 steps of fine-tuning.

Method	Val Acc @4k	Val Acc @16k	Val Acc @64k
RoPE	0.962 ± 0.003	0.000 ± 0.000	0.000 ± 0.000
p-RoPE	0.966 ± 0.001	0.250 ± 0.020	0.000 ± 0.000
NoPE	0.964 ± 0.000	0.303 ± 0.239	0.000 ± 0.000
RoPE+NTK	0.962 ± 0.001	0.217 ± 0.078	0.000 ± 0.000
YaRN	$\mathbf{0.969} \pm 0.001$	0.000 ± 0.000	0.000 ± 0.000
LogN+RoPE	0.965 ± 0.002	0.276 ± 0.010	0.064 ± 0.006
LogN+p-RoPE	$\mathbf{0.969} \pm 0.002$	0.962 ± 0.003	0.939 ± 0.009
LogN+NTK	0.962 ± 0.001	0.253 ± 0.015	0.056 ± 0.008
ALiBi	0.957 ± 0.002	0.020 ± 0.001	0.003 ± 0.000
Scale-invariant p-RoPE (ours)	0.965 ± 0.000	$\mathbf{0.969} \pm 0.004$	$\mathbf{0.969} \pm 0.005$

5 Limitations

In this work, we evaluated our methods by pretraining with 162M and 304M parameter models, and investigated 7B parameter models via continual pretraining in Appendix I.5. Ideally, we would have pretrained from scratch at the multi-billion parameter scale, in line with contemporary state-of-the-art LLMs. While compute resource constraints prevented this, our results, together with the natural theoretical approach, offer no indication that our conclusions would fail to generalise to larger, commercially deployed models.

We focused on scale-invariant p-RoPE with dense attention, but the extension to other settings is a promising direction for future investigation.

6 Conclusions

We have proposed two desirable properties of attention mechanisms, "scale-invariant total attention" and "scale-invariant attention sparsity", and presented a straightforward modification to attention logits that enables these properties in practice. In our experiments, we found that our scale-invariant attention modification, especially when paired with p-RoPE, substantially improves long-context language modelling performance and gives zero-shot generalisation from training on short contexts to testing on long contexts.

7 Acknowledgements

We sincerely thank the Engineering and Physical Sciences Research Council and the COMPASS CDT for funding Ben Anson at the University of Bristol. Many of our experiments used computational resources at the Advanced Computing Research Centre at the University of Bristol, with GPUs generously funded by Dr. Stewart.

References

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Bai, Y., Lv, X., Zhang, J., He, Y., Qi, J., Hou, L., Tang, J., Dong, Y., and Li, J. Longalign: A recipe for long context alignment of large language models. *arXiv preprint arXiv:2401.18058*, 2024.

Barbero, F., Banino, A., Kapturowski, S., Kumaran, D., Madeira Araújo, J., Vitvitskyi, O., Pascanu, R., and Veličković, P. Transformers need glasses! information over-squashing in language tasks. *Advances in Neural Information Processing Systems*, 37:98111–98142, 2024a.

Barbero, F., Vitvitskyi, A., Perivolaropoulos, C., Pascanu, R., and Veličković, P. Round and round we go! what makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*, 2024b.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

bloc97. Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation. Reddit, 2023. URL https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/.

Bulatov, A., Kuratov, Y., Kapushev, Y., and Burtsev, M. Beyond attention: breaking the limits of transformer context length with recurrent memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Chen, J., Peng, S., Luo, D., Yang, F., Wu, R., Li, F., and Chen, X. Edgeinfinite: A memory-efficient infinite-context transformer for edge devices. *arXiv preprint arXiv:2503.22196*, 2025.

Chen, S., Wong, S., Chen, L., and Tian, Y. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Chiang, D. and Cholak, P. Overcoming a theoretical limitation of self-attention. *arXiv preprint arXiv:2202.12172*, 2022.

Cohere Team. Command a: An enterprise-ready large language model. *arXiv preprint arXiv:2504.00698*, 2025.

Ding, J., Ma, S., Dong, L., Zhang, X., Huang, S., Wang, W., Zheng, N., and Wei, F. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023.

Dong, J., Feng, B., Guessous, D., Liang, Y., and He, H. Flex attention: A programming model for generating optimized attention kernels. *arXiv preprint arXiv:2412.05496*, 2024.

Fountas, Z., Benfeghoul, M. A., Oomerjee, A., Christopoulou, F., Lampouras, G., Bou-Ammar, H., and Wang, J. Human-like episodic memory for infinite context llms. *arXiv preprint arXiv:2407.09450*, 2024.

Gao, T., Wettig, A., Yen, H., and Chen, D. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*, 2024.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Gu, X., Pang, T., Du, C., Liu, Q., Zhang, F., Du, C., Wang, Y., and Lin, M. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781*, 2024.

Han, C., Wang, Q., Peng, H., Xiong, W., Chen, Y., Ji, H., and Wang, S. Lm-infinite: Zero-shot extreme length generalization for large language models. *arXiv preprint arXiv:2308.16137*, 2023.

Hu, Z., Liu, Y., Zhao, J., Wang, S., Wang, Y., Shen, W., Gu, Q., Luu, A. T., Ng, S.-K., Jiang, Z., et al. Longrecipe: Recipe for efficient long context generalization in large language models. *arXiv preprint arXiv:2409.00509*, 2024.

Jianlin, S. Attention's scale operation from the perspective of entropy invariance, 2021. URL https://www.kexue.fm/archives/8823.

Jin, H., Han, X., Yang, J., Jiang, Z., Liu, Z., Chang, C.-Y., Chen, H., and Hu, X. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*, 2024.

Jordan, K., Bernstein, J., Rappazzo, B., @fernbear.bsky.social, Vlado, B., Jiacheng, Y., Cesista, F., Koszarsky, B., and @Grad62304977. modded-nanogpt: Speedrunning the nanogpt baseline, 2024a. URL https://github.com/KellerJordan/modded-nanogpt.

Jordan, K., Jin, Y., Boza, V., Jiacheng, Y., Cesista, F., Newhouse, L., and Bernstein, J. Muon: An optimizer for hidden layers in neural networks, 2024b. URL https://kellerjordan.github.io/posts/muon/.

Kamradt, G. Needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023.

Karpathy, A. Discussion #481. https://github.com/karpathy/llm.c/discussions/481, 2024.

Kazemnejad, A., Padhi, I., Natesan Ramamurthy, K., Das, P., and Reddy, S. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36:24892–24928, 2023.

Li, K., Kong, Y., Xu, Y., Su, J., Huang, L., Zhang, R., and Zhou, F. Information entropy invariance: Enhancing length extrapolation in attention mechanisms. *arXiv preprint arXiv:2501.08570*, 2025.

Lieber, O., Lenz, B., Bata, H., Cohen, G., Osin, J., Dalmedigos, I., Safahi, E., Meirom, S., Belinkov, Y., Shalev-Shwartz, S., et al. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Liu, H., Zaharia, M., and Abbeel, P. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023a.

Liu, X., Yan, H., Zhang, S., An, C., Qiu, X., and Lin, D. Scaling laws of rope-based extrapolation. *arXiv preprint arXiv:2310.05209*, 2023b.

Liu, X., Li, R., Guo, Q., Liu, Z., Song, Y., Lv, K., Yan, H., Li, L., Liu, Q., and Qiu, X. Reattention: Training-free infinite context with finite attention scope. *arXiv preprint arXiv:2407.15176*, 2024b.

Llama 4 Team. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025. Accessed: 2025-04-08.

Martins, P. H., Marinho, Z., and Martins, A. F. ∞-former: Infinite Memory Transformer. *arXiv preprint arXiv:2109.00301*, 2021.

Munkhdalai, T., Faruqui, M., and Gopal, S. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*, 101, 2024.

Nakanishi, K. M. Scalable-softmax is superior for attention. *arXiv preprint arXiv:2501.19399*, 2025.

Penedo, G., Kydlíček, H., Lozhkov, A., Mitchell, M., Raffel, C. A., Von Werra, L., Wolf, T., et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Peng, B., Quesnelle, J., Fan, H., and Shippole, E. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

Press, O., Smith, N. A., and Lewis, M. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Roberts, A., Raffel, C., Lee, K., Matena, M., Shazeer, N., Liu, P. J., Narang, S., Li, W., and Zhou, Y. Exploring the limits of transfer learning with a unified text-to-text transformer. *Google Research*, 2019.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Torchtune. Torchtune: Pytorch's finetuning library, April 2024. URL https://github.com/pytorch/torchtune.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Van der Schaaf, v. A. and van Hateren, J. v. Modelling the power spectra of natural images: statistics and information. *Vision research*, 36(17):2759–2770, 1996.

Wang, S., Kobyzev, I., Lu, P., Rezagholizadeh, M., and Liu, B. Resonance rope: Improving context length generalization of large language models. *arXiv preprint arXiv:2403.00071*, 2024.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. 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, 2020.

Wortsman, M., Liu, P. J., Xiao, L., Everett, K., Alemi, A., Adlam, B., Co-Reyes, J. D., Gur, I., Kumar, A., Novak, R., et al. Small-scale proxies for large-scale transformer training instabilities. *arXiv preprint arXiv:2309.14322*, 2023.

Xiao, C., Zhang, P., Han, X., Xiao, G., Lin, Y., Zhang, Z., Liu, Z., Han, S., and Sun, M. Infllm: Unveiling the intrinsic capacity of llms for understanding extremely long sequences with training-free memory. *arXiv e-prints*, pp. arXiv–2402, 2024.

Xiong, W., Liu, J., Molybog, I., Zhang, H., Bhargava, P., Hou, R., Martin, L., Rungta, R., Sankararaman, K. A., Oguz, B., et al. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*, 2023.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., Zhang, J., Huang, F., and Zhou, J. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

Zhu, D., Yang, N., Wang, L., Song, Y., Wu, W., Wei, F., and Li, S. Pose: Efficient context window extension of llms via positional skip-wise training. *arXiv preprint arXiv:2309.10400*, 2023.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our contributions of the notion of scale-invariant attention, a method for implementing it in practice, alongside empirical evidence that it is effective.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: See our Limitations section (Sec. 5).

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All of our theoretical results (Lemmas 1, and 2, and Theorem 1) all clearly state assumptions, and we provide proofs and auxiliary results in Appendices B, C, D, E, F.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We give details of the experiments in Section 4, and in Appendix G, which enables reproducibility.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We provide code in the supplementary materials along with instructions to reproduce the main results.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: We provide extensive details in Section 4 and Appendix G.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: We provide mean validation results $\pm$ 1 stderr for the results in Table 1 and Table 2, clearly stating the number of seeds. We do not provide error bars for the plots as that makes them too messy to be easily interpreted.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We discuss resources required in Appendix G, under Section G.3.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics, and the research conducted in the paper conforms in every respect.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [NA]

    Justification: The paper contains foundational research pertaining to the attention mechanism in neural networks. While this has the potential to improve the LLMs in the long-term, the guidelines suggest it is not necessary to speculate on such matters.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: The paper does not release any artifacts at high risk for misuse.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We properly credit and reference all assets used in the reference section. We declare all licenses in the Appendix.

    Guidelines:

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

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

13. **New assets**

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

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: We do not crowdsource or perform research with human subjects.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: We do not crowdsource or perform research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: LLMs were not used for the core method development in this research, but they were used for visualizing data (e.g. creating tables/figures).

    Guidelines:

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

A Relating the Indicies in the Paper to Standard Indices

The standard form for attention is,

$$S^{ij} = \frac{1}{\sqrt{d}} \sum_{\lambda=1}^{d} Q_{i\lambda} K_{j\lambda}, \tag{17}$$

where λ indexes the feature, i indexes the query (i.e. the token we're generating now) and j indexes the key (i.e. the token we're attending to). We fix i, and take $t = i - j$, so

$$S^{ij} = S_{t=i-j} \tag{18}$$

where $S_{t=i-j}$ is given in Eq. 1 in the main text. Then,

$$L^{ij} = L(S^{ij}; i-j) = L(S_t; t) = L_{t=i-j}, \tag{19}$$

where $L_{t=i-j}$ is given in Eq. 1 in the main text. Then the unnormalised attention weights, $\tilde{A}^{ij}$, normalised attention weights, A^{ij} and normalisers, Z^i are,

$$\tilde{A}^{ij} = \exp\left(L^{ij}\right), \tag{20}$$

where, $\tilde{A}_{t=i-j}$ is given in Eq. 3 in the main text.

$$A^{ij} = \frac{\tilde{A}^{ij}}{Z^i}, \tag{21}$$

where, $A_{t=i-j}$ is given in Eq. 3 in the main text.

$$Z^i = \sum_{j=1}^{i} \tilde{A}^{ij} = Z, \tag{22}$$

where, Z is given in Eq. 3 in the main text.

B $\mathrm{E}\left[X^k \exp(\alpha X)\right]$ where X is Gaussian

Assume $X \sim \mathcal{N}(\mu, \sigma^2)$, then we can compute the expectation in terms of a moment of a Gaussian,

$$\begin{aligned}
\mathbb{E}[X^k e^{\alpha X}] &= \int_{-\infty}^{\infty} x^k e^{\alpha x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx && (23)\\
&= \int_{-\infty}^{\infty} x^k \frac{1}{\sqrt{2\pi\sigma^2}} e^{\alpha x - \frac{(x-\mu)^2}{2\sigma^2}} dx && (24)\\
&= \int_{-\infty}^{\infty} x^k \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x^2 - 2\mu x + \mu^2 - 2\alpha\sigma^2 x)} dx && (25)\\
&= \int_{-\infty}^{\infty} x^k \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x^2 - 2(\mu+\alpha\sigma^2)x + \mu^2)} dx && (26)\\
&= e^{\frac{\alpha^2\sigma^2}{2} + \alpha\mu} \int_{-\infty}^{\infty} x^k \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-(\mu+\alpha\sigma^2))^2}{2\sigma^2}} dx && (27)\\
&= e^{\frac{\alpha^2\sigma^2}{2} + \alpha\mu} \mathrm{E}\left[\tilde{X}^k\right], && (28)
\end{aligned}$$

where $\tilde{X} \sim \mathcal{N}(\mu', \sigma'^2) = \mathcal{N}(\mu + \alpha\sigma^2, \sigma^2)$. In the case $\alpha = k = 1$, we have,

$$\mathrm{E}\left[X \exp(X)\right] = (\mu + \sigma^2) \exp(\mu + \sigma^2/2). \tag{29}$$

In the case $\alpha = k = 2$, we have,

$$\mathrm{E}\left[X^2 \exp(2X)\right] = ((\mu + 2\sigma^2)^2 + \sigma^2) \exp(2\mu + 2\sigma^2). \tag{30}$$

C Deriving a_t and m_t in the Logit Transformation

From Eq. (15), we wish to find a_t and m_t such that,

$$\frac{\alpha}{\frac{t}{\tau}+1} = e^{m_t + a_t^2/2} \tag{31}$$

$$\frac{\beta}{\frac{t}{\tau}+1} = (m_t + a_t^2)e^{m_t + a_t^2/2}. \tag{32}$$

We begin by dividing Eq. (32) by Eq. (31),

$$\frac{\beta}{\alpha} = m_t + a_t^2, \tag{33}$$

which can be rearranged to,

$$m_t = \frac{\beta}{\alpha} - a_t^2, \tag{34}$$

$$\text{and } a_t^2 = \frac{\beta}{\alpha} - m_t. \tag{35}$$

Now, taking the log of Eq. (31), we have,

$$m_t + \tfrac{1}{2}a_t^2 = -\log\left(\tfrac{t}{\tau}+1\right) + \log\alpha. \tag{36}$$

Define $f_t = \log\left(\frac{t}{\tau}+1\right) - \log\alpha$. Then to solve for a_t^2, we substitute m_t from Eq. (34) into Eq. (36),

$$\left(\tfrac{\beta}{\alpha} - a_t^2\right) + \tfrac{1}{2}a_t^2 = -f_t \tag{37}$$

$$\tfrac{\beta}{\alpha} - \tfrac{1}{2}a_t^2 = -f_t \tag{38}$$

$$a_t^2 = 2\left(f_t + \tfrac{\beta}{\alpha}\right). \tag{39}$$

To solve for m_t, we substitute a_t^2 from Eq. (35) into Eq. (36)

$$m_t + \tfrac{1}{2}\left(\tfrac{\beta}{\alpha} - m_t\right) = -f_t \tag{40}$$

$$\tfrac{1}{2}m_t + \tfrac{\beta}{2\alpha} = -f_t \tag{41}$$

$$m_t = -2\left(f_t + \tfrac{\beta}{2\alpha}\right), \tag{42}$$

which can also be written as,

$$m_t = -a_t^2 + \tfrac{\beta}{\alpha}. \tag{43}$$

Thus, we have,

$$a_t^2 = 2\left[\log(t/\tau+1) + \beta/\alpha - \log\alpha\right] \tag{44}$$

$$m_t = -2\left[\log(t/\tau+1) + \beta/\alpha - \log\alpha\right] + \beta/\alpha \tag{45}$$

$$= -2\log(t/\tau+1) - \beta/\alpha + 2\log\alpha. \tag{46}$$

D Proofs of Lemmas 1 and 2

Lemma 1. *Consider a set of random variables, $\{L_t\}$, representing attention logits. Let $\tau > 0$ be a lengthscale parameter, and $\alpha > 0$ a multiplicative constant. If the attention logits satisfy,*

$$\mathrm{E}\left[\tilde{A}_t\right] = \frac{\alpha}{t/\tau+1}, \tag{12}$$

then we have scale-invariant total attention (Def. 3.1).

Proof. We want to show that $\mathrm{E}\left[Z_{t_1}^{t_1\Delta}\right] = \Theta(1)$ as $t_1 \to \infty$. By definition and linearity of expectation:

$$\mathrm{E}\left[Z_{t_1}^{t_1\Delta}\right] = \mathrm{E}\left[\sum_{t=t_1}^{t_1\Delta-1} \tilde{A}_t\right] = \sum_{t=t_1}^{t_1\Delta-1} \mathrm{E}\left[\tilde{A}_t\right]. \tag{47}$$

Using the given condition $\mathrm{E}\left[\tilde{A}_t\right] = \frac{\alpha}{t/\tau+1}$:

$$\mathrm{E}\left[Z_{t_1}^{t_1\Delta}\right] = \sum_{t=t_1}^{t_1\Delta-1} \frac{\alpha}{t/\tau+1} = \alpha\tau \sum_{t=t_1}^{t_1\Delta-1} \frac{1}{t+\tau} = \alpha\tau \sum_{k=t_1+\tau}^{t_1\Delta-1+\tau} \frac{1}{k}. \tag{48}$$

Using the standard integral bounds for the harmonic sum derived by comparison with the integral (see Appendix E):

$$\ln\left(\frac{t_1\Delta+\tau}{t_1+\tau}\right) \leq \sum_{k=t_1+\tau}^{t_1\Delta-1+\tau} \frac{1}{k} \leq \ln\left(\frac{t_1\Delta-1+\tau}{t_1+\tau-1}\right). \tag{49}$$

As $t_1 \to \infty$:

$$\frac{t_1\Delta+\tau}{t_1+\tau} \to \frac{t_1\Delta}{t_1} = \Delta \tag{50}$$

$$\frac{t_1\Delta-1+\tau}{t_1+\tau-1} \to \frac{t_1\Delta}{t_1} = \Delta \tag{51}$$

So, the logarithm terms in both the lower and upper bounds approach $\ln(\Delta)$. By the Squeeze Theorem:

$$\sum_{k=t_1+\tau}^{t_1\Delta-1+\tau} \frac{1}{k} \to \ln(\Delta) \tag{52}$$

Since α, τ, and $\Delta > 1$ are constants, $\ln(\Delta)$ is a positive constant. Thus:

$$\mathrm{E}\left[Z_{t_1}^{t_1\Delta}\right] \to \alpha\tau\ln(\Delta) = \Theta(1) \tag{53}$$

This satisfies the condition for scale-invariant total attention in expectation. □

Lemma 2. *Consider a set of random variables, $\{L_t\}$, representing attention logits. Let $\tau > 0$ be a lengthscale parameter, and $\beta > 0$ a multiplicative constant. If the attention logits satisfy,*

$$\mathrm{E}\left[\tilde{A}_t \log \tilde{A}_t\right] = \frac{\beta}{t/\tau+1}, \tag{13}$$

then we have scale-invariant unnormalised attention sparsity (Def. 3.2).

Proof. We want to show that $\mathrm{E}\left[\tilde{N}_{t_1}^{t_1\Delta}\right] = \Theta(1)$ as $t_1 \to \infty$. By definition and linearity of expectation:

$$\mathrm{E}\left[\tilde{N}_{t_1}^{t_1\Delta}\right] = \mathrm{E}\left[\sum_{t=t_1}^{t_1\Delta-1} \tilde{A}_t \log \tilde{A}_t\right] = \sum_{t=t_1}^{t_1\Delta-1} \mathrm{E}\left[\tilde{A}_t \log \tilde{A}_t\right] \tag{54}$$

Using the given condition $\mathrm{E}\left[\tilde{A}_t \log \tilde{A}_t\right] = \frac{\beta}{t/\tau+1}$:

$$\mathrm{E}\left[\tilde{N}_{t_1}^{t_1\Delta}\right] = \sum_{t=t_1}^{t_1\Delta-1} \frac{\beta}{t/\tau+1} = \beta\tau \sum_{t=t_1}^{t_1\Delta-1} \frac{1}{t+\tau} \tag{55}$$

This sum is exactly the same form as in the proof of Lemma 1, just with β instead of α. Following the same steps using the integral bounds for the harmonic sum derived in Appendix E:

$$\sum_{k=t_1+\tau}^{t_1\Delta-1+\tau} \frac{1}{k} \to \ln(\Delta) \quad \text{as} \quad t_1 \to \infty \tag{56}$$

Therefore, as $t_1 \to \infty$:

$$\mathrm{E}\left[\tilde{N}_{t_1}^{t_1\Delta}\right] \to \beta\tau\ln(\Delta) = \Theta(1) \tag{57}$$

This satisfies the condition for scale-invariant unnormalised attention sparsity in expectation. □

E Bounds for Harmonic Sums

For the proofs of Lemma 1 and Lemma 2, we need bounds on the partial harmonic sum $S = \sum_{k=a}^{b} \frac{1}{k}$, where $a = t_1 + \tau$ and $b = t_1\Delta - 1 + \tau$. We can obtain these bounds by comparing the sum to the integral of $f(x) = 1/x$.

Since $f(x) = 1/x$ is a decreasing function for $x > 0$, we have,

$$\int_a^{b+1} \frac{1}{x} dx \leq \sum_{k=a}^{b} \frac{1}{k} \leq \int_{a-1}^{b} \frac{1}{x} dx \tag{58}$$

Evaluating the integrals gives,

$$\ln\left(\frac{b+1}{a}\right) \leq \sum_{k=a}^{b} \frac{1}{k} \leq \ln\left(\frac{b}{a-1}\right). \tag{59}$$

F Proof of Theorem 1

Theorem 1. *Suppose attention logits $\{L_t\}$ are marginally Gaussian with mean m_t and standard deviation a_t defined by Eq. (16). Assuming $\alpha, \beta, \tau > 0$, $\beta \geq \alpha \log \alpha$, then we have scale-invariant total attention and scale-invariant unnormalised attention sparsity.*

Proof. We need to show that the given conditions are sufficient for scale-invariant total attention (Definition 3.1) and scale-invariant unnormalised attention sparsity (Definition 3.2).

1. Scale-invariant total attention: We need to show that $\mathrm{E}\left[Z_{t_1}^{t_1\Delta}\right] = \Theta(1)$ as $t_1 \to \infty$. By Lemma 1, this holds if $\mathrm{E}\left[\tilde{A}_t\right] = \frac{\alpha}{t/\tau+1}$. Since $L_t \sim \mathcal{N}(m_t, a_t^2)$, the unnormalised attention weight $\tilde{A}_t = e^{L_t}$ follows a log-normal distribution. The expectation of a log-normal variable e^X where $X \sim \mathcal{N}(\mu, \sigma^2)$ is $e^{\mu+\sigma^2/2}$. Therefore,

$$\mathrm{E}\left[\tilde{A}_t\right] = \mathrm{E}\left[e^{L_t}\right] = e^{m_t + a_t^2/2} \tag{60}$$

Substituting the given expressions for $m_t = -a_t^2 + \beta/\alpha$ and $a_t^2 = 2[\log(t/\tau + 1) - \log\alpha + \beta/\alpha]$:

$$\begin{aligned}
\mathrm{E}\left[\tilde{A}_t\right] &= e^{(-a_t^2+\beta/\alpha)+a_t^2/2} && (61)\\
&= e^{-a_t^2/2+\beta/\alpha} && (62)\\
&= e^{-[\log(t/\tau+1)-\log\alpha+\beta/\alpha]+\beta/\alpha} && (63)\\
&= e^{-\log(t/\tau+1)+\log\alpha-\beta/\alpha+\beta/\alpha} && (64)\\
&= e^{\log\alpha-\log(t/\tau+1)} && (65)\\
&= e^{\log\left(\frac{\alpha}{t/\tau+1}\right)} && (66)\\
&= \frac{\alpha}{t/\tau+1} && (67)
\end{aligned}$$

Since this condition matches the requirement of Lemma 1, scale-invariant total attention holds in expectation. The condition $\beta \geq \alpha \log \alpha$ ensures $a_t^2 \geq 0$ for all $t \geq 0$.

2. Scale-invariant unnormalised attention sparsity: We need to show that $\mathrm{E}\left[\tilde{N}_{t_1}^{t_1\Delta}\right] = \Theta(1)$ as $t_1 \to \infty$. By Lemma 2, this holds if $\mathrm{E}\left[\tilde{A}_t \log \tilde{A}_t\right] = \frac{\beta}{t/\tau+1}$. We need to compute $\mathrm{E}\left[\tilde{A}_t \log \tilde{A}_t\right] = \mathrm{E}\left[L_t e^{L_t}\right]$. Using the formula for $\mathrm{E}\left[Xe^X\right]$ where $X \sim \mathcal{N}(\mu, \sigma^2)$ from Appendix B, which is $(\mu + \sigma^2)e^{\mu+\sigma^2/2}$:

$$\mathrm{E}\left[L_t e^{L_t}\right] = (m_t + a_t^2)e^{m_t + a_t^2/2} \tag{68}$$

Substitute $m_t = -a_t^2 + \beta/\alpha$:

$$\mathrm{E}\left[L_t e^{L_t}\right] = (-a_t^2 + \beta/\alpha + a_t^2)e^{m_t + a_t^2/2} \tag{69}$$

$$= (\beta/\alpha)e^{m_t + a_t^2/2} \tag{70}$$

From the previous step, we know $e^{m_t + a_t^2/2} = \frac{\alpha}{t/\tau + 1}$. Substituting this:

$$\mathrm{E}\left[L_t e^{L_t}\right] = (\beta/\alpha)\left(\frac{\alpha}{t/\tau + 1}\right) \tag{71}$$

$$= \frac{\beta}{t/\tau + 1} \tag{72}$$

Since this condition matches the requirement of Lemma 2, scale-invariant unnormalised attention sparsity holds.

Therefore, under the given conditions, we have both scale-invariant total attention and scale-invariant unnormalised attention sparsity. □

G Further experimental details

We provide experimental details here, and we provide the code used in the supplementary materials.

G.1 Pretraining from scratch

Our base model is a `modded-nanogpt` (Jordan et al., 2024a) variant, which is similar to GPT-2 (Radford et al., 2019), but has the following main differences from GPT-2: RMSNorm for layer normalization (applied before the attention and MLP blocks, as well as to the token embeddings and the final output layer), squared ReLU activations, and QK Normalization (RMSNorm on query and key projections). All models used a vocabulary size of 50304, with text tokenized with the GPT-2 tokenizer (Wolf et al., 2020). We implemented scale-invariant attention and ALiBi using FlexAttention (Dong et al., 2024).

162M parameter model. The smaller model had 12 layers, 768 hidden dimension, and 6 heads. We optimized embedding parameters using Adam with learning rate $\gamma = 0.3$, $\beta = (0.9, 0.95)$. We optimized linear layers with Muon, with no weight decay, $\gamma = 0.02$ and momentum 0.95. For remaining parameters (unembedding and LogN scalings, if LogN trick was active), we optimized with Adam, using $\gamma = 0.002$, $\beta = (0.9, 0.95)$. We trained for 4578 steps. The batch size was $8 \times 65536/L_{\text{tr}}$, with more gradient accumulation for shorter training lengths. We scheduled the learning rate with a linear schedule for all parameters, with no warmup, constant learning rate for 3270 steps, and linear cooldown for the remaining 1308 steps. We vary the training context length when pretraining with this model, from 4096(4k), to 16384(16k), and 65536(64k). We validate at 4k, 16k, and 64k sequence lengths every 125 steps.

304M parameter model. For the larger model, we used the same settings as the smaller model, but with the following changes. The larger model had 16 layers, 1024 hidden dimension, and 8 heads. We trained the model for 10900 steps, processing approximately 10B tokens. We trained for 2 accumulation steps on 4 GPUs, with 28 sequences per batch. We scaled learning rates following μParam (Wortsman et al., 2023), which involves multiplying learning rates of the linear layers by $768/1024$, to adjust for changing the model width. Learning rates of other layers (embedding, unembedding, and LogN scalings if active) were not changed. We used a cosine learning rate schedule, with no warmup, and a minimum learning rate of 0 (at the end of training). All runs on the 304M parameter models were with 4096 training sequence length. We validated at 4k, 16k, 64k sequence lengths every 250 steps.

RoPE hyperparameters. We used a base θ of $10,000$ for RoPE, and an effective base of 1024 for the angular frequencies in p-RoPE. For RoPE+NTK scaling (note the scaling applies only when the inference sequence length is longer than the training sequence length), we scaled θ by $(L_{\text{inf}}/L_{\text{tr}})^{d/(d-2)}$, where L_{tr} is the training sequence length, L_{inf} is the inference sequence length, and d is the RoPE head dimension (128 for both models).

Dataset. We used subsets of FineWeb (Penedo et al., 2024) for pretraining from scratch: a 10B token subset for the 162M model, and a 100B token subset for the 304M model (from which $\sim$10B tokens were used for its training). We keep the training data ordering fixed, and we keep the validation set identical across all runs to reduce variance.

Seeds. We repeat the experiment for 3 different seeds when training at 4k on the 162M parameter model. We train with 1 seed at 4k on the 304M parameter model, and at 16k/64k on the 304M parameter model due to compute limitations.

Compute. We trained the smaller (162M) models on single A100 80G GPUs. We trained the 304M models on 4xH100 grace hopper nodes using distributed data parallelism.

G.2 Needle-in-a-haystack

The Needle in a Haystack experiments were conducted by fine-tuning the pre-trained 162M parameter models. We fine-tune with the same learning rate as above, but with 100 warmup, 100 constant, and 100 warmdown steps (decaying to $\gamma = 0$). We use the same optimizers as in the pretraining phase. We train for 300 steps on sequences of length 4096, and batch size 8 with 8 accumulation steps. We repeated with 3 seeds.

The task is to generate responses of the form "`city1=needle1;city2=needle2;city3=needle3`", where the cities and needles are embedded uniformly at random into samples from C4 in the form "`The special magic <city> number is <7_digit_number>.`". We sample several times from the C4 dataset (concatenating samples) and remove tokens until we have the necessary number of tokens — 4096(4k) for training, and 4096(4k), 16384(16k), 65536(64k) for validation. Only the expected response tokens are included in the loss (the prompt/context tokens are masked). We repeat the task for three different seeds. The validation accuracies presenting in Table 1 are calculated by the proportion of times that cities *and* numbers are output correctly.

G.3 Resources required to reproduce experiments

To reproduce results, 80G GPUs are required. We used 80G A100s, and 80G H100 grace hopper nodes. In terms of time taken to execute each experiment type, we give estimates of the resources required:

- pretraining 162M parameter model w/o flex attention takes roughly 4/8/22 A100 hours at 4k/16k/64k;
- pretraining 162M parameter model with flex attention takes roughly 8/16/44 A100 hours at 4k/16k/64k;
- pretraining 304M parameter model takes roughly 36 4xH100 node hours;
- fine-tuning on needle-in-a-haystack task takes roughly 3 A100 hours.

To obtain our results, each experiment is executed several times. In particular, needle-in-a-haystack with 3 seeds per method, pretraining at 4k with 162M model is 3 seeds per positional encoding method. The remaining experiments are executed once per positional encoding method. This gives a total of $\sim$ 100 4xH100 hours, $\sim$ 550 1xA100 hours. We estimate that very roughly that amount again was spent configuring the experiments correctly, and on preliminary/failed experiments.

H The optimal lengthscale, τ, is around 10

By introducing scale-invariant attention, we introduce one extra lengthscale hyperparameter, τ, which represents the size of the 'chunks' we attend over. To select τ, we trained at 4k for $\tau \in \{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$, and compared validation losses (shown in Fig. 5). Validation performance is very similar amongst different τ at the training context length (though $\tau = 10$ is strictly best), but as we extend to out of distribution context lengths (64k, 16$\times$ the training context length) the benefit of $\tau = 10$ becomes clearer.

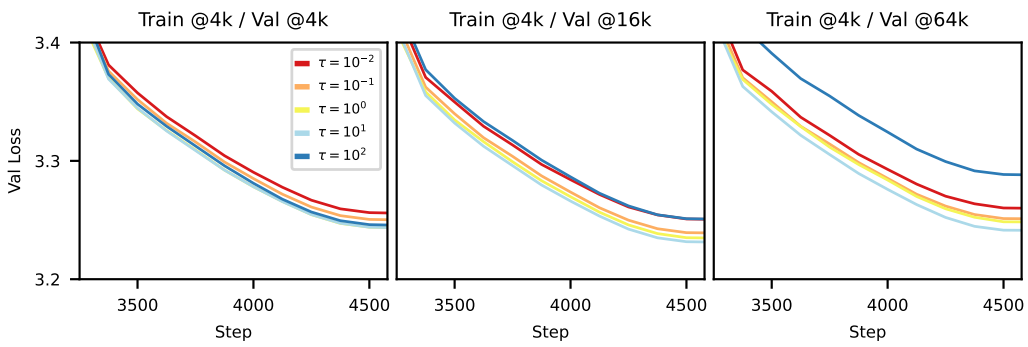


Figure 5: Validation losses at 4k (left), 16k (middle), and 64k (right) context lengths, for a GPT-2-like model trained with scale-invariant attention for varying τ. The models were trained at 4k context length.

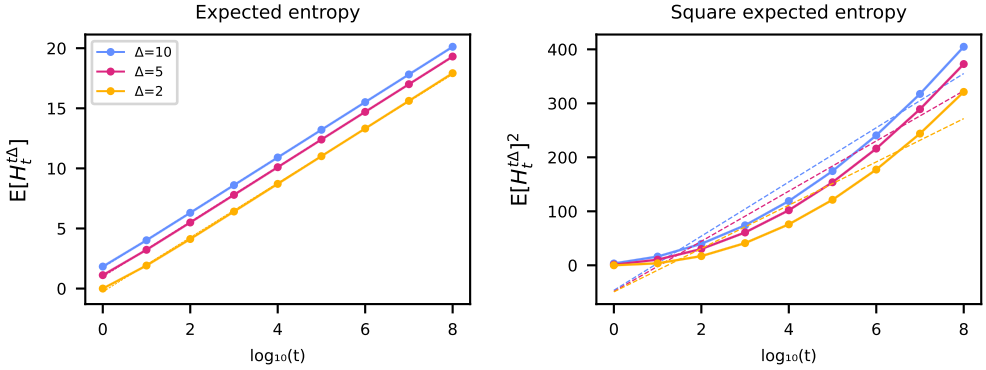


Figure 6: Scaling of expected entropy-in-range for standard attention with an independent Gaussian assumption on the logits. We calculate expected entropies in the range $[t, t\Delta]$ for different t and $\Delta \in \{2, 5, 10\}$, with the lengthscale τ set to 10. The dashed lines show the best linear fit of the data. We empirically find that standard Gaussian logits give logarithmic entropy (left panel).

I Extra Results

I.1 Entropy scaling of regular attention

In the main text, we illustrated in Fig. 2 that scale-invariant attention method under a Gaussian assumption has sub-logarithmic expected entropy. Fig. 6 empirically shows that expected entropy in a range t to $t\Delta$ for standard/unscaled attention instead scales logarithmically with t.

I.2 Pretrain @16k

For completeness, we include validation losses when training at 16k in Fig. 7. We see again that scale-invariant p-RoPE outperforms other methods over a range of validation lengths, with the improvements becoming more noticable at the longest validation length of 64k.

I.3 Alternative scale-invariant attention variants

In Section 3 we presented scale-invariant p-RoPE as our proposed method. In preliminary experiments however, it was very natural to also consider scale-invariant RoPE and scale-invariant NoPE. We show results when training at 4k in Fig. 8. Scale-invariant RoPE performs almost as well as scale-invariant p-RoPE when evaluating at the training context length, but underperforms more as we move to 16k and 64k. On the other hand, scale-invariant NoPE underperforms scale-invariant p-RoPE, yet generalises

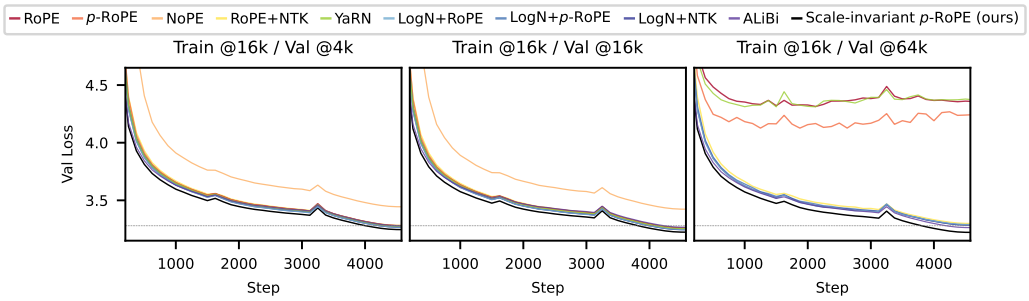


Figure 7: Validation losses at 4k / 16k / 64k context lengths, for a 162M parameter GPT-2-style model trained at 16k. NoPE omitted on the right-most plot to avoid excessive zooming.

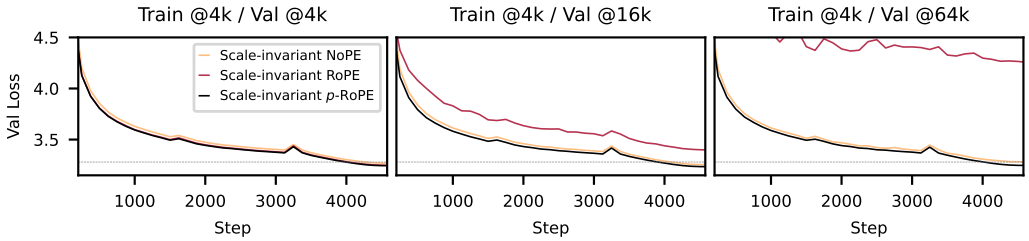


Figure 8: Validation losses at 4k / 16k / 64k of scale-invariant NoPE, RoPE, and p-RoPE, for a 162M parameter GPT-2-style model trained at 4k .

to long contexts. We hypothesised that scale-invariant RoPE does not long-context generalise due to RoPE's low frequences (i.e. high wavelengths) interfering with the position-dependent scale-invariant transformation, $L_t \mapsto a_t L_t + m_t$.

I.4 Infini-Attention

During the review process, we were asked to compare scale-invariant attention to Infini-attention (Munkhdalai et al., 2024). Since infini-attention is a method for compressing the KV-cache, it is slightly different to the other methods we compare to (which primarily control entropy), and so we include these results separately. See Table 3, which shows that while infini-attention gives long-context generalization, it is not as strong as our method.

I.5 Larger models

We also investigated the ability for scale-invariant attention to generalise to longer contexts in larger models by continual pretraining Llama 2 7B (Touvron et al., 2023). We chose Llama 2 for this experiment because it is one of the few models that have not already mid-trained at a longer context length.

Specifically, we continually pretrained at 4k context length after replacing the default attention mechanism with various other methods, and we looked at validation loss at out-of-distribution lengths (4k/16k/64k). We trained using the Torchtune (2024) library, using data from FineWeb (Penedo et al., 2024), with AdamW and a learning rate of 2×10^{-5}.

We see in Table 4 that changing the attention mechanism degraded the loss at the training context length, but the performance of scale-invariant p-RoPE far exceeds the other methods at 16k and 64k.

Table 3: Final mean validation losses (± 1 standard error across 3 seeds) for different methods when training with 4k context length on a 162M parameter GPT-2-style model.

Method	Val @ 4k	Val @ 16k	Val @ 64k
RoPE	3.261 ± 0.001	3.936 ± 0.010	5.260 ± 0.014
Scale-invariant p-RoPE (ours)	$\mathbf{3.244} \pm 0.001$	$\mathbf{3.235} \pm 0.001$	$\mathbf{3.247} \pm 0.001$
Infini-RoPE	3.296 ± 0.003	3.302 ± 0.004	3.310 ± 0.008
Infini-p-RoPE	3.295 ± 0.003	3.303 ± 0.005	3.311 ± 0.009

Table 4: Validation performance when fine-tuning Llama-2 7B with different attention methods on $\sim$50M tokens. RoPE and RoPE+NTK (denoted *) were not fine-tuned.

Method	Val loss @4k	Val loss @16k	Val loss @64k
RoPE*	1.968	7.036	8.815
p-RoPE	2.029	3.504	6.800
NoPE	4.152	6.172	7.730
RoPE+NTK*	1.988	3.090	7.523
YaRN	**1.957**	7.004	8.763
LogN+RoPE	1.966	6.932	8.726
LogN+p-RoPE	2.049	2.984	6.224
LogN+NTK	1.966	3.063	7.413
ALiBi	2.750	2.745	2.744
Scale-invariant p-RoPE (ours)	2.163	**2.193**	**2.252**

J Checking Gaussianity of attention logits

In our analysis in Section 3.2 we assume that the unmodified logits (query-key products), $\bar{L}_t$'s, are standard Gaussian. In this section, we empirically verify that the logits are Gaussian by looking at QQ-plots.

We consider several sizes of model, including 1B and 8B Llama variants (Grattafiori et al., 2024), and Gemma 2 27B Team et al. (2024). We calculate logits with the introduction paragraph of a Wikipedia page [1] as the input. In Figs. 9,10,11,12,13,14 we show quantiles of $\{\bar{L}_t\}_{t>0}$'s (i.e. lower triangular part of the QK^T matrix) for each layer, aggregated over the input and the attention heads in each layer. Note that we do not aggregate over the 'beginning of sequence' token, as it is an outlier attention sink (Gu et al., 2024).

[1] https://en.wikipedia.org/wiki/New_England

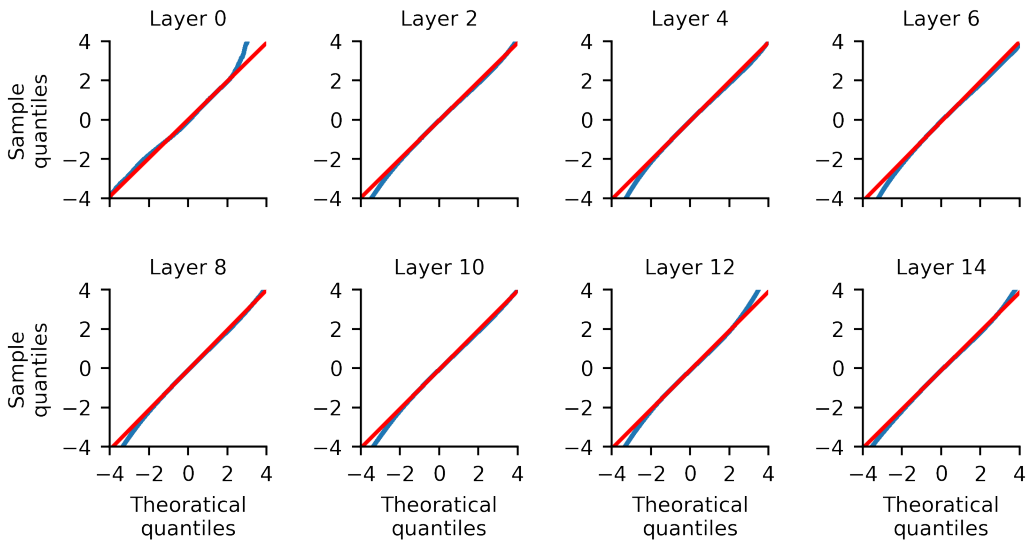


Figure 9: Quantile-Quantile (QQ) plots of attention logits (without RoPE applied) in a Llama-1B model (blue line), with theoretical quantiles of a Gaussian shown by the red line.

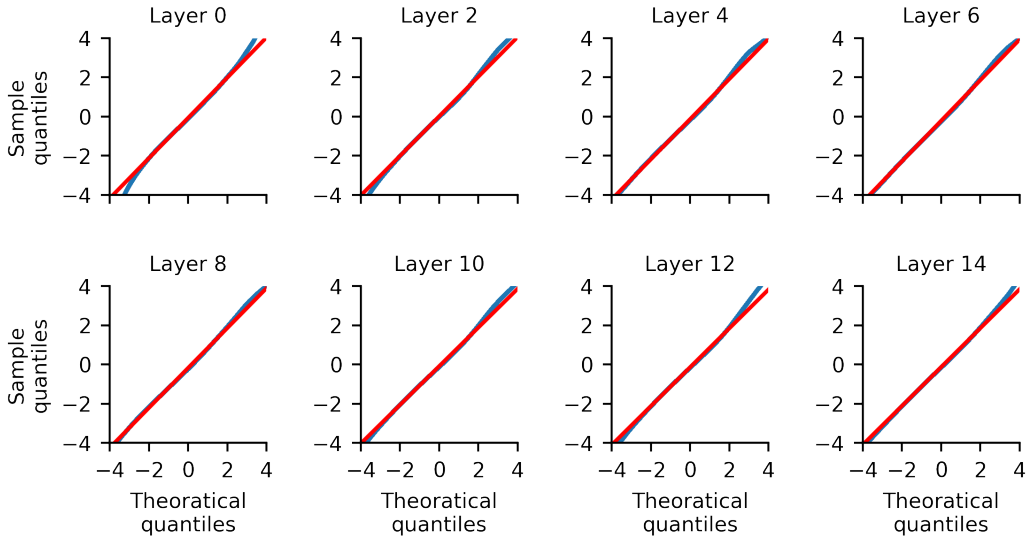


Figure 10: Quantile-Quantile (QQ) plots of attention logits (with RoPE applied) in a Llama-1B model (blue line), with theoretical quantiles of a Gaussian shown by the red line.

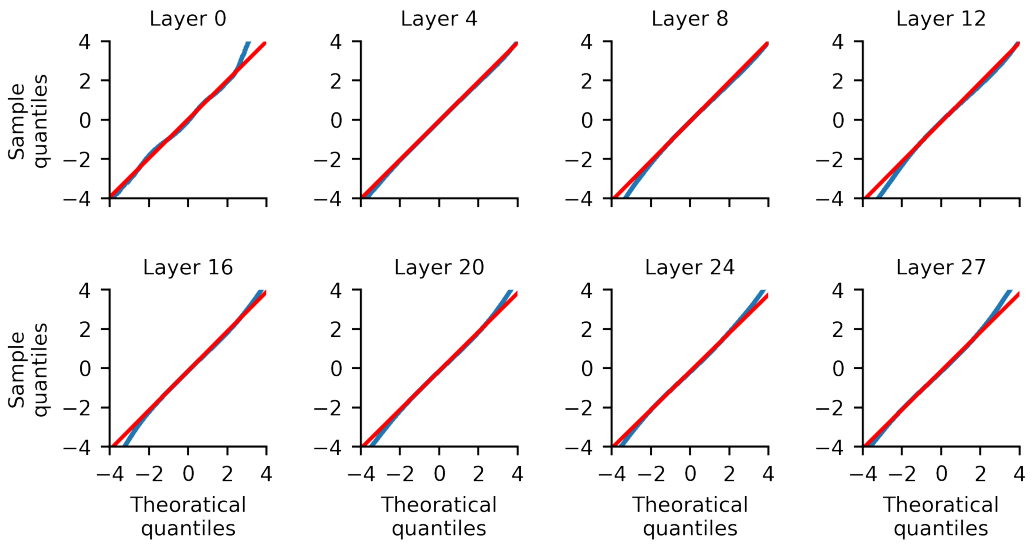


Figure 11: Quantile-Quantile (QQ) plots of attention logits (without RoPE applied) in a Llama-8B model (blue line), with theoretical quantiles of a Gaussian shown by the red line.

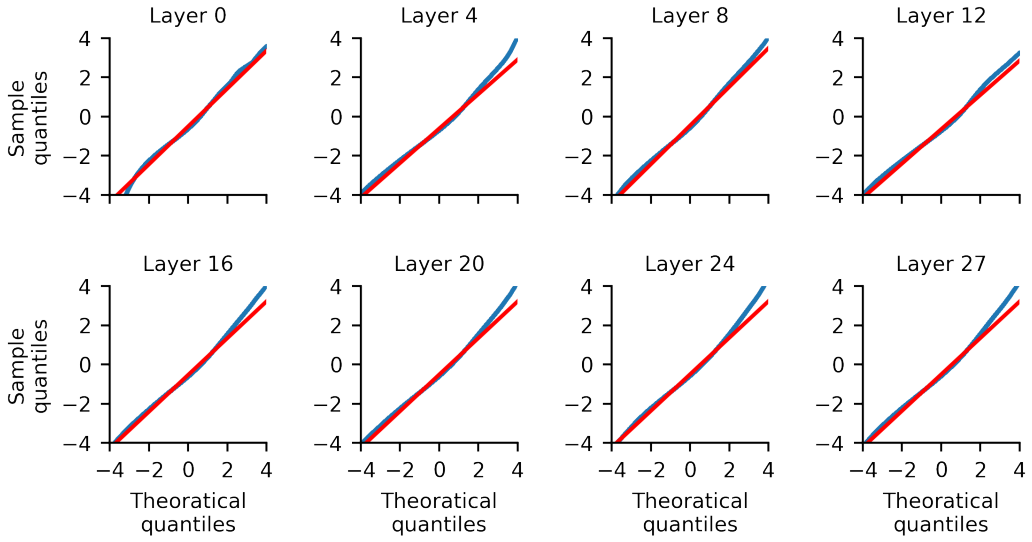


Figure 12: Quantile-Quantile (QQ) plots of attention logits (with RoPE applied) in a Llama-8B model (blue line), with theoretical quantiles of a Gaussian shown by the red line.

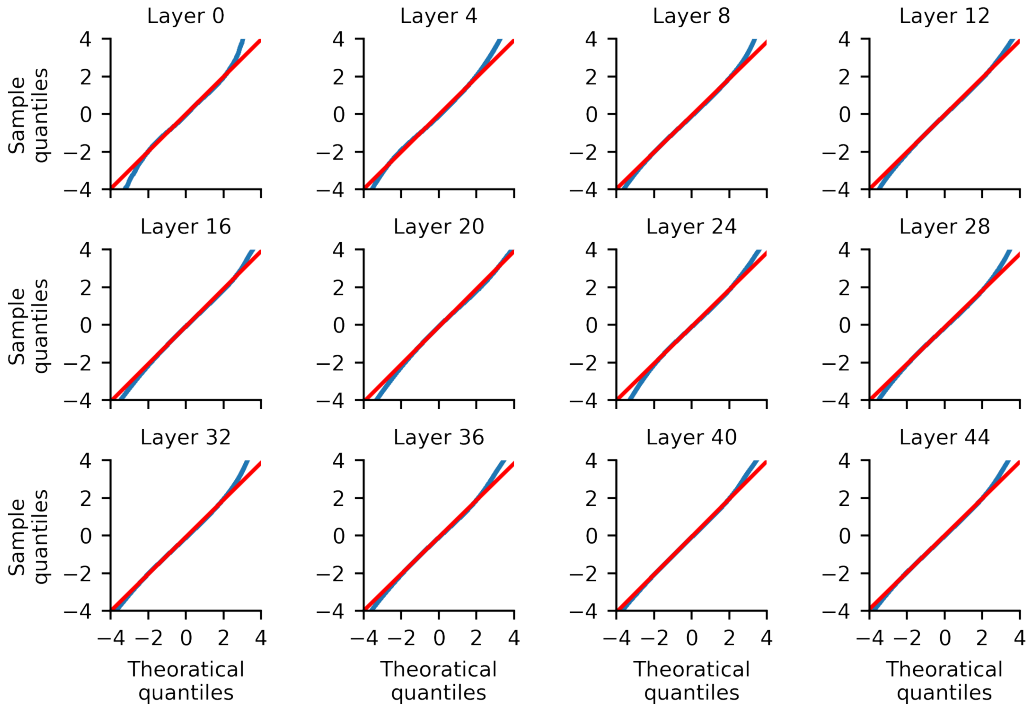


Figure 13: Quantile-Quantile (QQ) plots of attention logits (without RoPE applied) in a Gemma 2 27B model (blue line), with theoretical quantiles of a Gaussian shown by the red line.

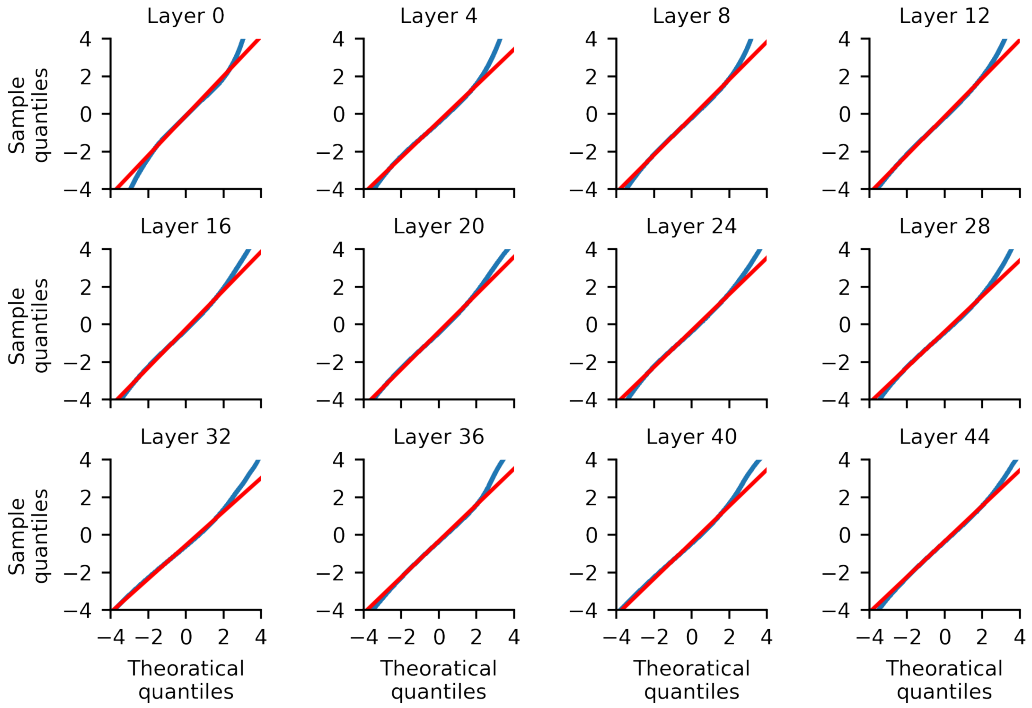


Figure 14: Quantile-Quantile (QQ) plots of attention logits (with RoPE applied) in a Gemma 2 27B model (blue line), with theoretical quantiles of a Gaussian shown by the red line.

K Licenses

- This project uses a modified version of `modded-nanogpt`, https://github.com/KellerJordan/modded-nanogpt which is MIT licensed.
- This project uses a 10B subset of the fineweb dataset, https://huggingface.co/datasets/kjj0/fineweb10B-gpt2, which is MIT licensed.
- This project uses a 100B subset of the fineweb dataset, https://huggingface.co/datasets/kjj0/fineweb100B-gpt2, which is MIT licensed.
- This project uses the C4 dataset, https://huggingface.co/datasets/kjj0/fineweb100B-gpt2, which licensed under the Open Civic Data Attribution License (OCD-BY).