# RoPE Knows Its Limits: Frequency Bands Show Effective Context Length in LLMs

**Anonymous authors**
Paper under double-blind review

## Abstract

Recent large language models claim to support extremely long contexts. However, as demonstrated by the RULER benchmark, the maximum sequence length a model claims to accept can diverge substantially from its effective context length, namely the length over which the model can actually process contextual information. This observation raises a natural question: is effective context length merely an empirical property that can only be measured by running long-context benchmarks, or can it be predicted from a model's internal structure? In this work, we focus on the frequency bands induced by Rotary Position Embeddings (RoPE), which are widely used in modern LLMs. Through our analysis, we find that the trends in effective context length observed in RULER are proportional to the base frequencies of the dimensions in which these frequency bands appear. This suggests that the context length a RoPE-based LLM can truly exploit may be determined by the scale of the frequency bands it actually uses. In hybrid attention architectures that combine Full Attention and Sliding Window Attention, the frequency bands differ depending on the attention type, and these bands indicate the effective context length available to each layer. Our findings recast effective context length as a model-specific spectral quantity, enabling its dataset-independent estimation in RoPE-based LLMs.

## 1 Introduction

Recent large language models increasingly advertise long context windows, ranging from tens of thousands to even millions of tokens (Yang et al., 2025; OpenAI, 2025; Grattafiori et al., 2024). However, the maximum length that a model can accept as input does not necessarily match the length over which it can effectively use information. Long-context benchmarks such as RULER (Hsieh et al., 2024) make this gap explicit by evaluating models across multiple sequence lengths and reporting an effective context length. While such benchmarks are essential for measuring long-context behavior, they require many long-sequence inference runs and treat effective context length primarily as an external behavioral property. This raises a natural question: can we estimate a model's usable context scale from its internal representations?

We study this question through Rotary Position Embedding (RoPE) (Su et al., 2023), a positional encoding method widely used in modern LLMs. RoPE represents positions through frequency-specific rotations of query and key vectors, and its frequencies are known to play an important role in long-context extension through positional interpolation (Peng et al., 2024; bloc97, 2023; Ding et al., 2024). Prior work has shown that pretrained models do not use all RoPE frequency components uniformly: instead, certain high-norm frequency bands emerge in query and key representations and can strongly affect attention logits (Barbero et al., 2025; Oka et al., 2026). However, it remains unclear whether these frequency bands are merely an internal representational pattern, or whether they correspond to the context length that a model can actually use.

In this work, we propose **RoPE-RULER** (Figure 1), a spectral diagnostic that estimates the effective context length of RoPE-based LLMs from high-norm RoPE frequency bands. Our key hypothesis is that the RoPE frequency scale associated with the dominant high-norm band reflects the distance scale available to the model. Rather than using the maximum advertised context length or the global RoPE base frequency, RoPE-RULER identifies the RoPE chunk that is most consistently dominant across layers and heads, and converts its effective RoPE scale into an estimated context length. This



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

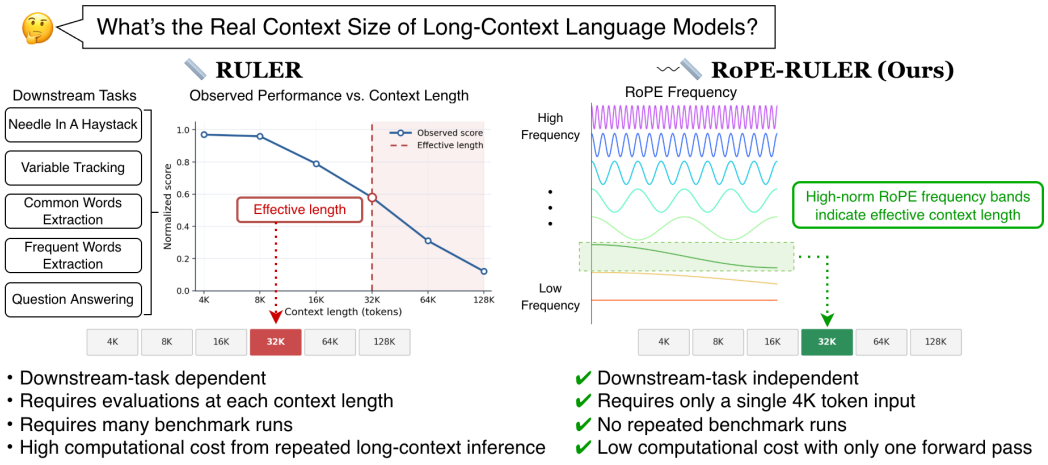



Figure 1: Overview of RULER and our **RoPE-RULER** for estimating effective context length. RULER measures effective context length from benchmark performance across context-length buckets, whereas RoPE-RULER estimates it from high-norm RoPE frequency bands in the model's internal query representations.

requires only unlabeled text and access to its internal query representations, and does not require running long-context benchmark tasks.

Empirically, RoPE-RULER's effective RoPE scale closely matches the RULER-measured effective context length across several RoPE-based long-context LLMs (§4). The identified frequency band is stable across input corpora and context lengths, suggesting that it is not input-specific (§4.3). In hybrid attention architectures, RoPE-RULER identifies different bands for full-attention and sliding-window layers (§5). These bands correspond to layer-specific usable distance scales, closely matching the global context scale for full attention and the local window size for sliding-window attention.

Thus, this study reframes effective context length as a spectral property reflected in the model's internal RoPE frequency structure, rather than only as an external benchmark measurement. Because RoPE-RULER reads out the usable distance from the model's internal RoPE frequency structure rather than relying on a benchmark threshold, it can estimate usable context length even for models whose effective length is not well defined by RULER. When the goal is to estimate usable context length itself, RoPE-RULER offers a lower-cost diagnostic that requires only unlabeled text and access to internal query representations.

## 2 Background and Related Work

**Rotary Position Embedding and position interpolation** Rotary Position Embedding (RoPE) (Su et al., 2023) is a widely used positional encoding method for Transformer-based models. RoPE injects positional information into attention logits by splitting query and key vectors into two-dimensional chunks and rotating each chunk with a frequency-specific rotation. Let $d$ be the head dimension. Each query or key vector is divided into $d/2$ two-dimensional chunks, and the RoPE scale for chunk $i$ is defined as $\theta_i = \theta^{2i/d}$, where $\theta$ is the base frequency. In practice, $\theta$ is typically set to a large value, such as 10,000 or 500,000 (Su et al., 2023; Grattafiori et al., 2024). Equivalently, the angular frequency of chunk $i$ is given by $\omega_i = 1/\theta_i$. The RoPE rotation for chunk $i$ at position $m$ can then be written as

$$R_i(m) = \begin{pmatrix} \cos(m/\theta_i) & -\sin(m/\theta_i) \\ \sin(m/\theta_i) & \cos(m/\theta_i) \end{pmatrix}. \tag{1}$$

Therefore, for the $i$-th query chunk $q_m^{(i)} \in \mathbb{R}^2$ and the $i$-th key chunk $k_n^{(i)} \in \mathbb{R}^2$, the RoPE-augmented attention logit is computed as

$$a_{mn} = \sum_{i=0}^{d/2-1} \left\langle R_i(m) q_m^{(i)}, R_i(n) k_n^{(i)} \right\rangle. \tag{2}$$

In long-context LLMs, positional interpolation is often applied to RoPE in order to support sequence lengths beyond the maximum length used during pretraining (Peng et al., 2024; bloc97, 2023; Ding et al., 2024).

**Frequency bands in RoPE** Although RoPE is widely used, how pretrained LLMs use its individual frequency components remains poorly understood. Barbero et al. (2025) questioned the common view that RoPE simply attenuates attention with relative distance, and analyzed the role of different RoPE frequencies in pretrained models. They used the norms of query/key RoPE chunks to identify frequency components that can strongly affect attention logits. More recently, Oka et al. (2026) extended this analysis to multiple LLMs, including models with positional interpolation. They showed that high-norm frequency bands consistently emerge across models, and reported that their locations are determined by the maximum RoPE base frequency $\theta$ and the pretraining context length. They also found that these bands form early in pretraining and are inherited after positional interpolation. Together, these studies identify high-norm RoPE frequency bands as a common feature of pretrained LLMs. However, it remains unclear whether these bands explain the effective context length measured by benchmarks such as RULER (Hsieh et al., 2024).

**Hybrid full and windowed attention for Long Context** To handle long contexts efficiently, recent LLMs sometimes adopt architectures that combine full attention (FA), often implemented with Grouped-Query Attention (GQA) (Ainslie et al., 2023), with local sliding window attention (SWA) (Child et al., 2019), rather than using FA in all layers (OpenAI, 2025; Google DeepMind, 2026). FA layers can attend to the entire input sequence, whereas SWA layers restrict each token to attend only to tokens within a predefined window of size $W$.

**Long-context benchmarks and effective context length** Recent LLMs increasingly claim very long context windows, from 32K or 128K tokens to even 1M tokens (Yang et al., 2025; OpenAI, 2025; Grattafiori et al., 2024). Long-context benchmarks such as LongBench (Bai et al., 2024), L-Eval (An et al., 2024), InfiniteBench (Zhang et al., 2024), and needle-in-a-haystack (NIAH) tests evaluate these capabilities. However, accepting long inputs does not necessarily mean effectively using the entire context. RULER (Hsieh et al., 2024) formalizes this gap by measuring effective context length across multiple sequence lengths. While these benchmarks quantify the gap between claimed and effective context lengths, they require many long-sequence inference runs. Furthermore, they primarily measure external task behavior; although they reveal failure modes, they do not directly explain the internal mechanisms underlying the degradation.

# 3 RoPE Knows Its Limits: A Spectral Diagnostic for Effective Length

We propose a diagnostic method that estimates the distance scale available to a trained model from its internal representations. Rather than introducing a new architecture or training method, our goal is to read out usable context length from the model itself. We consider the following hypothesis.

**Hypothesis.** *High-norm frequency bands encode information about the effective context length, and the associated RoPE frequency scales $\theta_i$ are predictive of the effective context length.*

Motivated by this hypothesis, we define RoPE-RULER, a spectral diagnostic method that estimates a model's effective context length from the RoPE frequency bands observed in its internal representations.

## 3.1 Why Focus on Frequency Bands?

Following Barbero et al. (2025), we observe that each RoPE-rotated contribution to the attention logit is upper-bounded by the Cauchy–Schwarz inequality:

$$\left|\left\langle R_i(m)q_m^{(i)}, R_i(n)k_n^{(i)}\right\rangle\right| \leq \|R_i(m)q_m^{(i)}\|_2\|R_i(n)k_n^{(i)}\|_2. \tag{3}$$

A RoPE chunk $i$ with large query/key $\ell_2$ norms suggests that the corresponding frequency component $\theta_i$ can substantially influence the attention logit.

Thus, we view high-norm RoPE frequency bands not merely as representational byproducts, but as prime candidates for the dimensions through which trained models use positional information in attention. Note that the maximum base frequency $\theta$, typically specified as a RoPE hyperparameter, merely determines the set of positional scales $\mathcal{S}_\theta = \left\{\theta_i = \theta^{2i/d} \;\middle|\; i = 0, 1, \ldots, \frac{d}{2} - 1\right\}$ available to the architecture, rather than directly corresponding to the maximum context length. Therefore, rather than comparing the maximum base frequency $\theta$, we estimate a model's effective context length from the base frequency scales that are actually used with high norms in its query/key representations.

### 3.2 RoPE-RULER

RoPE-RULER estimates the effective context length in two steps: (1) it identifies high-norm dimensions in the query representations, and (2) it computes the RoPE distance scale corresponding to the identified frequency band.

**(1) Measuring High-norm Frequency-band Index** Given a trained model $M$, we feed an unlabeled text sequence $\mathcal{X}$ into the model and extract the query representations after RoPE has been applied at each attention layer.[1] For each layer $\ell$, head $h$, token position $m$, and RoPE chunk $i$, we measure the strength of chunk $i$ by the $\ell_2$ norm of its two-dimensional query vector: $s_{\ell,h,i}(m) = \|q^{(i)}_{\ell,h,m}\|_2$. We then average this quantity over token positions to obtain a single score for each layer, head, and RoPE chunk:

$$S_{\ell,h,i} = \mathbb{E}_m \left[ s_{\ell,h,i}(m) \right]. \tag{4}$$

Thus, $S_{\ell,h,i}$ represents how strongly RoPE chunk $i$ is used in the query representations of head $h$ in layer $\ell$. Next, for each layer and head, we select the RoPE chunk with the largest average norm:

$$i^*_{\ell,h} = \arg\max_i S_{\ell,h,i}. \tag{5}$$

This gives one dominant chunk for every layer–head pair. To obtain a model-level summary, we then choose the RoPE chunk that is selected most often across all layers and heads:

$$i_{\text{band}} = \arg\max_i \sum_{\ell=1}^{L} \sum_{h=1}^{H} \mathbf{1}\left[i^*_{\ell,h} = i\right], \quad \mathbf{1}\left[i^*_{\ell,h} = i\right] = \begin{cases} 1, & \text{if } i^*_{\ell,h} = i, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

We refer to the resulting index $i_{\text{band}}$ as the representative high-norm frequency-band index of the model. This majority-based definition is intended to capture the frequency component that is consistently used across layers and heads. It avoids selecting a component solely because it has an exceptionally large norm in a small number of individual layers or heads. This differs from Oka et al. (2026), who average norm values across layers and heads; instead, we identify the RoPE chunk that is most frequently dominant across layer–head pairs.

**(2) Converting the Dominant RoPE Band into an Effective Context Length** After identifying the dominant RoPE chunk $i_{\text{band}}$, which is the chunk most frequently selected as high-norm across layers and heads, we convert this chunk into a positional scale. In standard RoPE, given the head dimension $d$ and the RoPE base frequency $\theta$, the scale associated with chunk $i_{\text{band}}$ is $\theta_{i_{\text{band}}} = \theta^{\frac{2 \times i_{\text{band}}}{d}}$. However, practical long-context LLMs often apply positional interpolation or RoPE scaling methods, such as YaRN, LongRoPE, or Llama scaling. Therefore, directly using the raw scale $\theta^{2i/d}$ may not reflect the actual RoPE frequencies used by the model. To account for this, we use the effective angular frequencies $\omega^{\text{eff}}$ after positional interpolation or RoPE scaling has been applied.[2] For the dominant chunk $i_{\text{band}}$, we define its effective positional scale as

$$\theta^{\text{eff}}_{i_{\text{band}}} = \frac{1}{\omega^{\text{eff}}_i} \tag{7}$$

In RoPE, the phase of chunk $i$ at position $m$ is given by $\phi_i(m) = m\,\omega^{\text{eff}}_i$. Although each RoPE component has period $2\pi$, our analysis focuses on the initial phase interval from 0 to $\frac{\pi}{2}$, where the rotational signal changes most rapidly. We then define the effective context length associated with chunk $i_{\text{band}}$ as

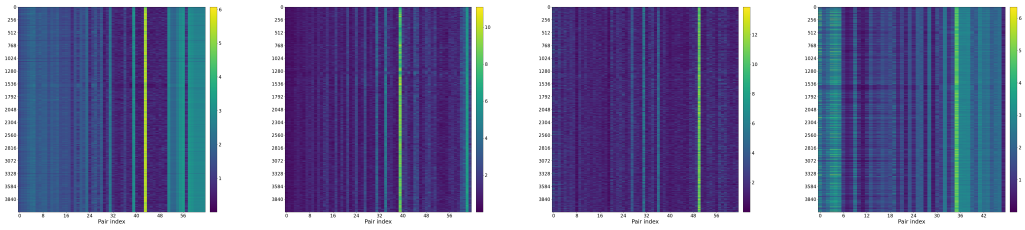
$$L^{\text{eff}}_{i_{\text{band}}} = \frac{\pi}{2} \theta^{\text{eff}}_{i_{\text{band}}} \tag{8}$$

[1] We also conducted the same analysis on key representations and observed no substantial differences between queries and keys. The results are provided in Appendix B.

[2] In Hugging Face implementations, $\omega^{\text{eff}}$ is stored as `inv_freq`.

Table 1: RULER scores across context lengths. The red downward triangle ▼ marks the RULER-estimated effective length, while the green downward triangle ▼ marks the RoPE-RULER-estimated effective frequency mapped to the nearest RULER length bucket.

| Models | Claimed Len. | Effective Len. (RULER ▼) | Effective Len. (Ours ▼) | 4K | 8K | 16K | 32K | 64K | 128K |
|---|---|---|---|---|---|---|---|---|---|
| Mistral-7B | 32K | 16K | 16879.7 | 93.6 | 91.2 | 87.2 ▼▼ | 75.4 | 49.0 | 13.8 |
| Llama3.1-8B | 128K | 32K | 37325.2 | 95.5 | 93.8 | 91.6 | 87.4 ▼▼ | 84.7 | 77.0 |
| Qwen3-4B | 128K | 32K | 61641.0 | 95.1 | 93.6 | 91.0 | 87.8 ▼▼ | 77.8 | 66.0 |
| Qwen3-8B | 128K | 32K | 61641.0 | 96.3 | 96.0 | 91.8 | 91.2 ▼▼ | 82.1 | 77.4 |
| Qwen3-14B | 128K | 64K | 76492.6 | 98.0 | 97.8 | 96.4 | 96.1 | 94.0 ▼▼ | 85.1 |
| Qwen3-32B | 128K | 64K | 76492.6 | 98.4 | 96.0 | 96.2 | 94.4 | 91.8 ▼ | 85.6 ▼ |
| Phi3-mini 3.8B | 128K | 32K | 52675.1 | 92.2 | 91.5 | 90.7 | 87.5 ▼▼ | 80.6 | 66.7 |



(a) Mistral-7B (b) Llama-3.1-8B (c) Qwen3-8B (d) Phi3-mini 3.8B

Figure 2: Examples of $\ell_2$-norm heatmaps for each model.The horizontal axis denotes the RoPE frequency index ($0$ to $d/2-1$), and the vertical axis denotes the token position ($0$ to $4096$). Yellow regions correspond to high-norm frequency bands.

## 4 Experiments

### 4.1 Settings

We evaluate three RoPE-based long-context LLMs: Mistral-7B-Instruct-v0.2, Llama-3.1-8B-Instruct, Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen3-32B, and Phi-3-mini-128k-instruct (3.8B). For each model, we use the first $4{,}096$ tokens of the Wikitext-103 test set (Merity et al., 2017) as input. Additional experimental details are provided in Appendix A.

In the RULER evaluation setting, for each model and each target sequence length, we use 500 examples per task and report the average score over all 13 tasks. The evaluated context-length buckets are 4K, 8K, 16K, 32K, 64K, and 128K tokens. Following RULER, the effective context length is defined as the largest evaluated length at which the average score remains above the qualitative threshold given by the Llama-2-7B 4K score, namely 85.6.

### 4.2 Results

Table 1 compares an estimated effective context length using Eq. (8) with the effective context length reported by RULER. Across the three models, our estimates fall into the same RULER length bucket as the benchmark-based effective length. Although the claimed context lengths differ from the effective lengths, the frequency band scale extracted from the model's internal query representations closely tracks the RULER effective context length.

Figure 2 shows examples of the query L2-norm heatmaps for each models. In all models, we observe high-norm frequency bands, shown as yellow bands in the heatmaps. Table 2 summarizes the model-specific frequency-band

Table 2: Model-specific frequency-band index $i_{\text{band}}$ and effective scales $\theta^{\text{eff}}_{i_{\text{band}}}$.

| Model | $i_{\text{band}}$ | $\theta^{\text{eff}}_{i_{\text{band}}}$ |
|---|---|---|
| Mistral-7B | 43 | 10746.0 |
| Llama-3.1-8B | 39 | 23762.9 |
| Qwen3-4B | 49 | 39241.8 |
| Qwen3-8B | 49 | 39241.8 |
| Qwen3-14B | 50 | 48696.7 |
| Qwen3-32B | 50 | 48696.7 |
| Phi3-mini 3.8B | 35 | 33534.0 |

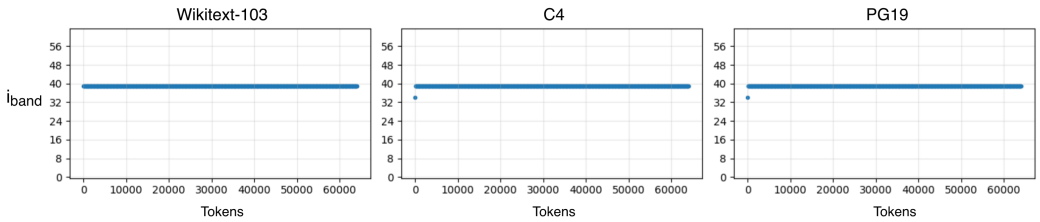



Figure 4: Robustness of the high-norm frequency-band index $i_{\text{band}}$ for Llama-3.1-8B. RoPE-RULER identifies the same $i_{\text{band}}$ across Wikitext-103, C4, and PG19 for input lengths from 1k to 64k tokens, showing that the dominant frequency band is stable across both input corpora and context lengths.

index $i_{\text{band}}$ and its corresponding effective RoPE scale $\theta^{\text{eff}}_{i_{\text{band}}}$. The identified $i_{\text{band}}$ aligns well with the frequency dimensions in which high-norm bands appear in Figure 2. [3] These results support our hypothesis that the effective context length of RoPE-based LLMs is reflected in the RoPE frequency band that the model actually uses. Importantly, the estimate is obtained without running long-context benchmark tasks: it only requires unlabeled text inputs and access to the model's query representations. This suggests that effective context length can be interpreted not only as an external behavioral property measured by benchmarks, but also as a model-specific spectral quantity encoded in RoPE representations.

### 4.3 Robustness Study

Because RoPE-RULER is based on norm statistics of query representations, we need to verify that the identified high-norm frequency band is not an artifact of a specific input corpus or context length. To test this, we compare the high-norm frequency-band index $i_{\text{band}}$ under different input settings. In addition to Wikitext-103, we use C4 (Raffel et al., 2020) and PG19 (Rae et al., 2020), and vary the input length from 1k to 64k tokens.

Figure 4 shows the results. Across most input lengths and corpora, RoPE-RULER consistently selects $i_{\text{band}} = 39$. This suggests that the detected high-norm frequency band is not specific to a particular dataset or sequence length, but instead reflects a model-intrinsic spectral structure. It also suggests that RoPE-RULER can remain effective even when applied to relatively short input sequences. Figure 3 further shows the $\ell_2$-norm heatmap for the 64k-token setting. The same high-norm frequency band remains visible even at this longer context length, indicating that the band is stable as the input length increases. Additional heatmaps and the corresponding $\theta^{\text{eff}}_{i_{\text{band}}}$ values for each dataset and context length are provided in Appendix C. These results further support that the identified frequency band is stable across input content and input length.

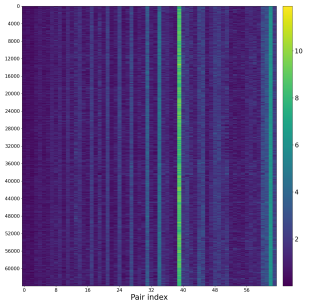



Figure 3: Query $\ell_2$-norm heatmap for Llama-3.1-8B on Wikitext-103 at 64k tokens.

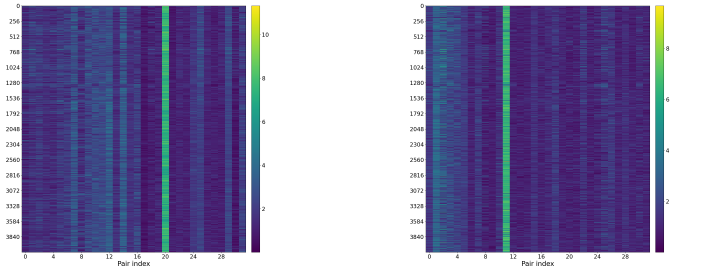
## 5 Estimating Effective Length Beyond RULER

In the previous section, we showed that high-norm RoPE frequency bands correspond to the effective context length measured by RULER. **Can RoPE-RULER still provide a useful estimate when the effective context length is not well defined under RULER's threshold criterion?** RULER defines the effective context length as the maximum sequence length at which the model score exceeds a threshold. However, some models do not reach this threshold even when they have a large claimed context length and strong downstream performance. For example, gpt-oss-120b supports a 128K

[3] Appendix B provides a layer- and head-wise analysis of the identified $i_{\text{band}}$. We also include the corresponding analysis for key representations. We observe no substantial difference between the query and key distributions.

Table 3: RULER scores across context lengths. The green downward triangle ▼ marks the RoPE-RULER-estimated effective frequency mapped to the nearest RULER length bucket.

| Models | Claimed Len. | Effective $\theta^{\text{eff}}_{i_{\text{band}}}$ (Ours ▼) | 128k | 256K | 512K | 1M |
|---|---|---|---|---|---|---|
| GPT-OSS 120B | 128k | 181899 | 55.01 ▼ | 52.30 | 46.70 | 22.30 |




(a) Full Attention (FA) (b) Sliding Window Attention (SWA)

Figure 5: Examples of L2-norm heatmaps for the GPT-OSS 120B model with FA and SWA.

context length, but has been reported not to satisfy the RULER threshold (Hsieh et al., 2024; NVIDIA et al., 2026). Moreover, gpt-oss adopts a hybrid architecture that combines Full Attention (FA) and Sliding Window Attention (SWA), suggesting that different layers may have different usable distance scales.

In this section, we examine whether RoPE-RULER can estimate the effective distance scale from internal representations for models that are difficult to characterize using RULER's threshold-based definition. We further compare FA and SWAlayers to test whether RoPE frequency bands reflect the distance scale actually available to each layer.

### 5.1 Settings

We apply RoPE-RULER to `openai/gpt-oss-120b` (OpenAI, 2025). This model uses a hybrid-attention architecture with interleaved FA and SWA layers, where SWA layers have a window size of 128 tokens. This setting allows us to compare RoPE-RULER estimates and L2-norm patterns between FA and SWA layers. We use the first $4{,}096$ tokens of the Wikitext-103 test set (Merity et al., 2017) as input.

### 5.2 Results

Table 3 shows the result for GPT-OSS 120B. Although GPT-OSS 120B claims a 128K context length, its RULER scores remain below the threshold, leaving its effective length undefined under the RULER criterion. In contrast, RoPE-RULER estimates the effective RoPE scale from the model's internal frequency band. The estimated value is $\theta^{\text{eff}}_{i_{\text{band}}} = 181899.9$, which is mapped to the nearest RULER length bucket, 128K. This suggests that, even when RULER scores are below the threshold, the RoPE frequency band still provides a meaningful estimate of the distance scale available to the model. This result indicates that RoPE-RULER can serve as a complementary diagnostic for models whose effective context length is difficult to determine from benchmark thresholds alone.

### 5.3 Attention-Type Robustness Study

We next apply RoPE-RULER layer by layer to GPT-OSS-120B, a hybrid model with Full Attention (FA) and Sliding Window Attention (SWA) layers. Because FA layers attend globally while SWA layers are limited to a window of size $W = 128$, their RoPE frequency bands should reflect different usable distance scales.

Figure 5 shows that both FA and SWA layers exhibit clear high-norm frequency bands. Thus, the frequency-band structure is not limited to global attention layers; it also appears in local-window layers. However, the locations of the bands differ substantially between the two attention types. As shown in Table 4, FA layers yield $L^{\text{eff}}_{i_{\text{band}}} = 181899$, which is on the scale of the model's claimed 128K context length. In contrast, SWA layers yield $L^{\text{eff}}_{i_{\text{band}}} = 135$, closely matching the SWA window size of 128. These results suggest that RoPE-RULER operates at the layer level in hybrid attention models. Therefore, the RoPE frequency band is better understood as a layer-specific spectral indicator of usable context length, rather than a quantity determined only by the model's nominal context length.

Table 4: RoPE-RULER estimates for Full Attention (FA) and Sliding Window Attention (SWA) layers in GPT-OSS-120B. The model has a claimed context length of 128K, while SWA layers use a window size of 128.

| Atten. | $i_{\text{band}}$ | $\theta^{\text{eff}}_{i_{\text{band}}}$ | $L^{\text{eff}}_{i_{\text{band}}}$ |
|---|---|---|---|
| FA | 20 | 115801.1 | 181899 |
| SWA | 11 | 86.2 | 135 |

# 6 Why Do Frequency Bands Correspond to Effective Context Length?

In this section, we provide a theoretical explanation for why the base frequency of the high-norm frequency band defined in Eq. 8 corresponds to the effective context length. In particular, we explain why the relation $L_{\text{eff}} \approx (\pi/2)\theta^{\text{eff}}_{i_{\text{band}}}$ naturally arises.

## 6.1 RoPE Dimensions Have Distance Scales

The $i$-th RoPE component represents each absolute position $m$ by a two-dimensional rotation. Therefore, for a sequence of length $L$, the phase range experienced by the query from token position $0$ to token position $L$ is given by

$$\phi_i(L-0) = \frac{L-0}{\theta^{\text{eff}}_i} = \frac{L}{\theta^{\text{eff}}_i} \tag{9}$$

The key point is not that a RoPE component has a full period of $2\pi$, but rather how much non-overlapping positional signal it provides over the absolute position range $1, \ldots, L$ observed during training. When the phase range satisfies $0 \le \phi_i(L) \le \pi/2$, $\cos\phi$ changes monotonically from 1 to 0, while $\sin\phi$ changes monotonically from 0 to 1. Thus, the two-dimensional RoPE representation stays within the first quadrant, without self-intersection or sign reversal, and spreads smoothly as a function of absolute position. Within this range, larger $\phi_i(L)$ leads to greater separation between the representations of the beginning and the end of the sequence. Therefore, $\phi_i(L) \approx \pi/2$ can be interpreted as the largest phase range that remains monotonic and non-overlapping as a positional signal. This yields

$$\frac{L}{\theta^{\text{eff}}_i} \approx \frac{\pi}{2} \quad \Longleftrightarrow \quad L \approx \frac{\pi}{2}\theta^{\text{eff}}_i \tag{10}$$

as a natural scale linking RoPE frequency bands to effective context length.

## 6.2 Low Frequency

Consider the case where $\frac{L}{\theta^{\text{eff}}_i} < \frac{\pi}{2}$, that is, where $\theta^{\text{eff}}_i$ is too large for the sequence length $L$. In this case, the RoPE phase change over the entire sequence remains within only part of the first quadrant. The positional representation is therefore monotonic and stable, but the separation between the beginning and the end of the sequence is not sufficiently large. In particular, when $\theta^{\text{eff}}_i$ is very large, we have

$$\cos(x/\theta^{\text{eff}}_i) \approx 1, \qquad \sin(x/\theta^{\text{eff}}_i) \approx x/\theta^{\text{eff}}_i$$

so the cosine component changes very little, and the sine component also varies only weakly. Such overly low-frequency components are smooth over long distances, but they provide weak signals for separating positions within the training range. Consequently, even if the query/key representations assign large norms to such frequency components, the usable positional difference in the attention logit remains limited.

### 6.3 HIGH FREQUENCY

Next, consider the case where $\frac{L}{\theta_i^{\text{eff}}} > \frac{\pi}{2}$, that is, where $\theta_i^{\text{eff}}$ is too small. In this case, the RoPE phase exceeds the first quadrant before reaching the end of the sequence, and the cosine component begins to move below zero. Although this means that the positional representation continues to change, the change is no longer a simple monotonic signal that smoothly spreads from nearby to distant positions. In particular, because the contribution of a RoPE component to the attention logit appears through a rotated query/key inner product, once the phase exceeds $\pi/2$, in-phase correlations are weakened; as the phase advances further, the representation enters a regime involving sign reversal. In this regime, the signal for distinguishing distant positions may appear large, but its interpretation becomes less stable. Thus, overly high-frequency components may be useful for local positional discrimination, but they tend to oscillate too rapidly to serve as a consistent scale for using long-range contextual information.

## 7 CONCLUSION AND LIMITATION

In this work, we proposed RoPE-RULER, a spectral diagnostic for estimating the effective context length of RoPE-based language models from their internal frequency-band structure. Across multiple RoPE-based long-context LLMs, the effective RoPE scale associated with the dominant high-norm frequency band closely matched the effective context length measured by RULER. Instead of treating effective context length solely as an external behavioral quantity measured by long-context benchmarks, we showed that it can also be interpreted as a model-specific spectral quantity encoded in the query/key representations. We further showed that this relationship extends to hybrid attention architectures: in full-attention layers, the identified frequency band corresponds to the global usable context scale, whereas in sliding-window attention layers, it corresponds to the local window size.

Our results provide a mechanistic interpretation of why RULER-measured effective context length is reflected in a particular RoPE frequency scale. Because the contribution of each RoPE component to the attention logits is controlled by the norm of the corresponding query/key chunk, high-norm frequency bands indicate the frequency components that the model can strongly use for attention. RoPE-RULER therefore explains the effective context length as the distance scale associated with the strongest RoPE frequency band, rather than as a property determined only by the advertised context window or the global RoPE base frequency.

We emphasize that RoPE-RULER is not intended to replace long-context benchmarks such as RULER. Benchmarks remain essential for evaluating downstream task performance, robustness, and the actual behavior of models under realistic long-context inputs. In contrast, RoPE-RULER does not measure task accuracy directly. Its purpose is complementary: it provides a low-cost internal diagnostic of the distance scale that a model appears to make available through its RoPE representations. In our experiments, this estimate can be obtained from unlabeled text and relatively short forward passes, rather than requiring many long-context benchmark evaluations across multiple sequence lengths.

A key limitation of RoPE-RULER is that it requires access to internal query/key representations and RoPE configurations. Therefore, the method is currently most applicable to open-source or otherwise inspectable models, and cannot be directly applied to closed models for which activations and implementation details are unavailable. Another limitation is that our analysis focuses on RoPE-based architectures. Extending the same idea to models with other positional encoding schemes remains an important direction for future work. Finally, although the dominant frequency band provides a strong indicator of usable context scale, it does not capture all factors that affect long-context performance, such as retrieval behavior, instruction following, memory composition across layers, or task-specific reasoning failures. Thus, RoPE-RULER should be viewed as a complementary spectral diagnostic for effective context length rather than a complete substitute for behavioral evaluation.

Overall, our findings suggest that effective context length is not only a benchmark-level property, but also has a measurable signature inside the model. By connecting long-context behavior to the dominant RoPE frequency band, RoPE-RULER provides a dataset-light and computationally efficient way to diagnose the usable context scale of RoPE-based LLMs.

## 8 AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting parts of the manuscript. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in writing proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results. We have reviewed all AI-assisted content. The co-authors have also verified that the manuscript does not contain plagiarized material from other works and that all citations and references are accurate and free from AI-generated hallucinations. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

## REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.298/.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14388–14411, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.776. URL https://aclanthology.org/2024.acl-long.776/.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.172.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *The Thirteenth*

*International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GtvuNrk58a.

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

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

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens, 2024. URL https://arxiv.org/abs/2402.13753.

Google DeepMind. Gemma 4 26b a4b it. https://huggingface.co/google/gemma-4-26B-A4B-it, 2026. Hugging Face model card. Accessed: 2026-05-04.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu,

Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

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

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

NVIDIA, :, Aakshita Chandiramani, Aaron Blakeman, Abdullahi Olaoye, Abhibha Gupta, Abhilash Somasamudramath, Abhinav Khattar, Adeola Adesoba, Adi Renduchintala, Adil Asif, Aditya Agrawal, Aditya Vavre, Ahmad Kiswani, Aishwarya Padmakumar, Ajay Hotchandani, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, Aleksandr Shaposhnikov, Alex Gronskiy, Alex Kondratenko, Alex Neefus, Alex Steiner, Alex Yang, Alexander Bukharin, Alexander Young, Ali Hatamizadeh, Ali Taghibakhshi, Alina Galiautdinova, Alisa Liu, Alok Kumar, Ameya Sunil Mahabaleshwarkar, Amir Klein, Amit Zuker, Amnon Geifman, Anahita Bhiwandiwalla, Ananth Subramaniam, Andrew Tao, Anjaney Shrivastava, Anjulie Agrusa, Ankur Srivastava, Ankur Verma, Ann Guan, Anna Shors, Annamalai Chockalingam, Anubhav Mandarwal, Aparnaa Ramani, Arham Mehta, Arti Jain, Arun Venkatesan, Asha Anoosheh, Ashwath Aithal, Ashwin Poojary, Asif Ahamed, Asit Mishra, Asli Sabanci Demiroz, Asma Kuriparambil Thekkumpate, Atefeh Sohrabizadeh, Avinash Kaur, Ayush Dattagupta, Barath Subramaniam Anandan, Bardiya Sadeghi, Barnaby Simkin, Ben Lanir, Benedikt Schifferer, Benjamin Chislett, Besmira Nushi, Bilal Kartal, Bill Thiede, Bita Darvish Rouhani, Bobby Chen, Boris Ginsburg, Brandon Norick, Branislav Kisacanin, Brian Yu, Bryan Catanzaro, Buvaneswari Mani, Carlo del Mundo, Chankyu Lee, Chanran Kim, Chantal Hwang, Chao Ni, Charles Wang, Charlie Truong, Cheng-Ping Hsieh, Chenhan Yu, Chenjie Luo, Cherie Wang, Chetan Mungekar, Chintan Patel, Chris Alexiuk, Chris Holguin, Chris Wing, Christian Munley, Christopher Parisien, Chuck Desai, Chunyang Sheng, Collin Neale, Cyril Meurillon, Dakshi Kumar, Dan Gil, Dan Su, Dane Corneil, Daniel Afrimi, Daniel Burkhardt Eliuth Triana, Daniel Egert, Daniel Fatade, Daniel Lo, Daniel Rohrer, Daniel Serebrenik, Daniil Sorokin, Daria Gitman, Daria Levy, Darko Stosic, David Edelsohn, David Messina, David Mosallanezhad, David Tamok, Deena Donia, Deepak Narayanan, Devin O'Kelly, Dheeraj Peri, Dhruv Nathawani, Di Wu, Dima Rekesh, Dina Yared, Divyanshu Kakwani, Dmitry Konyagin Brandon Tuttle, Dong Ahn, Dongfu Jiang, Dorrin Poorkay, Douglas O'Flaherty, Duncan Riach, Dusan Stosic, Dustin Van Stee, Edgar Minasyan, Edward Lin, Eileen Peters Long, Elad Segal, Elena Lantz, Elena Lewis, Ellie Evans, Elliott Ning, Eric Chung, Eric Harper, Eric Pham-Hung, Eric W. Tramel, Erick Galinkin, Erik Pounds, Esti Etrog, Evan Briones, Evan Wu, Evelina Bakhturina, Evgeny Tsykunov, Ewa Dobrowolska, Farshad Saberi Movahed, Farzan Memarian, Fay Wang, Fei Jia, Felipe Soares, Felipe Vieira Frujeri, Feng Chen, Fengguang Lin, Ferenc Galko, Fortuna Zhang, Frankie Siino, Frida Hou, Gantavya Bhatt, Gargi Prasad, Geethapriya Venkataramani, Geetika Gupta, George Armstrong, Gerald Shen, Giulio Borghesi, Gordana Neskovic, Gorkem Batmaz, Grace Lam, Grace Wu, Greg Pauloski, Greyson Davis, Grigor Nalbandyan, Guoming Zhang, Guy Farber, Guyue Huang, Haifeng Qian, Haran Kumar Shiv Kumar, Harry Kim, Harsh Sharma, Hayate Iso, Hayley Ross, Herbert Hum, Herman Sahota, Hexin Wang, Himanshu Soni, Hiren Upadhyay, Huy Nguyen, Iain Cunningham, Ido Galil, Ido Shahaf, Igino Padovani, Igor Gitman, Igor Shovkun, Ikroop Dhillon, Ilya Loshchilov, Ingrid Kelly, Itamar Schen, Itay Levy, Ivan Moshkov, Izik Golan, Izzy Putterman, Jain Tu, Jan Baczek, Jan Kautz, Jane Polak Scowcroft, Janica Rosenberg, Jared Casper, Jarrod Pflum, Jason Grant, Jason Sewall, Jatin Mitra, Jeffrey Glick, Jenny Chen, Jesse Oliver, Jiacheng Xu, Jiafan Zhu, Jialin Song, Jian Zhang, Jiaqi Zeng, Jie Lou, Jill Milton, Jim Chow, Jimmy Zhang, Jinhang Choi, Jining Huang, Jocelyn Huang, Joel Caruso, Joey Conway, Joey Guman, Johan Jatko, John Kamalu, Johnny Greco, Jonathan Cohen, Jonathan Raiman, Joseph Jennings, Joyjit Daw, Juan Yu, Julio Tapia, Junkeun Yi, Jupinder Parmar, Jyothi Achar, Kari Briski, Kartik Mattoo, Katherine Cheung, Katherine Luna, Keith Wyss, Kevin Shih, Kezhi Kong, Khanh Nguyen, Khushi Bhardwaj, Kirill Buryak, Kirthi Shankar Sivamani, Konstantinos Krommydas, Kris Murphy, Krishna C. Puvvada, Krzysztof Pawelec, Kumar Anik, Laikh Tewari, Laya Sleiman, Leo Du, Leon Derczynski, Li Ding, Lilach Ilan, Lingjie Wu, Lizzie Wei, Luis Vega, Lun Su, Maarten Van Segbroeck, Maer Rodrigues de Melo, Magaret Zhang, Mahan Fathi, Makesh Narsimhan Sreedhar, Makesh Sreedhar, Makesh Tarun Chandran, Manuel Reyes Gomez, Maor Ashkenazi, Marc Cuevas, Marc Romeijn, Margaret Zhang, Mark Cai, Mark Gabel, Markus Kliegl, Martyna Patelka, Maryam Moosaei, Matthew Varacalli, Matvei Novikov, Mauricio Ferrato, Mehrzad Samadi, Melissa Corpuz, Meng Xin, Mengdi Wang, Mengru Wang, Meredith Price, Micah Schaffer, Michael Andersch, Michael Boone, Michael Evans, Michael Z Wang, Miguel Martinez, Mikail Khona, Mike Chrzanowski, Mike Hollinger, Mingyuan Ma, Minseok Lee, Mohammad Dabbah, Mohammad Shoeybi, Mostofa

Patwary, Nabin Mulepati, Nader Khalil, Najeeb Nabwani, Nancy Agarwal, Nanthini Balasubramaniam, Narimane Hennouni, Narsi Kodukula, Natalie Hereth, Nathaniel Pinckney, Nave Assaf, Negar Habibi, Nestor Qin, Neta Zmora, Netanel Haber, Nick Reamaroon, Nickson Quak, Nidhi Bhatia, Nikhil Jukar, Nikki Pope, Nikolai Ludwig, Nima Tajbakhsh, Nir Ailon, Nirmal Juluru, Nirmalya De, Nowel Pitt, Oleg Rybakov, Oleksii Hrinchuk, Oleksii Kuchaiev, Olivier Delalleau, Oluwatobi Olabiyi, Omer Ullman Argov, Omri Almog, Omri Puny, Oren Tropp, Otavio Padovani, Ouye Xie, Parth Chadha, Pasha Shamis, Paul Gibbons, Pavlo Molchanov, Peter Belcak, Peter Jin, Pinky Xu, Piotr Januszewski, Pooya Jannaty, Prachi Shevate, Pradeep Thalasta, Pranav Prashant Thombre, Prasoon Varshney, Prerana Gambhir, Pritam Gundecha, Przemek Tredak, Qing Miao, Qiyu Wan, Quan Tran Minh, Rabeeh Karimi Mahabadi, Rachel Oberman, Rachit Garg, Rahul Kandu, Raina Zhong, Ran El-Yaniv, Ran Zilberstein, Rasoul Shafipour, Renee Yao, Renjie Pi, Richard Mazzarese, Richard Wang, Rick Izzo, Ridhima Singla, Rima Shahbazyan, Rishabh Garg, Ritika Borkar, Ritu Gala, Riyad Islam, Robert Clark, Robert Hesse, Roger Waleffe, Rohit Varma Kalidindi, Rohit Watve, Roi Koren, Ron Fan, Ruchika Kharwar, Ruisi Cai, Ruoxi Zhang, Russell J. Hewett, Ryan Prenger, Ryan Timbrook, Ryota Egashira, Sadegh Mahdavi, Sagar Singh Ashutosh Joshi, Sahil Modi, Samuel Kriman, Sandeep Pombra, Sanjay Kariyappa, Sanjeev Satheesh, Santiago Pombo, Saori Kaji, Satish Pasumarthi, Saurav Mishra, Saurav Muralidharan, Scott Hara, Sean Narenthiran, Sebastian Rogawski, Seonjin Na, Seonmyeong Bak, Sepehr Sameni, Seth Poulos, Shahar Mor, Shantanu Acharya, Shaona Ghosh Adam Lord, Sharath Turuvekere Sreenivas, Shaun Kotek, Shaya Gharghabi, Shelby Thomas, Sheng-Chieh Lin, Shibani Likhite, Shiqing Fan, Shiyang Chen, Shreya Gopal, Shrimai Prabhumoye, Shubham Pachori, Shubham Toshniwal, Shuo Zhang, Shuoyang Ding, Shyam Renjith, Shyamala Prayaga, Siddhartha Jain, Simeng Sun, Sirisha Rella, Sirshak Das, Smita Ithape, Sneha Harishchandra S, Somshubra Majumdar, Soumye Singhal, Sri Harsha Singudasu, Sriharsha Niverty, Stas Sergienko, Stefana Gloginic, Stefania Alborghetti, Stephen Ge, Stephen McCullough, Sugam Dipak Devare, Suguna Varshini Velury, Sukrit Rao, Sumeet Kumar Barua, Sunny Gai, Suseella Panguluri, Sushil Koundinyan, Swathi Patnam, Sweta Priyadarshi, Swetha Bhendigeri, Syeda Nahida Akter, Sylendran Arunagiri, Tailling Yuan, Talor Abramovich, Tan Bui, Tan Yu, Terry Kong, Thanh Do, Thomas Gburek, Thorgane Marques, Tiffany Moore, Tijmen Blankevoort, Tim Moon, Timothy Ma, Tiyasa Mitra, Tomasz Grzegorzek, Tomer Asida, Tomer Bar Natan, Tomer Keren, Tomer Ronen, Traian Rebedea, Trenton Starkey, Tugrul Konuk, Twinkle Vashishth, Tyler Condensa, Udi Karpas, Ushnish De, Vahid Noorozi, Vahid Noroozi, Vanshil Atul Shah, Veena Vaidyanathan, Venkat Srinivasan, Venmugil Elango, Victor Cui, Vijay Korthikanti, Vikas Mehta, Virginia Adams, Virginia Wu, Vitaly Kurin, Vitaly Lavrukhin, Vladimir Anisimov, Wan Seo, Wanli Jiang, Wasi Uddin Ahmad, Wei Du, Wei Ping, Wei-Ming Chen, Wendy Quan, Wenliang Dai, Wenwen Gao, Will Jennings, William Zhang, Xiaowei Ren, Xiaowen Xin, Xin Li, Yang Yu, Yangyi Chen, Yaniv Galron, Yashaswi Karnati, Yejin Choi, Yev Meyer, Yi-Fu Wu, Yian Zhang, Ying Lin, Yonatan Geifman, Yonggan Fu, Yoshi Suhara, Youngeun Kwon, Yuan Zhang, Yuki Huang, Zach Moshe, Zhilin Wang, Zhiyu Cheng, Zhongbo Zhu, Zhuolin Yang, Zihan Liu, Zijia Chen, Zijie Yan, and Zuhair Ahmed. Nemotron 3 super: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning, 2026. URL https://arxiv.org/abs/2604.12374.

Yui Oka, Itsumi Saito, Kyosuke Nishida, and Kuniko Saito. Frequency bands in roPE: Base frequency and context length shape the interpolation–extrapolation trade-off. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PR1PPxvG9Q.

OpenAI. gpt-oss-120b gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

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

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylKikSYDH.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified

text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

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

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. $\infty$Bench: Extending long context evaluation beyond 100K tokens. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.814.

# A EXPERIMENTAL DETAILS

All experiments were conducted on NVIDIA H100 GPUs with 80GB memory. Our experiments do not involve any training. We only feed unlabeled text into open-source pretrained language models, extract their internal representations, and analyze these representations.

## A.1 MODELS

**Mistral-7B** We use `mistralai/Mistral-7B-Instruct-v0.2` (Jiang et al., 2023) [4] as a 7B-scale instruction-tuned autoregressive Transformer. The model supports a maximum sequence length of 32,768 tokens. Positional information is encoded with rotary position embeddings using a RoPE base of $\theta = 10^6$. We do not apply any additional position interpolation or RoPE scaling for this model; instead, we use the long-context setting provided by the released model. The attention module follows grouped-query attention, with 32 query heads and 8 key–value heads. Since the hidden dimension is 4096, the dimension of each query head is 128. The network consists of 32 Transformer layers. The license is Apache-2.0.

**Llama-3.1-8B** We use `meta-llama/Llama-3.1-8B-Instruct` (Grattafiori et al., 2024) [5] as an 8B-scale instruction-tuned model with native long-context support. The model is designed for a maximum sequence length of 131,072 tokens. Its rotary position embeddings use a RoPE base of $\theta = 500000$. For long-context extrapolation, the model uses the Llama 3.1 RoPE scaling scheme, which extends the original 8,192-token positional range by a factor of 8. This scheme rescales different frequency bands non-uniformly: lower-frequency components are stretched more strongly to support long-range positions, while higher-frequency components are preserved more conservatively to reduce degradation on shorter contexts. The model uses grouped-query attention with 32 query heads and 8 key–value heads. With hidden dimension 4096, each query head has dimension 128. The model has 32 Transformer layers. The license is Llama 3.1 Community License.

**Qwen3-8B** We use `Qwen/Qwen3-8B` (Yang et al., 2025) [6] as an 8B-scale dense causal language model. In its default setting, the model is intended for long contexts on the order of 32k tokens, and its rotary position embeddings use a RoPE base of $\theta = 10^6$. For experiments requiring context lengths beyond the native range, we enable YaRN-based RoPE scaling (Peng et al., 2024). YaRN can be viewed as a position-interpolation method that expands the usable positional range by rescaling RoPE frequencies, while attempting to preserve the behavior of the original positional encoding at shorter lengths. In our long-context setting, we use a scaling factor of 4 from an original 32,768-token range, yielding a target context length of 131,072 tokens. The model uses grouped-query attention with 32 query heads and 8 key–value heads. Its hidden dimension is 4096, giving a query-head dimension of 128, and the network contains 36 Transformer layers. The license is Apache-2.0.

**Phi-3-mini** We use `microsoft/Phi-3-mini-128k-instruct` (Abdin et al., 2024) [7] as a 3.8B-scale instruction-tuned autoregressive Transformer. The model supports a maximum sequence length of 131,072 tokens. Positional information is encoded with rotary position embeddings using a RoPE base of $\theta = 10^4$. To enable the 128K-token context length, the released configuration applies LongRoPE scaling (Ding et al., 2024) from an original context length of 4,096 tokens. The attention module uses 32 query heads and 32 key–value heads, corresponding to standard multi-head attention rather than grouped-query attention. Since the hidden dimension is 3072, the dimension of each query head is 96. The network consists of 32 Transformer layers. The license is Apache-2.0.

**gpt-oss-120b** We use `openai/gpt-oss-120b` (OpenAI, 2025) [8] as a large open-weight mixture-of-experts language model. The model supports a maximum sequence length of 131,072 tokens. Positional information is represented with rotary position embeddings using a RoPE base of $\theta = 150000$. Long-context support is obtained through YaRN-style RoPE scaling (Peng et al., 2024):

[4] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2`
[5] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`
[6] `https://huggingface.co/Qwen/Qwen3-8B`
[7] `https://huggingface.co/microsoft/Phi-3-mini-128k-instruct`
[8] `https://huggingface.co/openai/gpt-oss-120b`

the original 4,096-token positional range is expanded by a factor of 32 to reach 131,072 tokens. This scaling modifies the effective RoPE frequencies so that positions beyond the original training range can be represented without directly changing the Transformer architecture. The model uses grouped-query attention with 64 query heads and 8 key–value heads. Each query head has dimension 64, and the model contains 36 Transformer layers. In addition, its attention pattern combines full attention and sliding-window attention (Child et al., 2019) across layers; the sliding-window attention uses a local window of 128 tokens. The license is Apache-2.0.

### A.2 DATASETS

To evaluate model behavior across diverse long-form text domains, we use three English-language corpora: WikiText-103, C4, and PG-19. These datasets cover different types of natural text, including encyclopedic articles, web-crawled documents, and full-length books, allowing us to assess performance under varied long-context settings. We use WikiText-103(Merity et al., 2016), `Salesforce/wikitext` with the `wikitext-103-raw-v1` configuration[9]. WikiText-103 is a large collection of English Wikipedia articles designed for language modeling, and the raw version preserves the original article text with minimal preprocessing. Its document-level structure and coherent long-form text make it suitable for evaluating language models on long-range textual dependencies. The license isCC BY-SA 3.0 / GFDL. We also use the English subset of the C4 dataset (Raffel et al., 2020), `allenai/c4` with English configuration[10]. C4 is a large-scale cleaned English corpus derived from Common Crawl and introduced in the T5 study (Raffel et al., 2020). It was constructed by filtering web-crawled documents to remove duplicate, low-quality, and non-English content. The license is ODC-BY + Common Crawl Terms of Use. The PG-19 dataset (Rae et al., 2020) is a large-scale collection of full-length books extracted from Project Gutenberg. The corpus consists of English-language books published before 1919, and has been widely used as a standard benchmark for evaluating long-range dependency modeling in natural language text. The license is Apache-2.0.

### A.3 INFERENCE SETTINGS

**RoPE-RULER inference settings** For RoPE-RULER, we run each model in inference mode and feed the input sequence without applying any task-specific prompting or supervised labels. During the forward pass, we disable gradient computation and keep the model parameters fixed. For all models, tokenization is performed with the corresponding pretrained tokenizer, and the input is truncated to the target analysis length before being passed to the model. We use deterministic decoding with greedy token selection, temperature $0.0$, top-$p = 1.0$, and no sampling. The maximum number of newly generated tokens is set to 1, since RoPE-RULER only requires the internal query/key representations induced by the input context rather than long-form generation. We extract the query representations from each attention layer and head during this inference pass and compute the RoPE frequency-band statistics described in Section 3.

**RULER settings** RULER (Hsieh et al., 2024) consists of 13 synthetic tasks covering four categories: retrieval, multi-hop tracing, aggregation, and question answering. The license is Apache-2.0. The task set includes eight needle-in-a-haystack variants (`niah_single_1`, `niah_single_2`, `niah_single_3`, `niah_multikey_1`, `niah_multikey_2`, `niah_multikey_3`, `niah_multivalue`, and `niah_multiquery`), one variable-tracking task (`vt`), two aggregation tasks (`cwe` and `fwe`), and two question-answering tasks (`qa_1` and `qa_2`). For each model and each target sequence length, we use 500 examples per task and report the average score over all 13 tasks. The evaluated context-length buckets are 4K, 8K, 16K, 32K, 64K, and 128K tokens. Generation is performed with greedy decoding, using temperature $0.0$, top-$p = 1.0$, and top-$k = 32$. Following RULER, the effective context length is defined as the largest evaluated length at which the average score remains above the qualitative threshold given by the Llama-2-7B 4K score, namely 85.6.

[9] `https://huggingface.co/datasets/Salesforce/wikitext`

[10] `https://huggingface.co/datasets/allenai/c4`

## B Additional Analysis of Dominant RoPE Dimensions

We further analyze whether the observed high-norm frequency bands are specific to query representations or also appear in key representations. Figures 6 and 7 show the layer-wise distributions of dominant RoPE dimensions for the query and key representations of `Llama-3.1-8B`, respectively. For both representations, the dominant dimensions are concentrated in a similar range of RoPE indices across layers and heads. This indicates that the high-norm frequency-band structure is not unique to queries, but is also present in keys. Overall, the query and key representations exhibit no substantial difference in the locations of their dominant RoPE frequency bands.

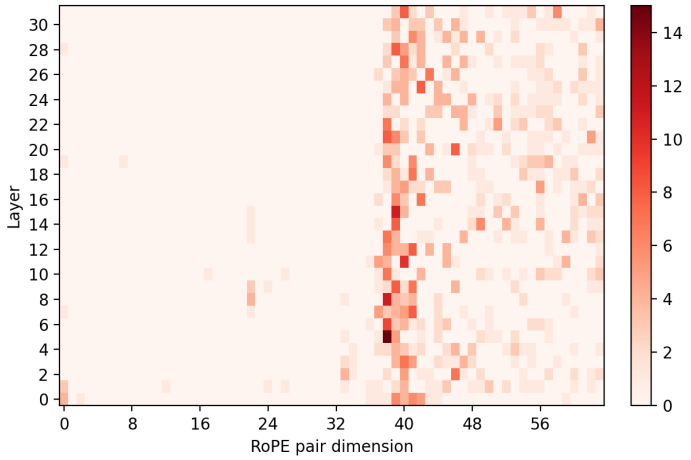



Figure 6: Layer-wise distribution of dominant RoPE dimensions in the **query** representations of `Llama-3.1-8B`. For each layer and head, we identify the RoPE dimension pair $(i, i + d/2)$ with the largest L2 norm over the first 4000 tokens. Each cell shows how frequently each RoPE dimension is selected across heads in that layer; darker colors indicate higher frequency.

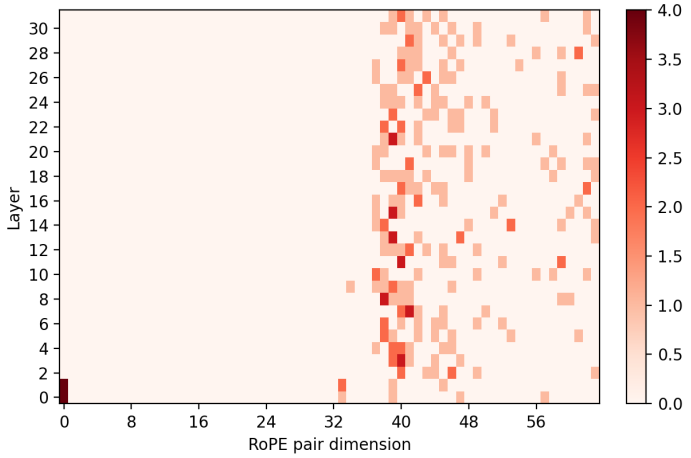



Figure 7: Layer-wise distribution of dominant RoPE dimensions in the **key** representations of `Llama-3.1-8B`. For each layer and head, we identify the RoPE dimension pair $(i, i + d/2)$ with the largest L2 norm over the first 4000 tokens. Each cell shows how frequently each RoPE dimension is selected across heads in that layer; darker colors indicate higher frequency.

## C Robustness to Input Texts and Context Lengths

Figure 8 shows L2-norm heatmaps for `Llama-3.1-8B` under nine evaluation settings, obtained by varying both the input corpus and the context length. In all settings, a high-norm frequency band is clearly visible. Moreover, the band appears at the same RoPE-index location across datasets and context lengths. Thus, the dominant RoPE frequency band does not shift with the choice of input text or the number of tokens used for inference. These results support the view that the frequency band detected by RoPE-RULER is a model-intrinsic spectral feature rather than an input-dependent artifact.

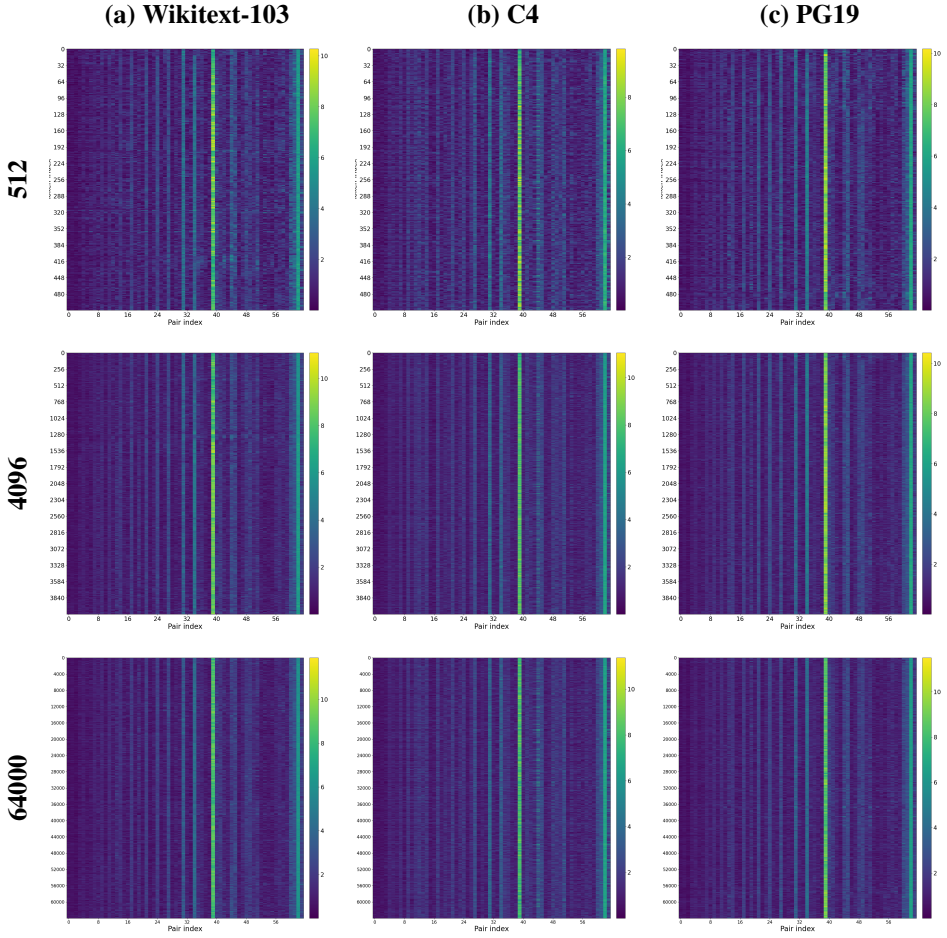



Figure 8: L2-norm heatmaps of Llama-3.1-8B across different datasets and context lengths. Columns correspond to datasets: (a) Wikitext-103, (b) C4, and (c) PG19. Rows correspond to context lengths: 512 (top), 4096 (middle), and 64000 (bottom). The horizontal axis represents the RoPE frequency index (0 to $d/2 - 1$), and the vertical axis represents the token position. Yellow regions indicate frequency bands with consistently high norm values.

Table 5: Robustness analysis for Llama-3.1-8B across input corpora and context lengths. Each entry reports the identified effective frequency $\theta^{\text{eff}}_{i_{\text{band}}}$.

| Input | 0.5k | 4k | 64k |
|---|---|---|---|
| Wiki. | 23762.9 | 23762.9 | 23762.9 |
| C4 | 23762.9 | 23762.9 | 23762.9 |
| PG19 | 23762.9 | 23762.9 | 23762.9 |

# D FUNCTIONAL RELEVANCE OF ROPE FREQUENCY BANDS

Our method builds on a line of prior work showing that RoPE frequency bands are not merely visual artifacts of the query/key representations, but correspond to frequency components that can substantially affect attention computation and model behavior. In this appendix, we clarify the role of these prior results and explain why our experiments focus on estimating the distance scale associated with the dominant band, rather than re-validating the functional importance of RoPE frequency bands through additional intervention experiments.

**Prior intervention evidence.** Prior work has already provided direct evidence that RoPE frequency bands are functionally relevant to model behavior. Barbero et al. (2025) showed that certain high-frequency RoPE components are used to construct positional attention heads, including diagonal and previous-token heads. They also showed theoretically that RoPE can construct such positional attention patterns, whereas an isolated attention head without positional encoding cannot robustly construct the same patterns. In addition, their p-RoPE experiments modify the set of RoPE-rotated dimensions and show that removing rotation from part of the low-frequency range can maintain, and in some settings improve, language-modeling performance. These results indicate that changing which frequency components are rotated is reflected in model performance. Oka et al. (2026) further extended this line of evidence across multiple pretrained LLMs and context-extension methods. They showed that high-norm frequency bands appear consistently across models such as Gemma, Llama, Qwen, and Phi, and that these bands persist after position-interpolation methods such as YaRN, Llama scaling, and LongRoPE. They also used p-RoPE at inference time to test the contribution of different frequency ranges, finding that replacing RoPE dimensions below the observed band with NoPE often leaves perplexity largely unchanged, whereas removing dimensions beyond this range can substantially degrade performance. This provides intervention-based evidence that the frequency band separates weakly used and functionally important RoPE dimensions. Moreover, Oka et al. (Oka et al., 2026) showed that changing the RoPE base frequency $\theta$ and the training context length shifts the band location, and that this shift leads to a measurable interpolation–extrapolation trade-off. These controlled experiments directly connect the frequency-band location to downstream performance.

**Scope of our contribution.** In this paper, we do not aim to re-establish that frequency bands affect model behavior; this has already been supported by theoretical analyses, p-RoPE interventions, and controlled pretraining experiments in prior work. Instead, our contribution is to clarify a previously unexplained role of RoPE frequency bands: their connection to the effective context length of a trained model. Prior work showed that frequency bands are functionally relevant, but it did not determine whether the positional scale of these bands corresponds to the distance range over which a model can actually use contextual information. We address this missing link by showing that the *scale* of the dominant high-norm frequency band corresponds to the effective context length measured by long-context benchmarks.

In other words, we ask a different question: given that pretrained RoPE-based models allocate attention-relevant norm to particular frequency bands, can the positional scale of the dominant band predict the distance over which the model can effectively use context? Our results suggest that the dominant frequency band is not only a functionally relevant component of the attention computation, but also a spectral indicator of the usable context scale of the model.

This distinction is important. Directly masking or rescaling the dominant band in a pretrained model would test whether that band is necessary for the model's current behavior, but such an intervention can introduce distribution shift and may disrupt multiple computations at once, including local positional heads, semantic matching, and long-range retrieval. RoPE-RULER instead uses the learned frequency allocation as a diagnostic signal. It estimates the usable distance scale already encoded in the model's RoPE representation, without modifying the model or requiring task-specific long-context evaluation.

## E BROADER IMPACTS

This work introduces RoPE-RULER, an internal spectral diagnostic for estimating the usable context scale of RoPE-based language models. The main positive impact of this work is to reduce the cost of preliminary long-context analysis. Long-context benchmarks require many inference runs over long sequences, whereas RoPE-RULER can provide a lightweight complementary estimate from internal query/key representations. This may make long-context model analysis more accessible to researchers with limited computational resources.

This work may also improve the transparency of long-context models. Advertised context windows do not always reflect the length over which a model can effectively use information. By connecting effective context length to identifiable RoPE frequency bands, RoPE-RULER provides an interpretable diagnostic for understanding how usable distance scales are represented inside RoPE-based models.

We emphasize that RoPE-RULER is intended as a diagnostic tool, not as a direct method for improving model capability or safety. It does not measure downstream task performance, factual reliability, harmfulness, bias, or robustness. A model with a favorable spectral signature may still fail on tasks requiring retrieval, reasoning, instruction following, or safe behavior over long contexts. Therefore, RoPE-RULER should not be used as a substitute for behavioral evaluation, safety evaluation, or task-specific benchmarking.

A potential risk is that spectral diagnostics could be overinterpreted as a complete measure of long-context capability. To avoid such misuse, we position RoPE-RULER as a complementary analysis tool to be used together with established long-context benchmarks such as RULER and with application-specific evaluations. The method also requires access to internal activations and RoPE configurations, so it is mainly applicable to open or inspectable models.