UNDERSTANDING ROPE THROUGH HEAD-WISE RETRIEVAL DISTANCE PROFILING

Anonymous authors
Paper under double-blind review

ABSTRACT

Rotary Position Embedding (RoPE) is widely used in large language models (LLMs) and is central to context-length extension. However, existing analyses mainly infer the importance of RoPE frequencies from magnitude-based statistics such as the ℓ_2 norms of query and key representations, which do not reveal how attention heads function across different query–key relative positions. We introduce Head-wise Retrieval Distance Profiling (HRDP), which explicitly controls the query–key relative offset and measures, for each attention head, whether the target key achieves the highest query–key score among all preceding keys. HRDP provides an offset-dependent retrieval profile for each attention head. Our analysis reveals that approximately 20% of heads successfully retrieve targets at short offsets, whereas long-range retrieval relies on only a small subset of heads. We further find that better-performing positional interpolation more faithfully preserves the original head-wise retrieval structure, whereas poorer interpolation substantially disrupts it. These results provide a functional view of RoPE beyond frequency magnitude and offer a new basis for analyzing and designing context-extension methods.

1 INTRODUCTION

Rotary Position Embedding (RoPE) (Su et al., 2023) is widely used to encode positional information in Transformer-based large language models (LLMs) (Vaswani et al., 2017). Extending pretrained models to longer contexts often relies on positional interpolation (bloc97, 2023; Chen et al., 2023; Peng et al., 2024; Ding et al., 2024), which modifies the rotation frequencies of RoPE, making the choice of frequency scaling critical to long-context performance. Therefore, it is important to understand how pretrained LLMs effectively use different RoPE frequencies to inform the design of context-extension methods.

Recent studies have begun to characterize how pretrained models utilize RoPE frequencies. Because the RoPE attention logit can be expressed as a function of the relative offset between a query and a key, earlier work analyzed its distance-dependent behavior, including the attenuation of attention scores with increasing relative distance (Xiong et al., 2024). However, Barbero et al. (2025) showed that RoPE does not generally guarantee such distance-dependent decay for query and key representations learned by actual LLMs. Instead, they analyzed the block-wise ℓ_2 norms of query and key representations as a proxy for the potential contribution of each RoPE frequency block to the attention logit. Such magnitude-based analyses have revealed the emergence of high-norm frequency bands (Oka et al., 2026) and the concentration of large norms of query and key in a limited subset of RoPE frequency channels across layers (Gopalakrishnan et al., 2026). In contrast, Wang et al. (2026) reported substantial heterogeneity in frequency utilization across attention heads and further showed, using a synthetic retrieval task, that the utilized frequency range changes with the width of the target window. At first glance, these observations appear to be in tension: magnitude-based analyses suggest that large norms are concentrated in a limited set of frequency channels, whereas head-wise analyses suggest that the frequency ranges functionally utilized by the model vary across heads and positional scales. Therefore, how these two observations relate to each other remains unclear.

From the perspective of attention-head function, Wu et al. (2025) showed that long-context retrieval is concentrated in a sparse subset of attention heads, termed Retrieval Heads. However, their retrieval score aggregates behavior across multiple context lengths and needle positions, and therefore does

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.

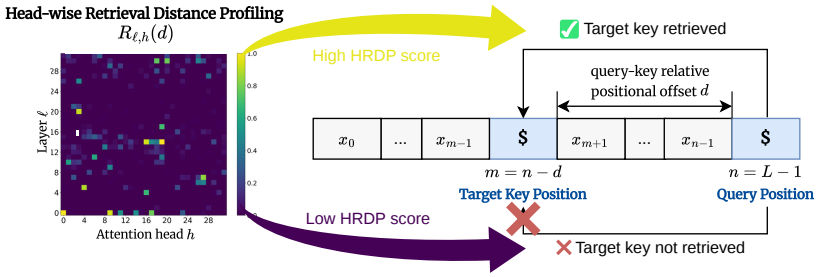


Figure 1: An overview of Head-wise Retrieval Distance Profiling (HRDP). HRDP measures, for each layer and attention head, how reliably the target key at relative offset d is retrieved, yielding the offset-dependent score $R_{\ell,h}(d)$.

not directly characterize how the retrieval capability of an individual head varies as a function of the exact query–key relative offset. Thus, although prior work has characterized frequency magnitude, head-specific frequency utilization, and the attention heads responsible for retrieval, it also remains unclear *which heads can retrieve a target key at a specific query–key relative offset and which RoPE frequency channels are functionally required for retrieval at that offset*.

To address this gap, we introduce **Head-wise Retrieval Distance Profiling (HRDP)**, which profiles retrieval as a function of the query–key relative positional offset. Figure 1 provides an overview of the proposed analysis framework. We precisely control the relative offset (d) between a query and its corresponding target key and, for each layer and attention head, measure the fraction of examples in which the target key attains the highest query–key score among all preceding keys. Rather than assigning each head a single retrieval score, this framework produces an offset-dependent retrieval profile that reveals which heads can retrieve target keys at different relative positions.

Our analysis reveals a clear offset-dependent structure in head-wise retrieval. At short relative offsets, approximately 20% of attention heads successfully retrieve the target key, whereas the number of successful heads rapidly decreases as the offset grows, eventually concentrating on only a small subset of heads at long distances. Moreover, within the pretrained context range, the head distribution at a fixed query–key offset remains largely stable as the overall sequence length changes. This suggests that the observed structure reflects a functional sensitivity of individual heads to relative position, rather than merely a response to input length. Furthermore, by selectively disabling specific frequency bands of RoPE, we found that retrieval cannot be explained by a simple correlation such as "low frequency = long range" or "high frequency = short range," and that both low and high frequencies affect distances across a range. In previous studies (Barbero et al., 2025; Oka et al., 2026; Gopalakrishnan et al., 2026), the ℓ_2 norms of query and key have been used as a metric to identify important RoPE frequencies; however, our results show that the magnitude of a frequency does not necessarily reflect its functional importance in retrieval. Finally, we use HRDP to analyze positional interpolation. In our comparisons, interpolation methods with stronger long-context retrieval performance better preserve the pretrained model's original head-wise retrieval structure, whereas uniform interpolation substantially alters it. We further find that even within the same interpolation method, aggressive parameter choices can systematically distort the retrieval profiles.

Overall, our analysis connects RoPE frequency utilization to attention-head retrieval through query–key relative position. Our findings suggest that understanding RoPE and positional interpolation requires going beyond frequency magnitude to characterize how individual heads functionally use different frequency components for retrieval across relative positions.

2 Background

RoPE. RoPE (Su et al., 2023) is widely adopted in recent LLMs, including Llama (Touvron et al., 2023; Grattafiori et al., 2024), Qwen (Yang et al., 2025), and Gemma (Team et al., 2024). RoPE encodes positional information by rotating the query and key representations according to their absolute positions. For a query $\mathbf{q}_n \in \mathbb{R}^d$ at position n and a key $\mathbf{k}_m \in \mathbb{R}^d$ at position m, the attention

logit is written as

$$A_{m,n} = (\mathbf{R}_{n,\theta}\boldsymbol{q}_n)^\top(\mathbf{R}_{m,\theta}\boldsymbol{k}_m) = \boldsymbol{q}_n^\top\mathbf{R}_{m-n,\theta}\boldsymbol{k}_m, \tag{1}$$

where $R_{n,\theta}$ is a block-diagonal rotation matrix consisting of $d/2$ two-dimensional rotations:

$$\mathbf{R}_{n,\theta} = \operatorname{diag}\left(\mathbf{R}(n\theta_0), \ldots, \mathbf{R}(n\theta_{d/2-1})\right), \qquad \mathbf{R}(\phi) = \begin{bmatrix}\cos\phi & -\sin\phi \\ \sin\phi & \cos\phi\end{bmatrix}. \tag{2}$$

The rotation frequency assigned to the j-th two-dimensional RoPE block is defined as $\theta_j = \theta^{-2j/d}$ for $j = 0, \ldots, d/2-1$, where d is the head dimension and θ is the RoPE base. Because $\mathbf{R}_{n,\theta}^\top\mathbf{R}_{m,\theta} = \mathbf{R}_{m-n,\theta}$, the resulting attention score depends on the relative positional offset $m-n$, despite the rotations being specified by absolute token positions.

Position Interpolation for Long Context. RoPE-based context extension has often been motivated by scaling down position indices to keep longer positions within the pretrained rotation range. Position Interpolation (PI) (Chen et al., 2023) uniformly rescales all RoPE frequencies, whereas NTK-aware interpolation (bloc97, 2023) modifies the RoPE base to produce dimension-dependent scaling that preserves higher frequencies while more strongly reducing lower ones. YaRN (Peng et al., 2024), one of the most widely adopted RoPE-based context-extension methods, further applies frequency-dependent interpolation: high frequencies are largely preserved, low frequencies are fully interpolated, and intermediate frequencies are smoothly blended. More recent methods adopt finer-grained RoPE scaling, including dimension-wise scaling (Ding et al., 2024; Tian et al., 2026; Lu et al., 2025) and head-specific scaling (Wang et al., 2026).

RoPE Frequency. Given its impact on long-context performance, the choice of the RoPE base θ during pretraining has also received increasing attention. Based on earlier findings that increasing the RoPE base can improve long-context performance (Xiong et al., 2024), lower effective rotation frequencies have often been associated with better long-range behavior. However, Barbero et al. (2025) showed that RoPE does not generally guarantee the assumed distance-dependent decay and instead analyzed frequency usage through the ℓ_2 norms of query and key components. By the Cauchy–Schwarz inequality, the contribution of the j-th RoPE frequency block to the attention logit is bounded as

$$\left|\left\langle \mathbf{R}_{n,\theta}^{(j)}\boldsymbol{q}_n^{(j)}, \mathbf{R}_{m,\theta}^{(j)}\boldsymbol{k}_m^{(j)}\right\rangle\right| \le \|\mathbf{R}_{n,\theta}^{(j)}\boldsymbol{q}_n^{(j)}\|_2\|\mathbf{R}_{m,\theta}^{(j)}\boldsymbol{k}_m^{(j)}\|_2. \tag{3}$$

Thus, the block-wise ℓ_2 norms provide a proxy for the potential contribution of each RoPE frequency to the attention logit. This line of analysis has revealed that different frequency ranges are associated with distinct attention patterns (Barbero et al., 2025), that query and key norms concentrate in specific RoPE frequency bands (Gopalakrishnan et al., 2026), and that these bands are determined by the RoPE base and training sequence length and remain largely preserved after context extension with positional interpolation (Oka et al., 2026). Jin et al. (2025) further reports that analogous structured concentration is not observed in value representations. However, ℓ_2 norms characterize only the potential contribution of a frequency to the attention logit; they do not reveal whether an individual head can actually retrieve a target key at a particular query–key relative offset.

Head-wise Retrieval and Positional Sensitivity. Prior work has shown that attention heads exhibit specialized retrieval and positional behaviors. Wu et al. (2025) identified a sparse set of Retrieval Heads that are primarily responsible for retrieving information from long contexts. Positional specialization has also been observed at the attention-head level: Barbero et al. (2025) showed that RoPE can support purely positional attention patterns, including diagonal and previous-token heads. More recently, Wang et al. (2026) demonstrated substantial heterogeneity in frequency utilization and effective attention spans across heads, and showed that the useful frequency range changes with the width of a target window in a synthetic retrieval task. However, these analyses do not directly characterize whether an individual pretrained attention head can retrieve a specific target key at a precisely controlled query–key relative offset.

3 Head-wise Retrieval Distance Profiling

Existing analyses do not directly characterize how the retrieval capability of an individual attention head changes with the relative position between a query and a key. Query–key relative offset is a particularly important variable for analyzing RoPE because the RoPE attention logit is explicitly

a function of the relative position between a query and a key. Evaluations based only on context length, needle position, or target-window width therefore conflate the distance of the target from the query with other properties of the input. We therefore introduce a controlled evaluation that measures, for each attention head, whether its query–key matching mechanism can identify a target key among all preceding keys as a function of the query–key relative positional offset. Our evaluation is intentionally defined at the attention-logit level rather than from the final model output. It measures the retrieval capability encoded in query–key matching: whether a head assigns the highest query–key score to the target key over all competing past keys. HRDP is computed in two steps and produces a layer-by-head heatmap of retrieval scores for each query–key relative offset.

Step 1: Controlled Query–Key Retrieval. Given a token sequence of length L, we place the query at the final position $n = L - 1$ and the target key at $m = n - d$, where d denotes the *query–key relative positional offset*. We insert the same single-token marker \$ at positions m and n and ensure that \$ does not occur elsewhere in the sequence. All other preceding tokens therefore serve as distractor keys. By varying d, we can measure how the retrieval capability of each head changes with the query–key offset. Importantly, d denotes the token-level offset between the query and target key, rather than the overall sequence length or the relative depth of a needle in the context.

Step 2: Head-wise Retrieval Score. For layer ℓ and attention head h, let $q_n^{\ell,h}$ denote the query at position n and $k_t^{\ell,h}$ the key at a preceding position $t < n$. We compute

$$S_{\ell,h}(n,t) = \left(\mathbf{R}_{n,\theta}\boldsymbol{q}_n^{\ell,h}\right)^\top \mathbf{R}_{t,\theta}\boldsymbol{k}_t^{\ell,h}. \tag{4}$$

We omit the standard attention scaling factor $1/\sqrt{d_h}$ (Vaswani et al., 2017) because it is constant within a head and therefore does not affect the ranking of candidate keys. For an evaluation example i, retrieval is successful when the target key at position m achieves the highest query–key score among all preceding keys:

$$H_{\ell,h}^{(i)}(d) = \mathbb{I}\left[m = \arg\max_{0\le t<n} S_{\ell,h}^{(i)}(n,t)\right]. \tag{5}$$

where $\mathbb{I}[\cdot]$ is 1 if the condition holds and 0 otherwise. The head-wise retrieval score at offset d is then

$$R_{\ell,h}(d) = \frac{1}{N}\sum_{i=1}^{N} H_{\ell,h}^{(i)}(d), \tag{6}$$

where N is the number of evaluation examples. Because softmax is monotonic, the key with the largest query–key score also receives the largest attention logit within the head. Thus, $R_{\ell,h}(d)$ directly measures how consistently head (ℓ, h) retrieves the target key at relative offset d. Evaluating the score over multiple values of d and L yields retrieval profiles across relative offsets and sequence lengths, allowing us to characterize which heads support retrieval under different positional conditions.[1]

4 Experiments

4.1 Experimental Settings

We primarly evaluate Llama-3-8B (Grattafiori et al., 2024), whose pretrained context length is 8,192 tokens. [2] As the base text into which the marker token \$ is inserted, we use the test split of WikiText-103 (Merity et al., 2016). We evaluate $N = 100$ samples for each configuration, with sequence lengths $L \in \{4{,}096, 8{,}192, 16{,}384, 32{,}768\}$ and query–key relative positional offset $d \in \{2, 8, 32, 128, 512, 2048\}$. For each (L, d) pair, we compute $R_{\ell,h}(d)$ for every layer and attention

[1] We use exact Top-1 matching to assess whether a head ranks the designated target key highest among all competing keys, rather than merely localizing its neighborhood. Allowing positional tolerance would credit selections at offsets other than d. For example, a one-token tolerance at $d = 2$ would count a fixed previous-token head as successful. Therefore, this criterion directly measures target selection at the specified offset without introducing either a tolerance window or an absolute logit threshold. Maximal-attention criteria have also been used to identify head-level copy-paste behavior (Wu et al., 2025).

[2] We conduct the same analysis on Qwen3-8B[11], Qwen3-4B[12], Gemma-7B[13] and Gemma4-E2B[14], with the results reported in Appendix D. We also analyze sliding-window attention.

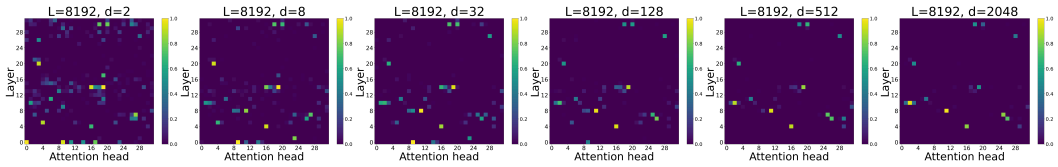


Figure 2: HRDP scores $R_{\ell,h}(d)$ at $L = 8{,}192$ for relative offsets $d \in \{2, 8, 32, 128, 512, 2048\}$, shown as heatmaps over layers ℓ and attention heads h.

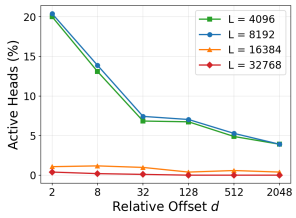


Figure 3: Percentage of active heads across relative offsets d.

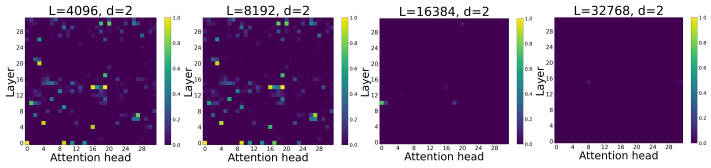


Figure 4: HRDP scores $R_{\ell,h}(d)$ at relative offset $d = 2$ for sequence lengths $L \in \{4\text{k}, 8\text{k}, 16\text{k}, 32\text{k}\}$, shown as heatmaps over layers ℓ and attention heads h.

head. We visualize the resulting Top-1 retrieval score over the 100 samples as a layer-by-head heatmap, where each cell represents the retrieval score of a particular attention head. [3]

4.2 Results

Distance-wise Results. Figure 2 shows the head-wise retrieval scores for different relative offsets d for the sequence length $L = 8{,}192$. As d increases, retrieval becomes concentrated in a progressively smaller subset of heads. Figure 3 quantifies this trend for the sequences $L \in \{4\text{k}, 8\text{k}, 16\text{k}, 32\text{k}\}$: when L falls within the pre-training context length the percentage of active heads, defined as heads that successfully retrieve the target at least once, decreases from approximately 20% at $d = 2$ to about 5% at $d = 2{,}048$, whereas for L beyond the pre-training length the percentage remains below 1% regardless of d. These results indicate head-specific distance sensitivity: some heads support short-range retrieval, while only a few remain effective at longer distances.

Sequnece Length-wise Results. Figure 4 shows the head-wise retrieval scores with the relative offset fixed to $d = 2$ while varying the sequence length over $L \in \{4{,}096, 8{,}192, 16{,}384, 32{,}768\}$. Within the pretrained context length of Llama-3-8B, the retrieval-head distribution is largely preserved between $L = 4{,}096$ and $8{,}192$. In contrast, once the sequence length exceeds the pretrained limit, the distribution collapses substantially, and retrieval score drops to nearly zero for most heads at $L = 16{,}384$ and $32{,}768$. Notably, this degradation occurs even though the query–key relative offset remains very short at $d = 2$. Although RoPE expresses attention scores in terms of relative position, this result suggests that such relative-position structure alone is insufficient to preserve head-wise retrieval behavior beyond the pretrained context range. In other words, retrieval capability appears to depend not only on the relative offset, but also on whether the query and key lie within the absolute-position range encountered during pretraining.

End-to-End Retrieval Performance. Because HRDP is a controlled head-level probe of query–key matching, we next test whether the same failure mode appears in an end-to-end retrieval task. The evaluation follows a Needle-in-a-Haystack-style setup (gkamradt, 2023), with the retrieval target and its position aligned with those used in HRDP. We evaluate 100 samples for each setting and report the resulting accuracy as a heatmap.[4] Figure 5 reports end-to-end marker retrieval accuracy based on the model's final next-token prediction. Retrieval accuracy remains stable within the pretrained context range but drops sharply beyond it, closely mirroring the HRDP degradation even at fixed relative offsets. This correspondence suggests that the extrapolation failure revealed by HRDP reflects a behaviorally relevant retrieval limitation rather than

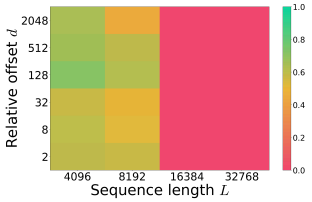


Figure 5: End-to-end QA accuracy.

[3] See Appendix A.1 for detailed experimental settings.

[4] Detailed experimental settings are provided in Appendix A.3.

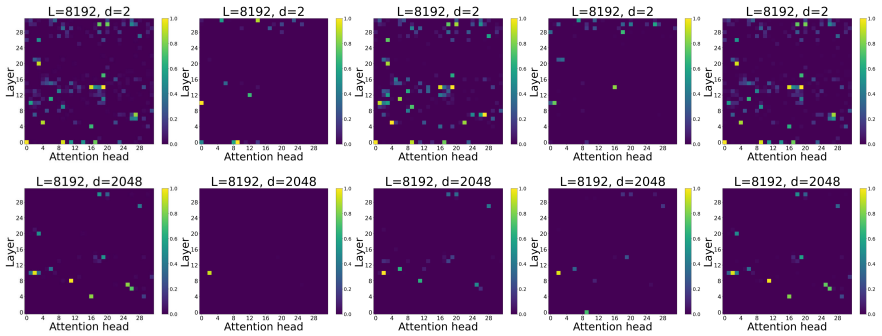


Figure 6: HRDP scores $R_{\ell,h}(d)$ after ablating each RoPE frequency band with NoPE at relative offsets $d \in \{2, 2{,}048\}$, shown as heatmaps over layers ℓ and attention heads h.

an artifact of the query–key-level probe. Because Llama-3-8B is not instruction-tuned, we focus on relative degradation across context lengths rather than absolute QA accuracy.

4.3 Frequency Ablation

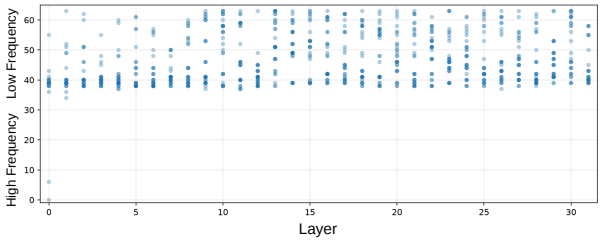
Next, we investigate whether frequency magnitude reflects functional importance for retrieval. We partition the 64 RoPE frequency blocks of Llama-3-8B into four equal-sized contiguous bands (`high`, `mid-high`, `mid-low`, and `low`), ordered from high to low frequency. We then ablate each band independently by disabling RoPE for that band while leaving the remaining bands unchanged. We fix the sequence length to $L = 8{,}192$ and compare HRDP heatmaps at $d = 2$ and $d = 2{,}048$ to examine which frequency ranges are required for short- and long-range retrieval.

Results. Figure 6 shows the HRDP results after ablating each frequency band. Ablating the `high` or `mid-low` band substantially reduces the number of active heads at both $d = 2$ and $d = 2{,}048$, indicating that these frequency ranges contribute to both short- and long-range retrieval. In contrast, ablating the `low` band produces little change in the head-wise retrieval pattern, suggesting only a limited contribution from the lowest-frequency components. Ablating the `mid-high` band has a smaller overall effect, but suppresses retrieval in a subset of heads, suggesting a more head-specific contribution. These results challenge the common intuition that lower RoPE frequencies primarily support longer-range dependencies (Xiong et al., 2024). Even at $d = 2{,}048$, ablating higher-frequency bands substantially degrades retrieval, whereas removing the lowest-frequency band has little effect. Therefore, retrieval distance cannot be explained by a simple frequency–distance correspondence; rather, multiple frequency ranges jointly support retrieval across distances.

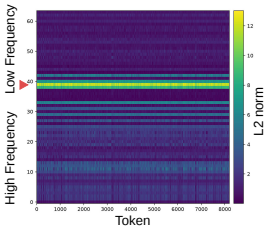
4.4 Comparison with ℓ_2 Norm Analysis

The frequency ablation results (§4.3) suggest that a simple frequency–distance correspondence does not directly explain functional importance. This motivates us to examine whether the ℓ_2-norm-based view adopted in prior work is consistent with the head-wise retrieval behavior revealed by HRDP. Specifically, we ask: *Can RoPE frequency magnitude, as measured by ℓ_2 norms, explain the retrieval distance of individual attention heads?* To answer this question, we compare the ℓ_2-norm-based analyses used in prior work (Barbero et al., 2025; Oka et al., 2026; Gopalakrishnan et al., 2026) against our HRDP results. In this section, we examine the head-wise distribution of ℓ_2 magnitude and the norm-dominant RoPE frequency, which has been treated as an important indicator of frequency usage in prior studies.

ℓ_2 norm distribution. RoPE decomposes each query and key into $d/2$ two-dimensional blocks, with the j-th block corresponding to frequency θ_j. By the Cauchy–Schwarz inequality, its contribution to the query–key dot product satisfies Eq (3). Thus, block-wise ℓ_2 norms provide a proxy for the potential contribution of each RoPE frequency to the attention logit. We aggregate these query ℓ_2 norms, $\|\mathbf{R}_{n,\theta}^{(j)} \boldsymbol{q}_n^{(j)}\|_2$, over token positions and samples to obtain a frequency-magnitude profile for each attention head.



(a) Norm-dominant RoPE frequency $j_{\ell,h}$ for each attention head and layer. The x-axis denotes the layer index, and the y-axis the RoPE frequency. Each point marks the frequency with the maximum ℓ_2 norm (the dimension marked by ▶ in (b)).



(b) Example of ℓ_2-norm distribution for head 0 in the 31st layer, ordered from high to low frequency from left to right.

Figure 7: Head-wise analysis of RoPE frequency magnitude in Llama-3-8B.

Norm-dominant RoPE frequency $j_{\ell,h}$. For each layer ℓ and attention head h, we define the norm-dominant frequency index $j_{\ell,h}$ as the frequency block that consistently exhibits the largest ℓ_2 norm across token positions. We calculate $j_{\ell,h}$ for all attention heads in each layer. For each layer, head, and frequency block j, we aggregate the norms over token positions and evaluation samples to obtain a head-wise frequency-magnitude profile.

Results. Figure 7 shows the norm-dominant RoPE frequencies and an example of the ℓ_2-norm distribution in Llama-3-8B. [5] Large norms are predominantly concentrated in the `mid-low` to `low` frequency ranges. However, this magnitude concentration does not align with the functional retrieval behavior revealed by HRDP. As shown in § 4.2, different subsets of heads remain active at different relative offsets, despite exhibiting similar norm concentration. Moreover, our frequency ablation shows that removing the lowest-frequency band has little effect on retrieval, whereas ablating higher-frequency bands substantially affects retrieval at both short and long offsets. These results indicate that ℓ_2-norm magnitude alone does not reliably identify which RoPE frequencies are functionally important for retrieval, nor does it explain the retrieval distance supported by individual attention heads.

Key Findings from Head-wise Retrieval Distance Profiling.

- ✓ Retrieval concentrates in fewer heads as the query–key offset increases, and retrieval degrades beyond the pretrained context length, even at short offsets (§4.2).
- ✓ Higher-frequency RoPE rotations contribute to retrieval at both $d = 2$ and $d = 2{,}048$, challenging a simple frequency–distance correspondence (§ 4.3).
- ✓ Despite being widely used in prior analyses, ℓ_2-norm magnitude does not reliably reflect the functional importance of a frequency for retrieval (§4.4).

5 Analyzing Positional Interpolation with HRDP

Positional interpolation (Chen et al., 2023; bloc97, 2023; Peng et al., 2024; Ding et al., 2024) modifies RoPE frequencies to extend a pretrained model beyond its original context length. While existing methods are typically compared using downstream long-context performance, it remains unclear how these frequency transformations affect the retrieval behavior already learned by individual attention heads. We therefore use HRDP to examine how positional interpolation changes the pretrained model's head-wise retrieval structure.

5.1 How Does Positional Interpolation Affect HRDP?

We compare the original Llama-3-8B model with RoPE-based context-extension methods using Linear Position Interpolation (Chen et al., 2023) and YaRN (Peng et al., 2024) [6]. For each method, we compute HRDP under the same query–key offsets as in § 4.2 and compare the resulting head-wise retrieval profiles with those of the original model.

[5] The complete ℓ_2-norm distributions for all attention heads are provided in Appendix C.

[6] A detailed description of YaRN and Linear PI is provided in Appendix A.4.

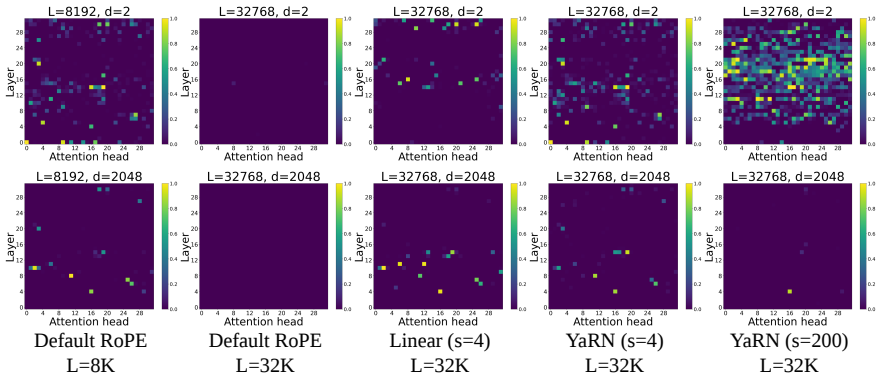


Figure 8: HRDP scores $R_{\ell,h}(d)$ for the default RoPE and positional interpolation methods at relative offsets $d = 2$ (top) and $d = 2{,}048$ (bottom), shown as heatmaps over layers ℓ and attention heads h. From left to right: Default ($L = 8{,}192$), Default ($L = 32{,}768$), Linear PI with $s = 4$, YaRN with $s = 4$, and YaRN with $s = 200$, all evaluated at $L = 32{,}768$ except the first.

Experimental Settings. We evaluate Llama-3-8B, pretrained with a context length of 8,192, at $L = 32{,}768$ using Default RoPE, Linear PI, and YaRN. For Linear PI and YaRN, we use a context-extension scale factor of $s = 4$, and additionally include YaRN with $s = 200$ as an extreme setting. We compare HRDP at $d = 2$ and $d = 2{,}048$.

Results. Figure 8 shows that positional interpolation can substantially change which attention heads retain their retrieval capability. Without positional interpolation, the original RoPE model exhibits almost no active retrieval heads at $L = 32{,}768$ across all tested relative offsets. In contrast, applying YaRN with a scale factor of $s = 8$ largely restores the head-wise retrieval patterns observed at $L = 8{,}192$. This suggests that effective positional interpolation can recover the pretrained retrieval structure even when the model is evaluated far beyond its original context length. Uniform interpolation with Linear PI alters the pretrained retrieval pattern even at short relative offsets. In contrast, YaRN, which achieves stronger long-context performance, more faithfully preserves the original head-wise retrieval structure. However, when the YaRN context-length extension scale factor s is set to an extreme value of 200, the original retrieval structure is substantially disrupted. At $d = 2$, retrieval becomes broadly activated across nearly all heads, whereas the heads responsible for long-range retrieval at $d = 2{,}048$ largely lose their retrieval capability. These results suggest that successful context extension is not simply a matter of enabling RoPE to represent larger positional indices. Rather, effective interpolation should preserve the retrieval behavior already encoded in pretrained attention heads while extending their functionality to longer contexts.

5.2 Which YaRN Parameters Best Preserve Retrieval?

The analysis in § 5.1 shows that the YaRN scale factor can substantially alter the pretrained head-wise retrieval structure, and that overly aggressive scaling can distort retrieval across relative offsets. This raises a natural question: *can preservation of the pretrained retrieval structure itself serve as a criterion for selecting the interpolation parameter?* We therefore investigate whether the YaRN scale factor that best preserves the original HRDP profile also leads to stronger long-context performance.

HRDP-based Parameter Selection. We select the YaRN scale factor s that best preserves the head-wise retrieval structure learned during pre-training. As in Section 4.1, we use WikiText-103 as the input. We first compute the reference head-wise retrieval scores $R^{\text{base}}_{\ell,h}(d; L = 8{,}192)$ using default RoPE at $L = 8{,}192$ for each relative offset $d \in \mathcal{D} = \{2, 8, 32, 128, 512, 2048\}$. We then evaluate YaRN at $L = 32{,}768$ for each candidate scale factor $s \in \mathcal{S} = \{2.0, 2.1, \ldots, 8.0\}$, obtaining $R^{(s)}_{\ell,h}(d; L = 32{,}768)$. For each candidate s, we compute the mean squared error (MSE) between the 32K YaRN retrieval profile and the 8K Default RoPE reference profile over all relative offsets, layers, and attention heads:

$$\text{MSE}\left(R^{(s)}_{32\text{K}}, R^{\text{base}}_{8\text{K}}\right) = \frac{1}{|\mathcal{D}|\,|\mathcal{L}|\,|\mathcal{H}|} \sum_{d\in\mathcal{D}} \sum_{\ell\in\mathcal{L}} \sum_{h\in\mathcal{H}} \left(R^{(s)}_{\ell,h}(d; L = 32768) - R^{\text{base}}_{\ell,h}(d; L = 8192)\right)^2, \tag{7}$$

Table 1: RULER scores.

	8K	16K	32K
Llama-3-8B	90.05	0	0
YaRN ($s = 4$)	90.10	86.09	47.94
YaRN ($s^* = 6.3$)	**90.45**	**86.32**	**62.99**

Table 2: Perplexity on the WikiText-103 test set.

	8K	16K	32K
Llama-3-8B	**6.17**	110.9	1136.7
YaRN ($s = 4$)	6.32	4.96	5.61
YaRN ($s^* = 6.3$)	6.27	**4.94**	**5.55**

Finally, we select

$$s^* = \arg\min_{s \in \mathcal{S}} \mathrm{MSE}\Big(R^{(s)}_{32\mathrm{K}}, R^{\mathrm{base}}_{8\mathrm{K}}\Big). \tag{8}$$

where s^* denotes the YaRN scale factor whose head-wise retrieval profile at 32K most closely matches the default RoPE profile observed at the pre-training context length of 8K.

Experimental Settings. We fine-tune and evaluate Llama-3-8B with YaRN at $s = 4$ and the HRDP-selected $s^* = 6.3$ for 32K context extension. [7] When extending the context length from 8K to 32K, a scale factor of 4 is typically used (Peng et al., 2024; Yang et al., 2025). We evaluate long-context performance using the RULER benchmark (Hsieh et al., 2024)[8] at context lengths of 8K, 16K, and 32K. We also evaluate perplexity on WikiText-103 test set (Merity et al., 2016).

Results. Table 1 shows the RULER results. The original Llama model without context extension achieves a score of 0 once the sequence length exceeds its maximum context length of 8K. YaRN with $s = 4$ works at 32K, while the HRDP-selected factor $s^* = 6.3$ consistently improves RULER performance at all evaluated lengths. Table 2 reports the perplexity results. The original Llama-3-8B performs best at 8K, whereas YaRN with $s^* = 6.3$ achieves the lowest perplexity at 16K and 32K. These results suggest that preserving the pretrained head-wise retrieval structure provides a useful signal for selecting YaRN parameters for context extension.

Key Findings from Positional Interpolation Analysis.

- ✓ Effective positional interpolation restores and preserves the pretrained head-wise retrieval structure beyond the original context length, whereas overly aggressive scaling disrupts it (§5.1).
- ✓ Preserving the pretrained HRDP profile provides a useful criterion for parameter selection: the HRDP-selected YaRN factor $s = 6.3$ achieves the best 32K RULER performance (§5.2).

6 Discussion and Conclusion

Our results show that long-context retrieval in RoPE-based models cannot be fully explained by assigning fixed roles to particular frequencies or Retrieval Heads (§4.2). While prior work has shown that retrieval is concentrated in a small subset of attention heads (Wu et al., 2025), HRDP further reveals that each head exhibits a distinct retrieval profile over query–key relative offsets. This distance-dependent view is also consistent with prior observations that different heads rely on different RoPE frequency ranges (Wang et al., 2026). Our frequency ablations further show that neither a simple frequency–distance correspondence nor frequency magnitude alone explains retrieval behavior (§4.3,§4.4). Thus, understanding RoPE requires a functional analysis of how modifying specific frequencies changes retrieval at the head level, rather than relying only on representation-level statistics such as ℓ_2 norms (Barbero et al., 2025; Oka et al., 2026; Gopalakrishnan et al., 2026).

Moreover, our results connect this functional view to context extension. Retrieval can fail beyond the pretrained context even at short relative offsets, indicating that extrapolating relative positions is not equivalent to preserving the retrieval computation learned during pretraining (§4.2). Positional interpolation works better when the pretrained HRDP structure is better preserved, and the HRDP-selected YaRN scale achieves the strongest 32K performance among the evaluated settings (§5.2). These findings suggest that context extension can be viewed not only as a problem of scaling RoPE frequencies, but also as one of preserving pretrained retrieval functions in a new positional regime.

Therefore, HRDP provides a common functional framework for connecting Retrieval Heads, RoPE frequency utilization, and context extension, while also offering a practical signal for evaluating and selecting context-extension configurations.

[7] Detailed settings for fine-tuning during context extension are provided in Appendix A.2.

[8] Details of the RULER benchmark are provided in Appendix B.

7 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

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/.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=OUIFPHEgJU.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. LongRoPE: Extending LLM context window beyond 2 million tokens. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11091–11104. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ding24i.html.

gkamradt. Needle in a haystack, 2023. URL https://github.com/gkamradt/needle-in-a-haystack.

Anand Gopalakrishnan, Róbert Csordás, Jürgen Schmidhuber, and Michael Curtis Mozer. Decoupling the "what" and "where" with polar coordinate positional embedding. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=I3Z9za1EkO.

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.

Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. Massive values in self-attention modules are the key to contextual knowledge understanding. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=1SMcxxQiSL.

Yi Lu, Wanxu Zhao, Xin Zhou, Chenxin An, Chenglong Wang, Shuo Li, Yuming Yang, Jun Zhao, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Effective length extrapolation via dimension-wise positional embeddings manipulation. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Tahpc3iAnO.

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

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.

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.

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.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau,

Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024. URL https://arxiv.org/abs/2403.08295.

Qingyuan Tian, Wenhong Zhu, Xiaoran Liu, Xiaofeng Wang, and Rui Wang. MrroPE: Mixed-radix rotary position embedding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1J63FJYJKg.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Shaowen Wang, Yuke Zheng, Tansheng Zhu, Shuang Chen, Shaofan Liu, Suncong Zheng, and Jian Li. AdaroPE: Not all attention heads should rotate and scale equally. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=xNuMtLaxo5.

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

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EytBpUGB1Z.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. Effective long-context scaling of foundation models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies (Volume 1: Long Papers), pp. 4643–4663, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.260. URL https://aclanthology.org/2024.naacl-long.260/.

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.

A Experimental Settings

In this section, we describe the models we analyze, the position-interpolation configurations applied to them. All models are obtained from the HuggingFace Hub, and all experiments are implemented on top of the HuggingFace `transformers` library (Wolf et al., 2020)[9].

A.1 Model Details

We analyze Llama-3-8B[10], whose pretrained context length is $L_{\text{train}} = 8{,}192$ tokens and whose RoPE base frequency is $\theta_{\text{base}} = 5 \times 10^5$. To examine how far these observations generalize, we repeat the same analysis on Qwen3-8B[11], Qwen3-4B[12], Gemma-7B [13], and Gemma4-E2B[14].

Llama-3-8B, Qwen3-8B, and Qwen3-4B use Grouped-Query Attention (GQA) with 32 query heads and 8 key/value heads (group size 4), and Gemma4-E2B likewise uses GQA but with 8 query heads and a single key/value head (group size 8), whereas Gemma-7B uses standard multi-head attention. Since our analysis is defined per query head, we expand the key/value heads to the query heads (i.e., `repeat_kv` in `transformers`)) before computing per-head scores; throughout the paper a head index therefore refers to a query head. Note that Qwen3-8B / Qwen3-4B are natively pretrained beyond our longest evaluation length ($L_{\text{train}} = 40{,}960 > 32{,}768$), so for these models context extension is not an extrapolation regime.

Table 3: Model Parameters.

Model	Attn	L_{train}	θ_{base}	Layers	Heads (Q/KV)	d_{head}
`meta-llama/Meta-Llama-3-8B`	Full	8,192	5.0×10^5	32	32/8 (GQA)	128
`Qwen/Qwen3-8B`	Full	40,960	1.0×10^6	36	32/8 (GQA)	128
`Qwen/Qwen3-4B`	Full	40,960	1.0×10^6	36	32/8 (GQA)	128
`google/gemma-7b`	Full	8,192	1.0×10^4	28	16/16 (MHA)	256
`google/gemma-4-E2B`	Full Sliding	131,072	1.0×10^6 1.0×10^4	35	8/1 (GQA)	512 256

A.2 Experimental Setting for Context Extension

We fine-tune on the training split of PG-19 (Rae et al., 2020). Each document is tokenized with an appended end-of-sequence token, and the resulting stream is packed into sequences of 32,768 tokens. We use QLoRA (Dettmers et al., 2023): the base model is loaded in 4-bit NF4 precision with

[9] We use `transformers` v5.12.0 throughout. Unless otherwise noted, models are run in bfloat16.

[10] https://huggingface.co/meta-llama/Meta-Llama-3-8B

[11] https://huggingface.co/Qwen/Qwen3-8B

[12] https://huggingface.co/Qwen/Qwen3-4B

[13] https://huggingface.co/google/gemma-7b

[14] https://huggingface.co/google/gemma-4-E2B

double quantization and an fp16 compute dtype. LoRA adapters are inserted only into the query and key projections (`q_proj`, `k_proj`) with rank $r = 8$, $\alpha = 16$, and dropout 0.1. No bias terms are trained. Restricting the adapters to the query and key projections keeps the update confined to the two projections that the position encoding acts on. We optimize with paged 8-bit AdamW at a learning rate of 2×10^{-4} and a weight decay of 0.01, using a linear decay schedule without warmup. Training is run in fp16 with gradient checkpointing enabled. Training is distributed over two GPUs with two sequences per device and two gradient accumulation steps, giving an effective batch of eight sequences, or $262{,}144$ tokens per optimizer step. We use the checkpoint at step $1{,}000$ for every condition, which corresponds to roughly 2.6×10^8 training tokens.

A.3 Experimental Setting for End-to-End Retrieval

We evaluate Llama-3-8B with $N = 100$ samples for each configuration, sweeping haystack lengths $L \in \{4{,}096, 8{,}192, 16{,}384, 32{,}768\}$ and distances $d \in \{2, 8, 32, 128, 512, 2048\}$. Each haystack is built from the WikiText chunks. We use `"§"` as the marker token, which is a single token for this tokenizer, and replace all of its pre-existing occurrences in the haystack with `"the"`, so that the marker token appears exactly once, at position $(L-1)-d$; unlike the head-level probe, no marker is placed at the query position. Note that L counts haystack tokens only: the question is appended after them, so the actual input length is L plus the question. We use a 3-shot prompt. After the haystack we append the line `"\nThe unusual marker token in the text above, copied verbatim, is:"` four times, filling in the first three with ¶, ±, and ©, for 67 question tokens in total. Like the marker token, these are single tokens. They are not correct answers for the haystack at hand; they only fix the output format, namely a single unusual symbol token. A sample is counted as correct when the $\arg\max$ of the next-token distribution at the final position equals the marker token id; we neither decode nor string-match, which is why the demonstration answers must be single tokens.

A.4 Position Interpolation

We consider two training-free context-extension methods, applied at inference time by overriding the RoPE configuration: Linear Positional Interpolation (PI) (Chen et al., 2023) and YaRN (Peng et al., 2024). In both cases the target context length is $L_{\text{target}} = 32{,}768$ and we sweep the scaling factor $s \in \{2, 4, 8\}$.

Linear PI. All frequencies are rescaled uniformly,

$$\theta_i^{\text{PI}} = \theta_i / s, \tag{9}$$

which is equivalent to compressing position indices as $m \mapsto m/s$.

YaRN. YaRN applies NTK-by-parts interpolation (bloc97, 2023), which interpolates only the low-frequency (long-wavelength) components while leaving high-frequency components extrapolated. With wavelength $\lambda_i = 2\pi/\theta_i$ and rotation count $r_i = L_{\text{train}}/\lambda_i$, the two regimes are blended by a linear ramp

$$\gamma(r_i) = \begin{cases} 0, & r_i < \beta_{\text{slow}}, \\ \dfrac{r_i - \beta_{\text{slow}}}{\beta_{\text{fast}} - \beta_{\text{slow}}}, & \beta_{\text{slow}} \le r_i \le \beta_{\text{fast}}, \\ 1, & r_i > \beta_{\text{fast}}, \end{cases} \tag{10}$$

$$\theta_i^{\text{YaRN}} = \frac{\theta_i}{s}\big(1 - \gamma(r_i)\big) + \theta_i\, \gamma(r_i). \tag{11}$$

We use the recommended values $\beta_{\text{fast}} = 32$ and $\beta_{\text{slow}} = 1$ (Peng et al., 2024).

Attention temperature. In addition to modifying the frequencies, YaRN rescales the attention logits by a temperature $m(s)$. In `transformers`, this term is an optional `attention_factor` field of the RoPE configuration; when it is left unset, the library falls back to the value recommended by the original work,

$$m(s) = 0.1\, \ln(s) + 1, \tag{12}$$

which we adopt in all of our experiments.

B RULER BENCHMARK

RULER (Hsieh et al., 2024) is a configurable benchmark for evaluating long-context LMs across different sequence lengths and task complexities. It contains 13 tasks in four categories: eight NIAH-based retrieval tasks, Variable Tracking for multi-hop tracing, Common/Frequent Words Extraction for aggregation, and two QA tasks based on SQuAD and HotpotQA. Unlike vanilla NIAH, RULER evaluates long-context capabilities beyond simple retrieval and measures performance degradation as the input length increases.

For each task, RULER generates examples at controlled sequence lengths by increasing the amount of distractor context while preserving the task objective. In the original evaluation, 500 examples are generated for each task and each sequence length from 4K to 128K. Performance is measured using recall-based accuracy, and the RULER score at each sequence length is obtained by averaging the accuracy across the 13 tasks. This design enables direct comparison of how long-context capability degrades as the input length increases.

In our experiments (§5.2), we evaluate RULER at 8K, 16K, and 32K to compare how different context-extension settings preserve performance as the input length increases.

C ℓ_2 NORM DISTRIBUTIONS ACROSS ALL ATTENTION HEADS

Figure 9 shows the complete ℓ_2-norm distributions for all attention heads in layer 31 of Llama-3-8B. Consistent with the analysis in § 4.4, large norms are primarily concentrated in the `mid-low` to `low` frequency ranges, while the dominant frequencies vary only slightly across heads.This full-layer view complements the representative example in Figure 7 and further illustrates that similar frequency-magnitude patterns can coexist with different distance-dependent retrieval behaviors observed by HRDP.

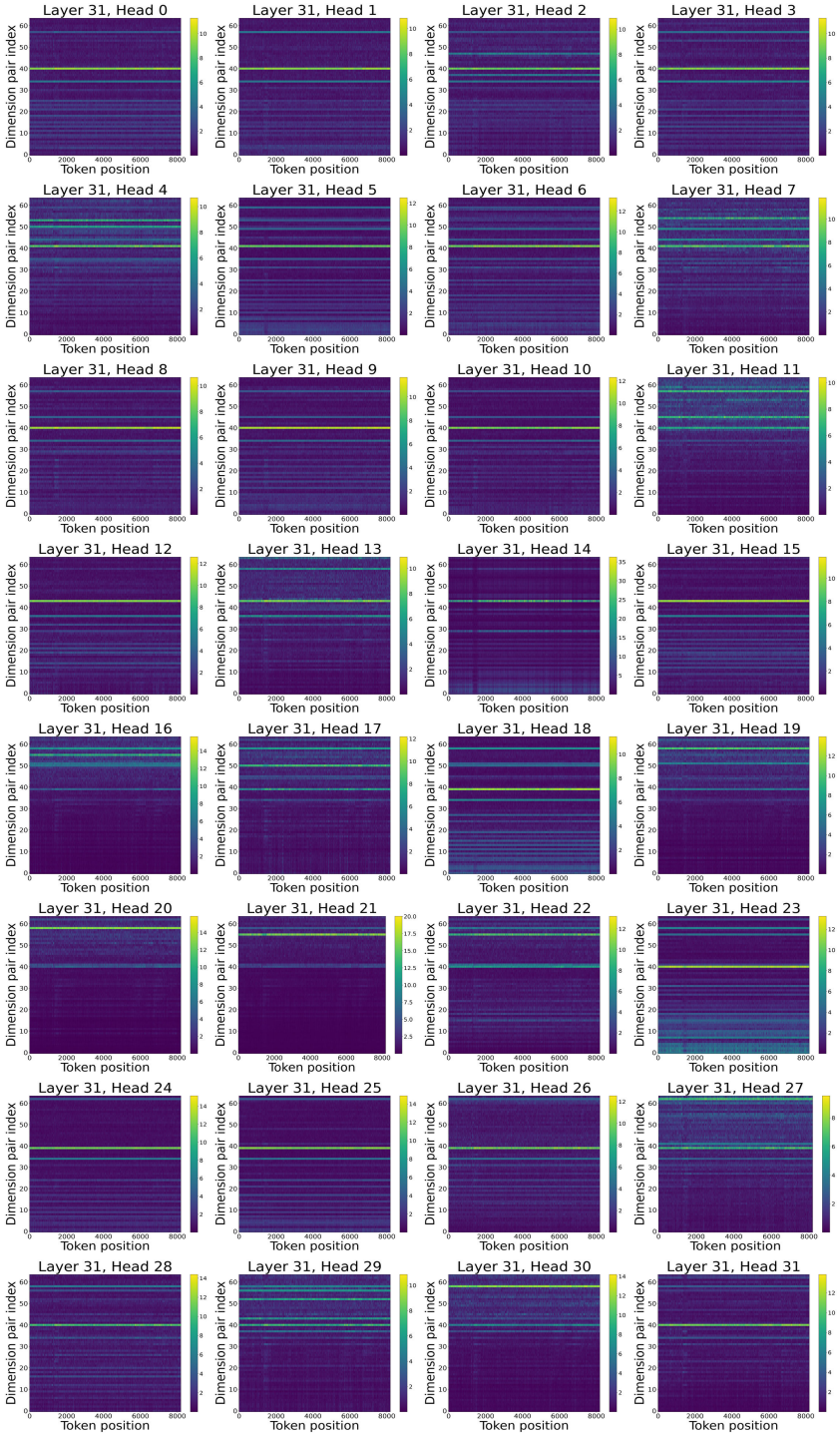


Figure 9: ℓ_2-norm distribution for head of Llama-3-8B in Layer 31. The horizontal axis represents token position, and the vertical axis represents RoPE dimension pair index, where higher indices correspond to lower rotational frequencies.

D HRDP RESULTS ON ADDITIONAL MODELS

In this section, we present the results of applying HRDP to Llama-3-8B, Qwen3-8B, Qwen3-4B, and Gemma-7B. Figure 10 shows the results for Llama-3-8B, Figure 11 shows the results for Qwen3-4B, Figure 12 shows the results for Qwen3-8B, Figure 13 shows the results for Gemma-7B, and Figure 14 shows the results for Gemma4-E2B.

From the experiment results, we observe two general trends across models. In terms of sequence length, no head can retrieve the marker token once the sequence is longer than the training length. Llama-3-8B was trained on 8,192 tokens, and accordingly almost no head retrieves the marker token at $L = 16{,}384$ and $32{,}768$. Gemma-7B behaves the same way beyond its own training length of 8,192 tokens. Qwen3-8B and Qwen3-4B show no such drop, simply because their training length of 40,960 tokens exceeds every length we evaluate. Gemma-4-E2B fits neither pattern. Its training length of 131,072 tokens likewise covers all of our lengths, but only 7 of its 35 layers use full attention: the remaining 28 use a sliding window of 512 tokens to handle long contexts efficiently. For $d \geq 512$ the marker token falls outside every window, so those layers cannot retrieve it at all, and the grid is carried entirely by the seven full-attention layers.

In terms of layer depth, retrieval is carried by a small number of heads rather than being spread across the network. In Llama-3-8B at $L = 8{,}192$, only 12 of its 1,024 (layer, head) pairs reach an score of 0.8 at any distance, and three of them, $(4, 16)$, $(8, 11)$ and $(14, 19)$, account for most of the retrieval across distances. No head above layer 20 reaches 0.8 at any distance, and layers 2, 3 and 23 of this model show essentially no retrieval at all. The responsible heads also specialize by distance: the head at $(0, 9)$ is accurate for $d \leq 32$ but collapses beyond it, whereas $(10, 2)$ appears only at $d \geq 512$. In Gemma-4-E2B, at $d \leq 128$ where both kinds of layers can reach the marker token, the heads that retrieve well are likewise concentrated in the shallow layers, and their fraction decreases as the layers get deeper.

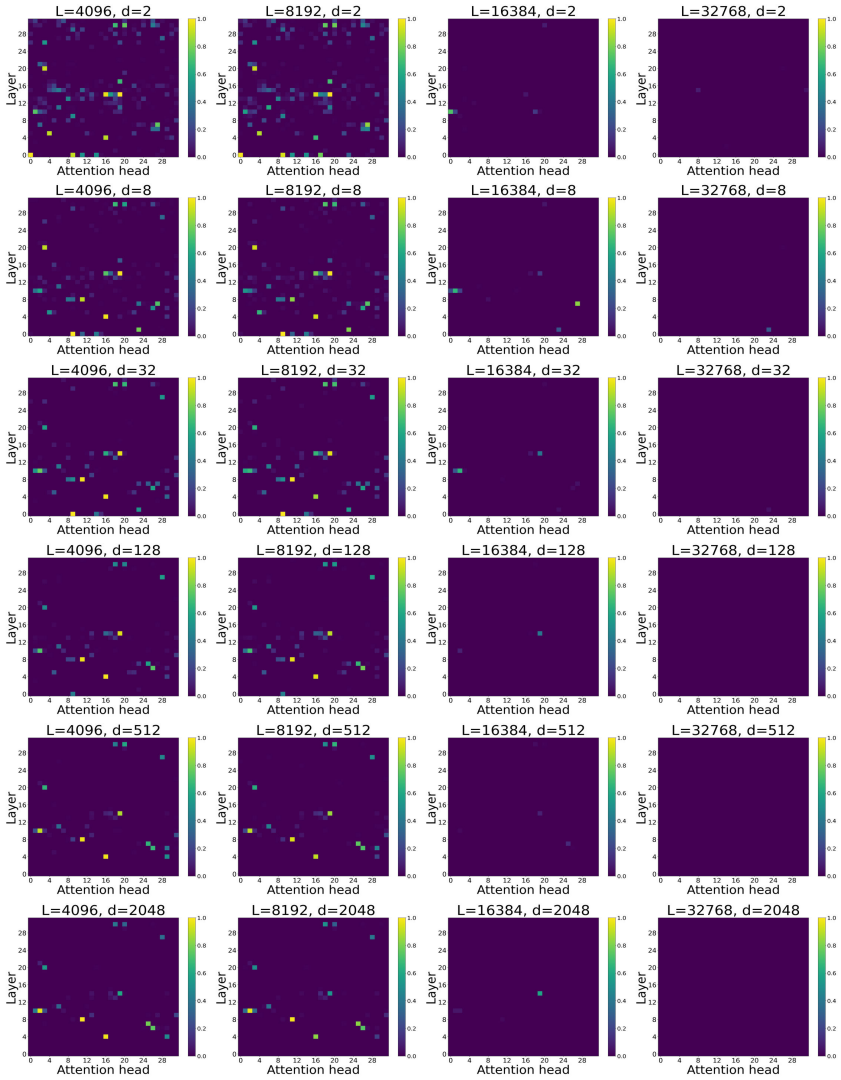


Figure 10: HRDP results of Llama-3-8B under default RoPE at relative offset $d \in \{2, 8, 32, 128, 512, 2048\}$ for sequence lengths $L \in \{4\text{k}, 8\text{k}, 16\text{k}, 32\text{k}\}$.

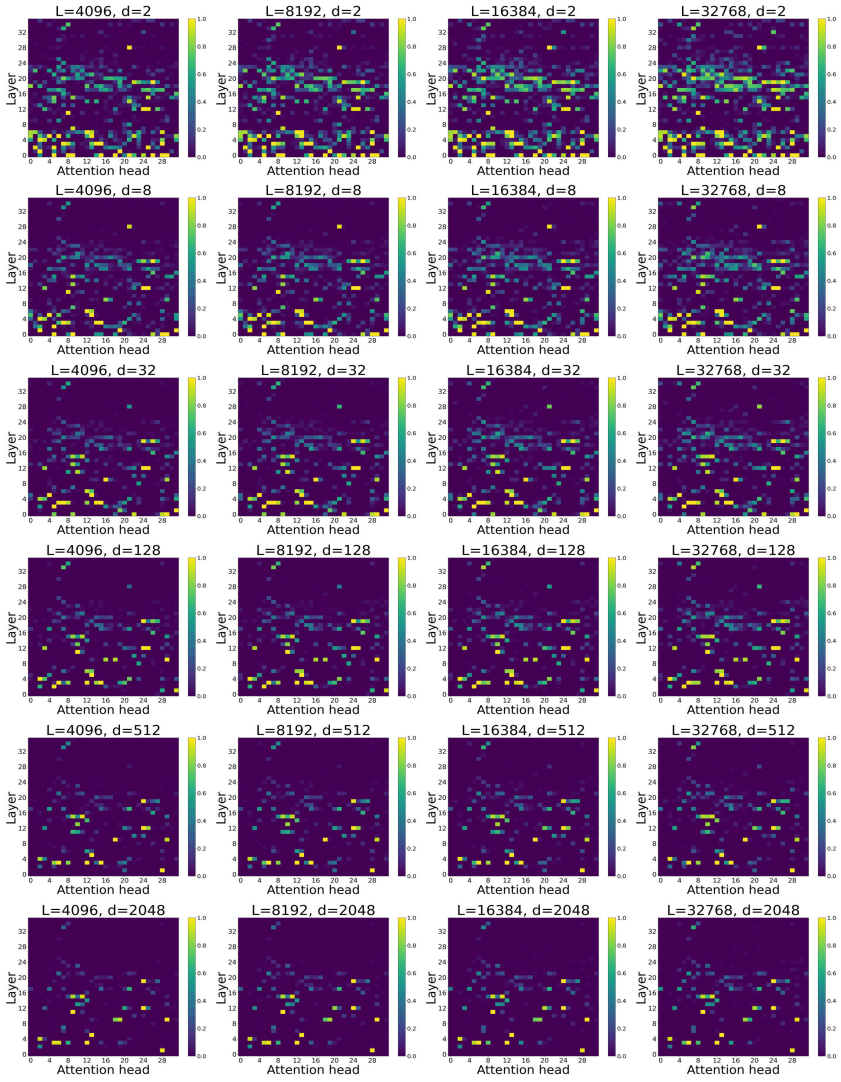


Figure 11: HRDP results of Qwen3-4B under default RoPE at relative offset $d \in \{2, 8, 32, 128, 512, 2048\}$ for sequence lengths $L \in \{4\text{k}, 8\text{k}, 16\text{k}, 32\text{k}\}$.

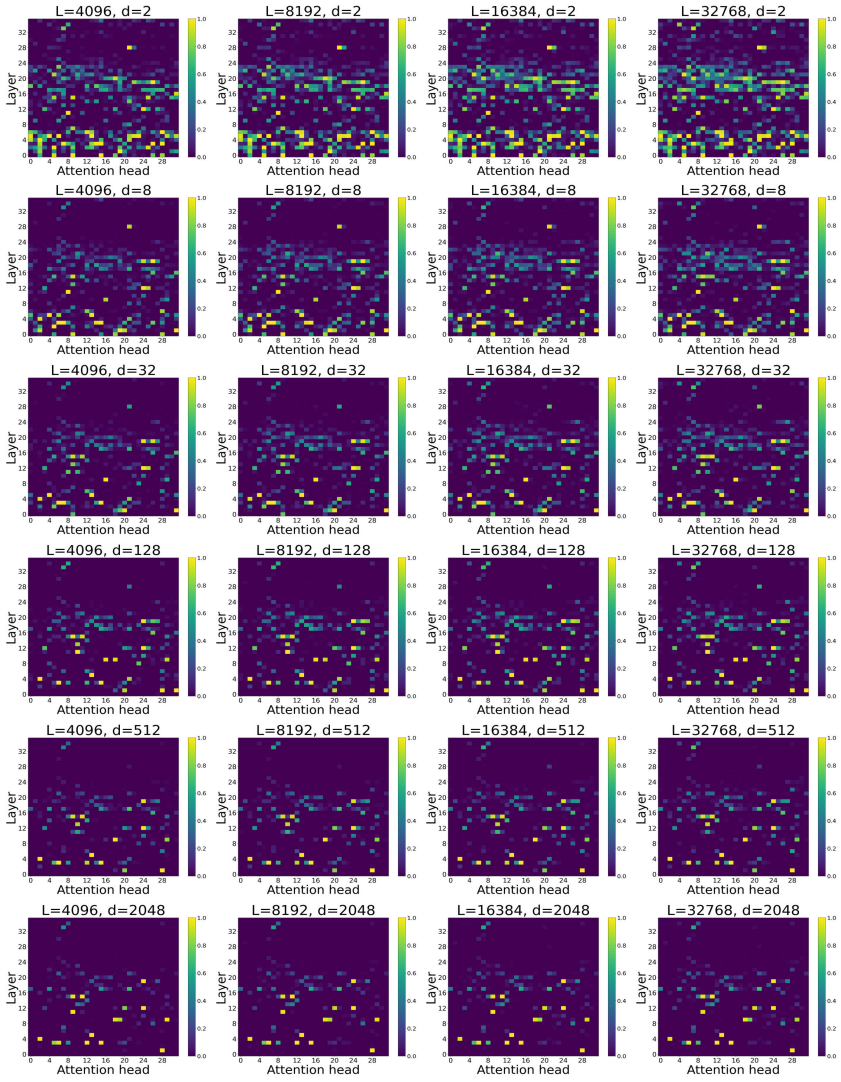


Figure 12: HRDP results of Qwen3-8B under default RoPE at relative offset $d \in \{2, 8, 32, 128, 512, 2048\}$ for sequence lengths $L \in \{4\text{k}, 8\text{k}, 16\text{k}, 32\text{k}\}$.

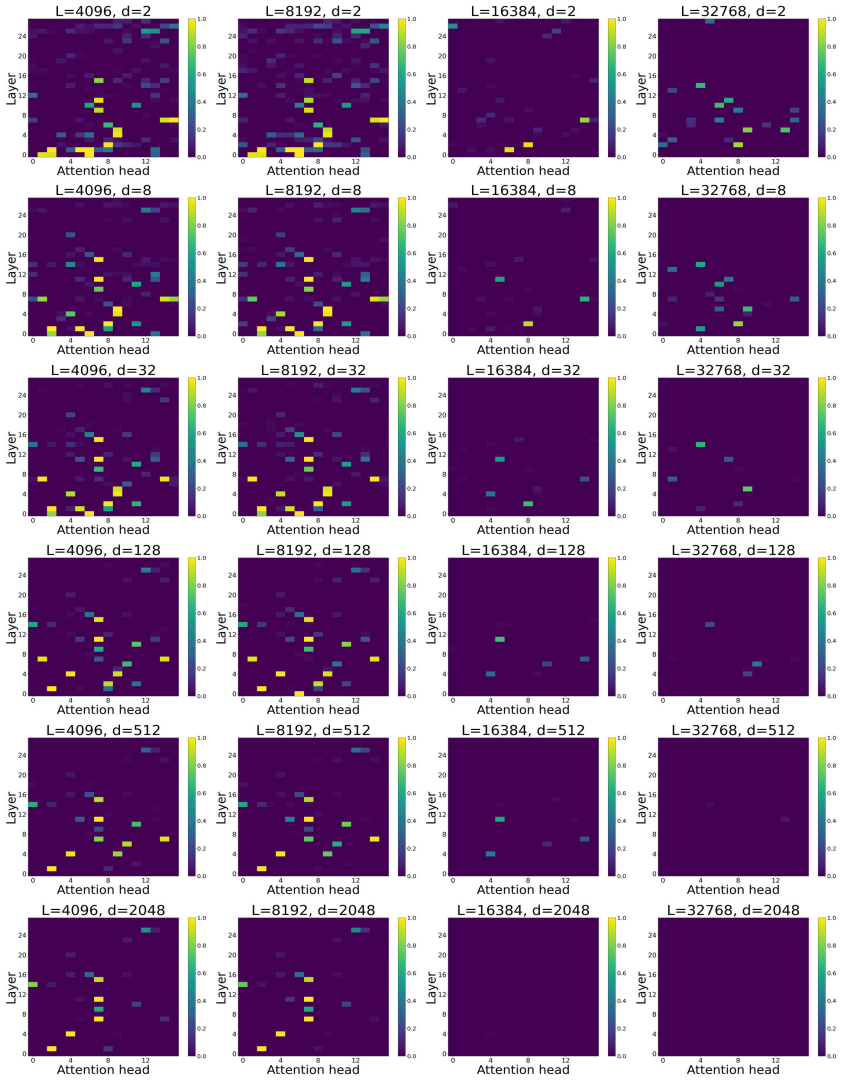


Figure 13: HRDP results of Gemma-7B under default RoPE at relative offset $d \in \{2, 8, 32, 128, 512, 2048\}$ for sequence lengths $L \in \{4\text{k}, 8\text{k}, 16\text{k}, 32\text{k}\}$.

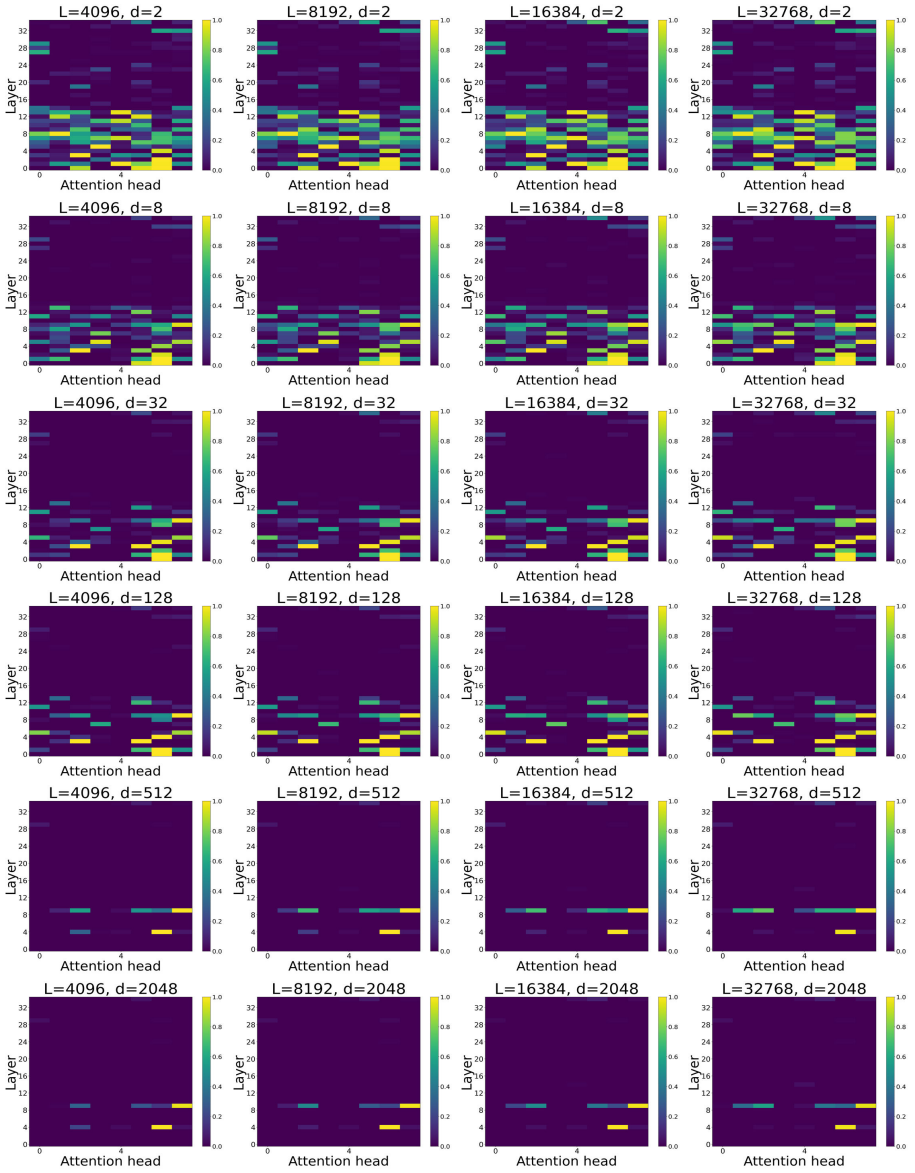


Figure 14: HRDP results of Gemma4-E2B under default RoPE at relative offset $d \in \{2, 8, 32, 128, 512, 2048\}$ for sequence lengths $L \in \{4\text{k}, 8\text{k}, 16\text{k}, 32\text{k}\}$.