Can Calibration of Positional Encodings Enhance Long Context Utilization?

Tom Zehle[1,2,3,4] **and Matthias Aßenmacher**[3,4]

[1]ELLIS Institute, Tübingen, Germany, [2]University of Freiburg, Germany,
[3]LMU Munich, Germany, [4]Munich Center for Machine Learning (MCML), Germany
Correspondence: tom.zehle@tue.ellis.eu

Abstract

Large language models suffer from positional biases like the "Lost in the Middle" (LiM) phenomenon and recency bias, which reduce the effective utilization of long contexts. In this work, we investigate the role of positional encodings in this context. Our empirical study confirms the persistence of these biases in modern large language models. Drawing on these findings, we introduce CALIOPE, a training-free framework for calibrating positional encodings at inference time. Our calibrators yield substantial improvements on needle-in-a-haystack and cross-chunk reasoning benchmarks, and offer a practical, lightweight method for improving long-context utilization.

1 Introduction

Large language models (LLMs) have demonstrated substantial capabilities across diverse natural language processing tasks, including text generation, logical reasoning, and question answering (Brown et al., 2020; OpenAI, 2023). These models leverage vast pre-training corpora to acquire broad knowledge, enabling generalist performance across multiple domains (Radford et al., 2019). However, their application in specialized settings faces a critical issue: The inability to access internal data outside of public training corpora, or real-time information that emerges after model training, is a fundamental limitation originating from the static nature of pre-training, as models can not access information that emerges from proprietary data or was created after their training cutoff date. Retrieval augmented generation (RAG) has emerged as a prominent approach to address these limitations by retrieving and incorporating relevant documents during inference time (Lewis et al., 2020). Modern LLMs advertise substantial context window capacities, suggesting that models should effectively process extensive retrieved information. However, empirical evidence reveals a systematic failure known as the "*Lost in the Middle*" (LiM) phenomenon, where model performance exhibits a U-shaped curve with respect to information position: models effectively utilize information at sequence boundaries while systematically underutilizing content in middle positions (Liu et al., 2024). This positional bias means that critical information becomes less visible solely due to its positioning in the context, undermining the practical utility of large context capacities. More generally, ineffective utilization of long contexts affects a broad range of applications beyond RAG, including extended conversational interactions with LLM assistants and tasks involving lengthy reasoning.

While current LLM research is mostly focused on various architectural and training improvements, positional encoding mechanisms remain comparatively understudied. Rotary position embeddings (RoPE; Xiong et al., 2023) have become the dominant encoding scheme in LLMs (Sun et al., 2021), yet mounting evidence suggests that these encodings correlate with long-context failures (Wang et al., 2024; Ding et al., 2024). We investigate whether calibration of (rotary) positional encodings can enhance the utilization of long context.

Our Contributions are the following:

1. We provide a comprehensive literature review of the LiM phenomenon and a categorization of possible existing mitigation strategies.
2. We conduct an empirical study to verify the existence of the LiM phenomenon, alongside recency biases in recently released LLMs.
3. We propose the **Cali**bration **o**f **P**ositional **E**ncodings (CALIOPE) framework, introducing a set of novel training-free position calibrators, as well as a rigorous evaluation.
4. We provide an open-source codebase for replication and extension of our experiments, available at `https://github.com/finitearth/caliope`.

Findings of the Association for Computational Linguistics: EACL 2026, pages 2268–2280
March 24-29, 2026

2 Related Work

Positional Encodings. The attention mechanism's permutation invariance necessitates explicit position information to distinguish between "the monkey ate the banana" and "the banana ate the monkey". Positional encoding schemes inject this ordering information by mapping each sequence position t to a vector representation $\psi(t)$. The original Transformer (Vaswani et al., 2017) employed sinusoidal functions to encode absolute positions without learnable parameters. BERT (Devlin et al., 2018) and GPT-2 (Radford et al., 2019) adopted trainable absolute position embeddings, parameterizing positions as a learnable matrix $\psi(t) \in \mathbb{R}^{T\times d}$, that – as in sinusoidal position encodings – is added to the token embeddings. Because this limits the model's applicability to sequences exceeding the training length T, Shaw et al. (2018) introduced relative position encodings (RPE), recognizing that syntactic and semantic relationships depend more on relative distances than absolute positions: $\alpha_{ts} = q_s^\top k_t + b_{t-s}$, where b_{t-s} represents a learned embedding for the relative distance $t - s$ between positions, clipped at a maximum distance.

Rotary Position Embeddings Xiong et al. (2023) build upon the concept of RPE while replacing learnable parameters in the positional encoding mechanism with a deterministic rotational transformation, by applying rotation matrices, constructed using sine and cosine functions. Similar to RPE, RoPE is applied exclusively to queries and keys, leaving values unmodified, ensuring that positional information solely affects attention weight computation, while preserving the semantic content aggregated through value vectors.

Lost in the Middle. The LiM phenomenon, as characterized by Liu et al. (2024), manifests as a U-shaped performance curve where LLMs demonstrate better utilization of information from beginning and end positions of provided context inside of the prompt while systematically under-performing on middle-positioned content. This pattern emerges consistently across major language model families, including GPT-3.5 and Claude 1.3 at the time of the original study, suggesting a fundamental representational limitation rather than a model-specific artifact. The core experimental paradigm employs a multi-document question-answering task with carefully controlled positioning of relevant information, inserting one relevant chunk among $k-1$ distractor chunks. These distractors are selected as the most relevant Wikipedia passages that do not contain the answer, retrieved using standard retrieval models. This design creates a challenging scenario where distractors maintain topical relevance but lack the specific information needed to answer the query, requiring the model to identify which retrieved chunk contains the pertinent details. In these controlled experiments, all conditions, except for the position of the relevant chunk, remain unchanged.

Mitigation Strategies. Existing attempts to mitigate the LiM phenomenon can be grouped into three levels of intervention: retrieved chunk processing, attention mechanism modifications, and positional encoding adjustments (cf. Tab. 1). Each category contains both training-free approaches that can be immediately applied to pre-trained models and training-based approaches that require fine-tuning or retraining. Although many of these methods are motivated by retrieval-augmented generation and Lost in the Middle specifically, the underlying phenomenon reflects a more general limitation in long-context utilization.

	Training-Free	**Requires Training**
Chunk Processing	Attention Sorting	LongLLMLingua*
Attention Mechanism	Found in the Middle SSMax Scale positional hidden states	Position-Agnostic Fine-tuning Curriculum Learning
Positional Encoding	CALIOPE *(ours)*	Position Interpolation YaRN LongRoPE NoPE p-RoPE

Table 1: Categorization of approaches addressing the LIM phenomenon.
*Requires training a lightweight compression model, not the main LLM.

Chunk processing methods address the problem at the input level through reordering or compression of retrieved information. *Attention Sorting* (Peysakhovich and Lerer, 2023) employs iterative retrieved chunk reordering, building on the hypothesis that models do attend more strongly to relevant retrieved chunks even when poorly positioned, as they simply fail to utilize this attended information effectively. *LongLLMLingua* (Jiang et al., 2024) implements question-aware prompt compression using smaller language models, finetuned for the selection of relevant context.

Attention mechanism modifications are direct interventions at inference time, which often require

architectural changes. *Found in the Middle Calibration* (Hsieh et al., 2024) addresses positional bias in attention by first estimating it through systematic shuffling of retrieved chunks. The measured bias is then subtracted from the attention scores during inference, assuming that positional effects are linearly separable. *Scale positional hidden states* (Yu et al., 2025) propose a method that mitigates positional bias by identifying and modifying hidden-state channels associated with position-dependent behavior induced by the causal attention mask. *Scalable-Softmax (SSMax)* (Nakanishi, 2025) introduces a modified softmax as a drop-in replacement for standard softmax that prevents attention flattening in long contexts. *Position-agnostic fine-tuning*, explored in An et al. (2024) and He et al. (2024), fine-tunes models on synthetic data with systematic position shuffling, forcing models to learn position-independent attention patterns. *Curriculum learning* with progressive context extension, investigated by Nagatsuka et al. (2023), and Li et al. (2022), gradually increases context length during training, allowing models to learn positional patterns incrementally.

Positional encoding modifications are targeted alterations of what may be the root of the representational bias. Existing methods (all requiring training) include *Position Interpolation* (Chen et al., 2023), which extends a model's context length by linearly compressing position indices to fit longer sequences. YaRN (Peng et al., 2023) advances beyond this by applying different scaling strategies to different frequency bands within the positional encoding. *LongRoPE* (Ding et al., 2024) generalizes this through evolutionary search for optimal per-dimension scaling factors. Barbero et al. (2024) show that high RoPE-frequencies encode positional attention for nearby tokens while low frequencies capture semantic relationships independent of distance. Their proposed p-RoPE removes the lowest frequencies with wavelengths exceeding context length. Wang et al. (2024) show that, removing positional encodings (NoPE) can improve length generalization.

CALIOPE occupies a distinct and direct point of intervention in this landscape, as it acts solely on the explicit inputs to the rotary positional encoding mechanism via deterministic position remapping at inference time. Unlike methods that probe or manipulate internal representations, CALIOPE requires no access to hidden states, no identification of bias-related channels, and no parameter tuning or retraining. It does not directly modify hidden-state tensors or learned parameters, making it a fully training-free, architecture-agnostic intervention that can be applied to existing RoPE-based models as a drop-in modification.

3 Experiments

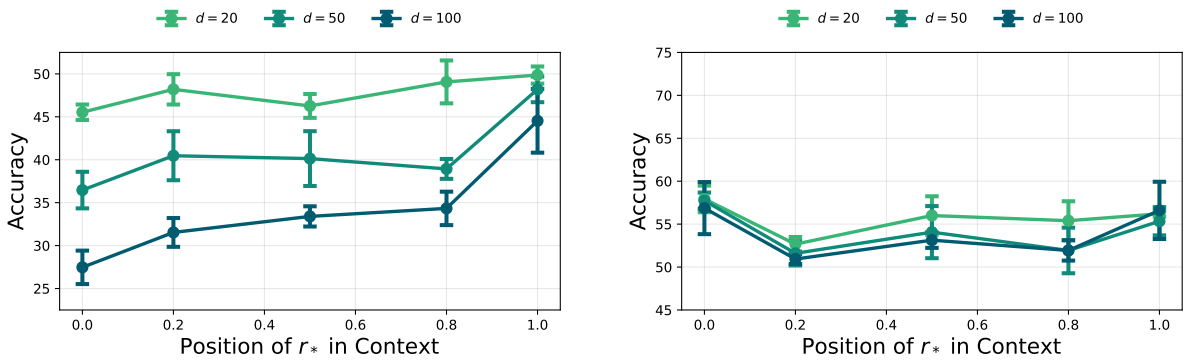
We evaluate positional biases and the effects of calibration using two complementary long-context datasets that capture distinct aspects of context utilization. First, we use the Natural-QA dataset (Kwiatkowski et al., 2019) to study single-fact retrieval under varying positional conditions, also known as Needle-in-a-Haystack experiments. Second, we use the Babilong-QA dataset (Kuratov et al., 2024) to assess cross-context reasoning, where answers require integrating multiple relevant facts distributed across a longer context.

Following (Liu et al., 2024), we evaluate the models on the Natural-QA dataset (Kwiatkowski et al., 2019), employing the classic needle-in-haystack paradigm, where a single relevant chunk r_* is placed among $d-1$ irrelevant distractors.

The retrieved information set $\mathcal{R} = \{r_1, ..., r_d\}$ contains exactly one chunk with the answer to the posed question. Key modifications from the original study include extended context scales with $d \in \{20, 50, 100\}$ chunks, corresponding to approximately 3K, 8K, and 16K tokens of context, respectively. We systematically vary the relevant chunk's position at $d_* \in \{0, 0.2d, 0.5d, 0.8d, d\}$ to test primacy effects ($d_* = 0$), recency effects ($d_* = d$), and varying degrees of middle positions. Prompt structure, question difficulty, and distractor content are kept constant across all position variations.

For each configuration, we run three random seeds with 500 questions sampled from the original datasets and report mean performance with standard deviation. Following Liu et al. (2024) and Kuratov et al. (2024), we use substring matching to check whether the correct answer appears in the model output. Since this metric does not assess semantic correctness (e.g., negation or misuse in context), we additionally employ an LLM-as-a-judge evaluation using Llama-3.3-70B to assess contextual and semantic validity.

To quantify the strength of the U-shaped positional bias, we calculate the Pearson correlation $\rho_{\delta,a}$ between a position's distance from the middle (δ_t) and its corresponding accuracy (a) multiplied



(a) Llama-3.1-8B

(b) Qwen-2.5-14B

Figure 1: Comparison of Accuracy across varying context window positions. The x-axis represents the normalized position of the relevant information within the context window. Note that y-axis ranges are adjusted per subplot to emphasize patterns while maintaining consistent scale spans across all plots.

Model	Position of r_* in retrieved information					$\rho_{\delta,a}$
	0%	**20%**	**50%**	**80%**	**100%**	
Llama-3.1-8B						
$d = 20$	45.53 ± 0.90	48.20 ± 1.77	46.27 ± 1.39	49.07 ± 2.50	49.87 ± 1.00	0.83
$d = 50$	36.47 ± 2.13	40.47 ± 2.85	40.13 ± 3.19	38.93 ± 1.16	48.20 ± 1.50	1.60
$d = 100$	27.47 ± 1.95	31.53 ± 1.68	33.40 ± 1.18	34.33 ± 1.95	44.53 ± 3.71	2.08
Qwen-2.5-14B						
$d = 20$	57.93 ± 1.55	52.67 ± 0.84	56.00 ± 2.24	55.40 ± 2.26	56.20 ± 0.33	1.06
$d = 50$	57.73 ± 0.94	51.60 ± 1.41	54.07 ± 3.03	51.93 ± 2.65	55.33 ± 1.64	2.14
$d = 100$	56.87 ± 3.03	50.93 ± 0.57	53.13 ± 0.90	51.93 ± 1.18	56.60 ± 3.34	2.97

Table 2: Retrieval accuracy (%) on Natural-QA measured by LLM-as-a-Judge.

by 100. We calculate the normalized distance from the middle as:

$$\delta_t = \frac{|t - 0.5T|}{0.5T}, \tag{1}$$

yielding values of 0 for the middle position ($d_* = 0.5d$) and 1 for border positions. Positive correlation indicates U-shaped patterns, near-zero correlation suggests no positional bias, while negative correlation reveals inverted-U patterns. We use this correlation measure to complement our visual inspection of the performance plots.

While the aforementioned experiment provides insight into the presence and strength of positional biases, it falls short of representing challenges present in real-world long-context use-cases: Oftentimes, answers are not contained within a single relevant paragraph, but scattered across the provided context. We conduct a complementary experiment on the BabiLong-QA cross-chunk reasoning dataset (Kuratov et al., 2024). Babilong-QA requires the model to identify and integrate two relevant facts distributed across different locations within a context of distractor documents. By varying the total context length from 0 to 8,000 tokens, we assess whether the model maintains its ability to reason across the provided context.

In these experiments, we employ Llama-3.1-8B-Instruct (Touvron et al., 2023), and Qwen-2.5-14B-Instruct (Bai et al., 2023), both of which support a context length of 128k tokens. A detailed description of the models can be found in Appendix B, and details about the hardware utilized in Appendix C. Additional details about the used datasets can be found in Appendix D.

4 Confirming Positional Biases

The original LiM study (Liu et al., 2024) demonstrated severe positional bias in language models available at the time. We replicate these experiments on modern LLMs to assess whether such biases persist and to establish a current baseline. Experiments are conducted at context lengths of $d \in 20, 50, 100$ chunks. Results from the LLM-as-a-Judge evaluation are reported in Figure 1 and Table 2, with additional substring-matching results provided in Appendix A.1 (Tab. 5, Fig. 4).

Findings. Llama-3.1-8B demonstrates a severe recency bias that manifests across all context lengths, meaning that positioning the relevant chunk in the first position results in a substantially worse performance (27.5 % using LLM-as-a-Judge at $d = 100$), than positioning it at the end of the context (44.5%). This model also degrades heavily in performance, when increasing the context length (46.3% for $d = 20$ at position $d_* = 0.5d$, 40.1% for $d = 50$, and 33.4% for $d = 100$).

Qwen-2.5-14B showcases a severe negative bias for middle positions, as the start (56.9% for $d = 100$, using LLM-as-a-Judge) and end (56.6%) positions perform vastly better than the middle position (53.1%). Not only does the general performance decrease with increased scale of context, but also the severity of the LiM U-Shape increases, as quantified by $\rho_{\delta,a}$ (1.1 for $d = 20$, 2.1for $d = 50$, and 3.0 for $d = 100$). Notably for this model, the accuracy measured by LLM-as-a-Judge is notably lower (ranging from 50% to 58%), than measured by substring matching (52% to 68%).

Having confirmed that the positional biases persist within recent LLMs, we now turn to developing a targeted solution. Our analysis suggests that middle positions suffer from the characteristic information underutilization that defines the LiM phenomenon, while later positions benefit from a recency bias. To address these issues, we present **CALIOPE**, a family of training-free calibration methods designed as drop-in replacements compatible with RoPE-based architectures.

5 CALIOPE: Calibration of RoPE

CALIOPE employs strictly monotonic transformation functions on the input to the positional encoding function while leaving all learned model parameters untouched.[1] The position calibration framework operates on a set of retrieved chunks $\mathcal{R} = \{r_1, r_2, \ldots, r_d\}$, where d denotes the total number of retrieved chunks in the context. To track the relationship between individual tokens and their respective chunks, we define a retrieved chunk membership function $m(t)$ that returns the index of the retrieved chunk preceding the token at index t, with $m(t) = 0$ for tokens appearing before any retrieved information.

Central to the calibration approach is the gap accumulation function $c : \{0, \ldots, d\} \to \mathbb{R}_0^+$, which maps each retrieved chunk index to a cumulative position offset. This function enables the position remapping operation $\Phi : \mathbb{N} \to \mathbb{R}_0^+$, defined as:

$$\Phi(t) = t + c(m(t)) \tag{2}$$

This transformation maps each token position t to a modified position that serves as input to the positional encoding function $\psi(\Phi(t))$, compared to uncalibrated encodings, where the encodings are solely based on t. A critical constraint for all calibration functions is strict monotonicity: $\forall t_1, t_2 : t_1 < t_2 \implies \Phi(t_1) < \Phi(t_2)$, ensuring that the order of tokens is preserved.

The calibrators are presented with untuned parameters, demonstrating their fundamental efficacy in an out-of-the-box configuration and providing clear baselines that avoid task-specific overfitting. The central question is whether strategic position remapping can reduce or eliminate the observed positional biases. To address this, we repeat the experiment from Section 4 with $d = 100$ retrieved chunks, applying each of the following calibrators during inference.

Moses Calibrator. The hypothesis underlying this calibrator posits that if positional encodings themselves cause the LiM phenomenon, avoiding these problematic encodings should improve downstream performance. This transformation fundamentally alters the relative positioning: tokens originally at the middle position become located at edges in an expanded position space. For instance, when mapping positions to ranges 0–500 and 1500–2000 instead of the original 0–1000, the calibrator ensures that retrieved chunks bypass the relative middle of typical position distributions entirely.

Moses Calibrator

Visualization:

Gap function:

$$c(m) = \begin{cases} 0 & \text{if } m \leq \lfloor d/2 \rfloor \\ \Delta_{\text{gap}} & \text{if } m > \lfloor d/2 \rfloor \end{cases} \tag{3}$$

Parametrization: $\Delta_{\text{gap}} = 10{,}000$.

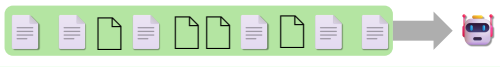
Hourglass Calibrator. Unlike the Moses Calibrator's discontinuous jump, the Hourglass approach provides gradual transitions that minimize middle position occupancy while maintaining local

[1] Implementation details are provided in Appendix E.

coherence at sequence boundaries. It does so by applying a gap function based on a parabola, which contains maxima at the middle positions, therefore making gaps in the middle bigger, ergo less tokens are mapped to "the middle".

Hourglass Calibrator

Visualization:



Gap function:

$$c(m) = \sum_{k=1}^{m-1} \left[\Delta_{\min} + 4 \cdot \frac{k}{d-1} \cdot \left(1 - \frac{k}{d-1}\right) \cdot (\Delta_{\max} - \Delta_{\min}) \right] \quad (4)$$

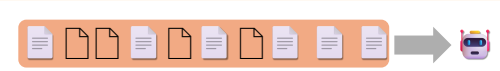
Parametrization:

$\Delta_{\min} = 5; \quad \Delta_{\max} = 1{,}000$

Decay Calibrator. Early chunks typically occupy positions far from the query, regions where models systematically ignore content. By introducing large initial gaps that decay exponentially, the calibrator effectively "pulls" early retrieved chunks closer in relative terms to the query position. Later retrieved chunks largely maintain their natural clustering in positions where attention mechanisms already function effectively. This redistribution places early retrieved chunks at positions with improved relative proximity to the question, compensating for the model's inherent bias toward recent content, observed in the Llama-3.1-8B in Section 4.

Decay Calibrator

Visualization:



Gap function:

$$c(m) = \sum_{k=1}^{m-1} \Delta_0 \cdot \lambda^k \quad (5)$$

Parametrization:

$\Delta_0 = 1{,}000; \quad \lambda = 0.95$

6 Experimental Results

6.1 Positional Biases after Calibration

The results measured by LLM-as-a-Judge of the LiM experiment with the aforementioned calibrators, showcased in Figure 2 and Table 3 paint a promising picture. For the results using Substring-Matching we refer to Appendix A.2.

For Llama-3.1-8B every calibrator increases the performance by a huge margin, and in the case of the Decay and Moses calibrator combat the positional bias effectively. In the case of Qwen-2.5-14B the calibrators stay competitive for the most part, with the exception of the Hourglass calibrator, that induces a severe recency bias into the model.

The Moses calibrator demonstrates the most consistent success in tackling the positional biases for Llama-3.1-8B, where Moses maintains accuracy between 50% and 53%. For Qwen-2.5-14B, Moses does not substantially alter performance compared to using unaltered RoPE; however, it decreases $\rho_{\delta,a}$ from 3.0 to 2.3, indicating a slight success in mitigating the LiM phenomena.

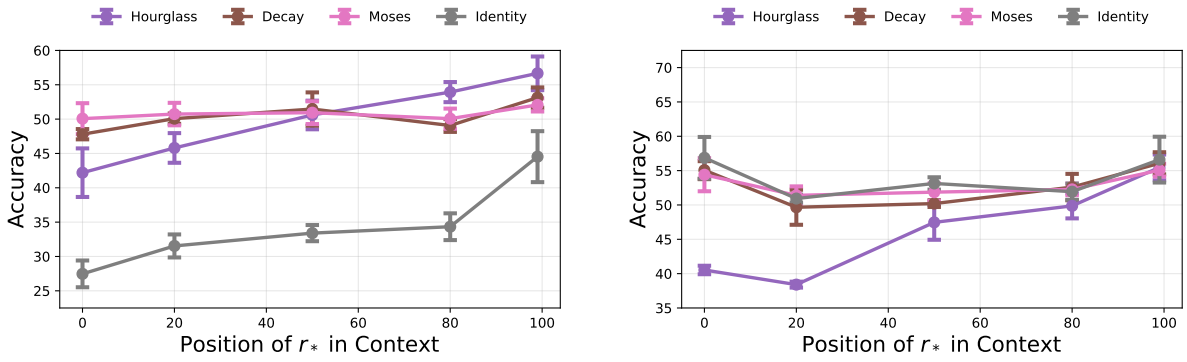
The Decay calibrator is competitive with the Moses calibrator, as for Llama-3.1-8B, it achieves severe improvement at middle positions while maintaining a slight recency bias. However, for Qwen-2.5-14B, the Decay calibrator in fact increases the severity of the U-shape, with $\rho_{\delta,a} = 4.1$ compared to the identity baseline's 3.0, representing an increase in the distance correlation.

The Hourglass calibrator produces an unexpected pattern: For Llama-3.1-8B, accuracy progressively improves from 42.2% at $d_* = 0$ to 56.7% at $d_* = d$, indicating a severe recency bias. Qwen-2.5-14B exhibits a similar pattern, with accuracy increasing from 40.5% to 55.5%.

6.2 Performance in Cross-Chunk Reasoning

The results of the cross-chunk reasoning experiments on Babilong-QA, shown in Figure 3 and Table 4, reveal further strengths of the calibrators.

The response to calibration varies dramatically between models, revealing important interactions between model capacity and calibration strategies. Under increasing distractor presence, the Moses calibrator demonstrates exceptional resilience for Qwen-2.5-14B. While the identity baseline degrades from 68.1% to 25.3% at 8K distractors, Moses degrades from 68.2% to 38.3%, resulting in a 13 %p improvement at high distractor presence versus unaltered positional encodings. Other



(a) Llama-3.1-8B

(b) Qwen-2.5-14B

Figure 2: Retrieval accuracy on Natural-QA measured by LLM-as-a-Judge using different calibration methods across varying context window positions. Colors indicate different calibration methods; y-axis ranges are adjusted per subplot.

Model	Position of r_* in retrieved information					$\rho_{\delta,a}$
	0%	**20%**	**50%**	**80%**	**100%**	
Llama-3.1-8B						
Hourglass	42.20 ± 3.54	45.80 ± 2.16	50.60 ± 2.08	53.93 ± 1.46	56.67 ± 2.45	-0.91
Decay	47.80 ± 0.75	50.07 ± 0.75	51.47 ± 2.42	49.07 ± 0.93	53.13 ± 1.48	-0.67
Moses	50.07 ± 2.25	50.73 ± 1.64	50.93 ± 1.65	50.07 ± 1.46	52.07 ± 0.96	0.13
Identity	27.47 ± 1.95	31.53 ± 1.68	33.40 ± 1.18	34.33 ± 1.95	44.53 ± 3.71	2.08
Qwen-2.5-14B						
Hourglass	40.53 ± 0.62	38.40 ± 0.43	47.47 ± 2.54	49.87 ± 1.82	55.47 ± 1.88	0.64
Decay	55.07 ± 1.31	49.67 ± 2.55	50.20 ± 0.49	52.60 ± 1.93	56.07 ± 1.61	4.13
Moses	54.40 ± 2.41	51.40 ± 1.31	51.87 ± 1.27	52.27 ± 0.84	55.07 ± 0.93	2.25
Identity	56.87 ± 3.03	50.93 ± 0.57	53.13 ± 0.90	51.93 ± 1.18	56.60 ± 3.34	2.97

Table 3: Retrieval accuracy (%) on Natural-QA measured by LLM-as-a-Judge matching for the different calibrators.

calibrators provide either modest benefits or slight degradation for this model, up to the point of 4k distractor tokens, where both Decay and Hourglass start outperforming the unaltered model.

In contrast, Llama-3.1-8B is not benefiting as much from Calibration. Even Moses, which provides substantial benefits to Qwen, slightly degrades Llama's reasoning performance. At 8K noise, Moses reduces Llama's accuracy from 12.5% to 10.8%, while Hourglass and Decay approach a low accuracy of 5%.

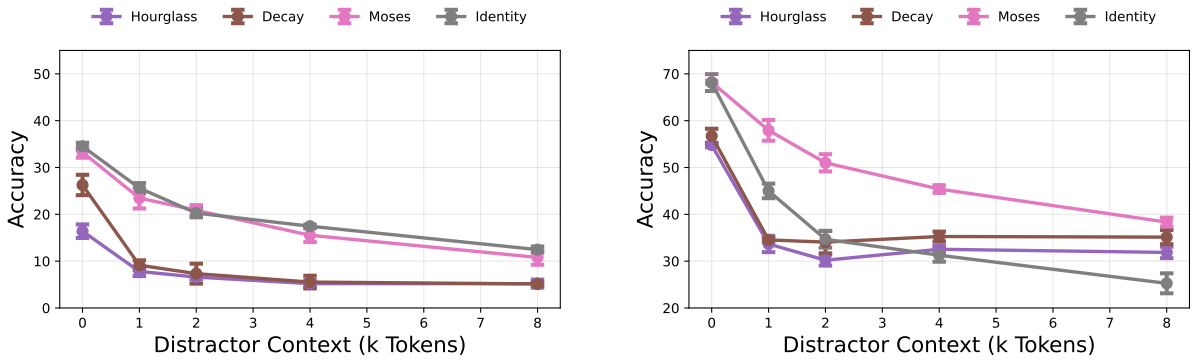
At 0K distractors, the Decay and Hourglass calibrators unexpectedly decrease performance. For Qwen-2.5-14B, the Hourglass calibrator reduces accuracy from 68.1% to 54.8%, while Decay drops to 56.7%. This effect is even more pronounced for Llama-3.1-8B, where Hourglass causes an 18.1 %p drop from 34.5% to 16.4%, and Decay results in an 8.3 %p decrease. A notable exception emerges with the Moses calibrator on Qwen-2.5-14B, which maintains performance at 68.2%, essentially matching the identity baseline of 68.1%.

7 Discussion

While models are frequently marketed with context windows exceeding hundreds of thousands of tokens, our baseline experiments confirm that context utilization degrades even at modest scales.

Across both experimental settings, our results demonstrate that calibrating positional encodings at inference time is a viable and surprisingly effective lever for improving long-context utilization. Despite leaving all model parameters untouched and using untuned calibration functions, CALIOPE can substantially mitigate positional biases in needle-in-a-haystack retrieval and can improve robustness under heavy distractor presence in cross-chunk reasoning, showcasing enhanced Long-Context Utilization. These findings suggest that positional encodings remain an underexploited control surface in modern LLMs, and that lightweight, training-free interventions can already yield meaningful gains.

At the same time, the heterogeneous effects observed across models and tasks point to non-ideal



(a) Llama-3.1-8B

(b) Qwen-2.5-14B

Figure 3: Cross-chunk reasoning accuracy on Babilong-QA as a function of distractor context length with different calibrators. Y-axis ranges are adjusted per subplot.

Model	Distractor noise (measured in k tokens)				
	0	**1**	**2**	**4**	**8**
Llama-3.1-8B					
Hourglass	16.40 ± 1.45	7.80 ± 1.00	6.60 ± 0.49	5.20 ± 0.75	5.20 ± 0.86
Decay	26.27 ± 2.17	9.13 ± 1.04	7.33 ± 2.13	5.53 ± 1.36	5.13 ± 0.34
Moses	33.27 ± 1.18	23.47 ± 2.22	20.87 ± 1.09	15.53 ± 1.46	10.80 ± 1.57
Identity	34.53 ± 0.75	25.60 ± 1.07	20.27 ± 0.90	17.47 ± 0.34	12.47 ± 0.66
Qwen-2.5-14B					
Hourglass	54.80 ± 0.49	33.67 ± 1.76	30.20 ± 1.14	32.53 ± 0.81	31.87 ± 1.23
Decay	56.73 ± 1.54	34.53 ± 0.47	34.07 ± 2.39	35.27 ± 1.05	35.13 ± 1.52
Moses	68.20 ± 0.33	57.93 ± 2.22	51.00 ± 1.84	45.40 ± 0.82	38.33 ± 0.96
Identity	68.13 ± 1.80	45.00 ± 1.56	34.67 ± 1.79	31.27 ± 1.39	25.27 ± 2.13

Table 4: Accuracy (%) measured by substring matching on Babilong-QA cross-chunk reasoning tasks for the different calibrators.

generalization of the proposed methods. The opposite trends between Llama-3.1-8B and Qwen-2.5-14B across retrieval and cross-chunk reasoning highlight that CALIOPE's benefits remain task- and model-specific. Taken together, our experiments position CALIOPE not as a silver bullet, but as a flexible and extensible framework that opens the door to task-aware, adaptive calibration strategies for long-context reasoning.

Crucially, the calibrators presented in this work were implemented with untuned parameters to establish their foundational efficacy. Therefore, the reported performance should be interpreted as a conservative lower bound on the potential of positional calibration in the CALIOPE framework. The performance shifts observed indicate that the models are highly sensitive to these modifications, strongly suggesting that a systematic hyperparameter search could unlock further gains. Future work should focus on optimizing these parameters, or potentially learn them via fine-tuning.

Another finding that is relevant for further investigation is the degraded performance of certain calibrators in cross-chunk reasoning when no distractor context is present. In the zero-distractor setting of the Babilong-QA dataset, the Hourglass and Decay calibrators reduce accuracy even when the two relevant facts constitute the only information provided. This behavior suggests that in short or minimally noisy contexts, in which relevant information is already compactly localized and positional biases are weak or absent, modifying positional encodings can unnecessarily perturb well-adapted attention patterns. Consequently, CALIOPE should be seen as a targeted intervention for long-context scenarios in which positional bias measurably impairs information utilization. For short-context tasks or standard reasoning problems without substantial distractor context, uncalibrated positional encodings may remain the preferable choice.

Real-world long-context applications span a broad spectrum, ranging from tasks that require

identifying a single relevant fact to more complex settings that demand integrating multiple distributed pieces of information. RAG represents one prominent instance of this spectrum, but similar challenges arise in document understanding, extended dialogue, and multi-step reasoning. Our results show that the CALIOPE framework provides a promising, training-free pathway for improving performance in such long-context scenarios. While the observed gains are task-dependent, they establish that strategic manipulation of positional encodings is a viable and effective tool for mitigating positional biases in practical long-context systems, including RAG.

8 Conclusion

This work investigated the "Lost in the Middle" phenomenon, confirming its persistence in modern LLMs, as well as another positional bias known as "recency bias", and examining the role of RoPE as a contributing factor. We introduced CALIOPE, a novel, training-free framework designed to mitigate these positional biases by recalibrating RoPE inputs at inference time.

Our empirical results demonstrate that strategic, training-free position remapping can substantially alleviate the LiM effect in single-document retrieval tasks on Natural-QA. The Moses and Decay calibrators, in particular, successfully flattened the performance curves of Llama-3.1-8B in the needle-in-a-haystack experiment. Additionally, the Hourglass substantially improved the performance of the Llama model, without flattening the performance curve. While not improving performance throughout positions, the calibrators remained competitive when evaluated on the Qwen-Model, with the exception of the Hourglass calibrator.

A more nuanced and complex picture emerged in our cross-chunk reasoning experiments on the Babilong-QA dataset. Llama-3.1-8B saw dramatic performance gains from calibration in the single-document retrieval task, but had its cross-chunk reasoning ability slightly impaired by the same methods. Conversely, Qwen-2.5-14B, which gained little from calibration in the simpler task, benefited substantially in the cross-chunk scenario.

In summary, this paper provides a systematic investigation of RoPE's contribution to the LiM phenomenon, offering a practical mitigation framework and key insights into the task-dependent nature of positional biases. While our findings present a viable path toward better context utilization, they also underscore the need for future research into task-aware calibration methods that can dynamically adapt to the specific reasoning demands of a given query.

Limitations

The experimental framework, while providing valuable insights into RoPE calibration strategies, is subject to several constraints that warrant careful consideration. The empirical evaluation was conducted exclusively on models with relatively modest parameter counts, specifically Llama-3.1-8B, and Qwen-2.5-14B. This choice of model scale potentially overlooks emergent behaviors that manifest only in larger architectures. Understanding whether and how the effectiveness of CALIOPE changes with increasing model size is an important direction for future work.

The rapidly evolving landscape of language model development presents an additional temporal constraint on this research. With new research emerging in high frequencies, the specific findings regarding RoPE calibration may face obsolescence as foundational architectural choices evolve.

A critical limitation of the current approach lies in its exclusive focus on post-training calibration that leave model parameters unchanged. This constraint, while ensuring computational efficiency and eliminating the need for additional training infrastructure, potentially overlooks calibration strategies that could achieve superior performance through minimal parameter adaptation. Fine-tuning approaches that allow models to adapt their internal representations to modified positional encodings may unlock performance improvements unattainable through parameter-frozen calibration alone. The interaction between learned attention patterns and modified positional encodings represents a complex optimization landscape that remains unexplored in this work.

Furthermore, an unexpected observation emerged during experimentation: certain calibrators demonstrated degraded performance even without the presence of distractors, suggesting that the calibration functions may introduce unintended perturbations to the model's baseline capabilities. This phenomenon requires further investigation through comprehensive ablation studies examining the interaction between different calibrator designs and model architectures.

A key limitation of our approach is that calibration hyperparameters were not systematically optimized. The performance observed with untuned parameters should therefore be interpreted as stablishing feasibility rather than optimal performance. Systematic hyperparameter search tailored to specific model architectures and context lengths represents a clear direction for improving calibration effectiveness.

Acknowledgments

Tom Zehle received funding by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. This work was supported by the European Union's Horizon Europe research and innovation programme under grant agreement No 101214398 (ELLIOT).

Matthias Aßenmacher received funding from the BERD@NFDI consortium in the context of the work of the National Research Data Infrastructure (NFDI) Association. NFDI is funded by the Federal Republic of Germany and the 16 federal states. The BERD@NFDI consortium is supported within NFDI by the German Research Foundation (DFG) – NFDI 27/1-2026, project number 460037581.

References

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. 2024. Make your llm fully utilize the context. *Preprint*, arXiv:2404.16811.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. 2024. Round and round we go! what makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*.

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, Liu YiBo, Sun Qianguo, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaxing Zhang. 2024. Never lost in the middle: Mastering long-context question answering with position-agnostic decompositional training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13628–13642, Bangkok, Thailand. Association for Computational Linguistics.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long T Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and 1 others. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. *arXiv preprint arXiv:2406.16008*.

Albert Qiaochu Jiang, Xiaohan Ren, Yixiao Chen, Michael Zhang, Tianle Wang, and 1 others. 2024. Longllmlingua: Compressing prompts for accelerating long-context llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yuri Kuratov, Mikhail Belyaev, Nikita Smetanin, Aleksandr Shevchenko, Aleksandr Krotov, Maxim Arkhipov, and Andrey Trusov. 2024. Babilong: Benchmarking long-context reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th*

symposium on operating systems principles, pages 611–626.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Changqun Li, Linlin Wang, Xin Lin, Gerard de Melo, and Liang He. 2022. Curriculum prompt learning with self-training for abstractive dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1096–1106. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Percy Liang, and Tatsunori Hashimoto. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Koichi Nagatsuka, Clifford Broni-Bediako, and Masayasu Atsumi. 2023. Length-based curriculum learning for efficient pre-training of language models. *New Generation Computing*, 41(1):109–134.

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

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting for in-context learning. In *Advances in Neural Information Processing Systems*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2, pages 464–468.

Zhen Sun, Yujie Zhang, Jian Zhang, Qiang Liu, Zhiqiang Wang, and 1 others. 2021. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Jie Wang, Tao Ji, Yuanbin Wu, Hang Yan, Tao Gui, Qi Zhang, Xuanjing Huang, and Xiaoling Wang. 2024. Length generalization of causal transformers without position encoding. *arXiv preprint arXiv:2404.12224*.

Ruochen Xiong, William Fedus, and 1 others. 2023. Effective long-range modeling with rotary position embeddings. In *NeurIPS*.

Yijiong Yu, Huiqiang Jiang, Xufang Luo, Qianhui Wu, Chin-Yew Lin, Dongsheng Li, Yuqing Yang, Yongfeng Huang, and Lili Qiu. 2025. Mitigate position bias in llms via scaling a single hidden states channel. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6092–6111.

Appendix

A Full Results

A.1 Reproduction Study

The results of the reproduction study measured by Substring-Matching can be found in Table 5 and Figure 4 respectively.

A.2 Evaluation of Calibrators

The results of the Needle-in-a-Haystack experiments, evaluated by Substring-Matching can be found in Table 6 and Figure 5 respectively.

B Model Details

Detailed specifications of the Large Language Models (LLMs) are provided in Table 7, including their HuggingFace identifiers and specific revisions. For local inference, we employed the vLLM library (version 0.7.3, Kwon et al., 2023), selected for its high-throughput serving capabilities, efficient memory management, and native support for quantized models. We configured the generation with a maximum output length of 2048 tokens.

Model	Huggingface ID
Llama-3.3-70B	shuyuej/Llama-3.3-70B-Instruct-GPTQ
Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct
Qwen-2.5-14B	Qwen/Qwen2.5-14B-Instruct-GPTQ-Int4

Table 7: Overview of the utilized LLMs.

C Hardware Details

All experiments were conducted on a GPU cluster equipped with NVIDIA A100 (80GB) and H100 (94GB) GPUs. Each model was hosted on a single GPU, and we leveraged the cluster to run multiple experimental configurations in parallel. The total computation for these experiments amounted to 13 GPU days.

D Dataset Details

In our experiments, we utilize the following two datasets:

1. **Natural-QA** (Kwiatkowski et al., 2019): An open-domain question answering dataset. To evaluate long-context retrieval, we adopt the Needle-in-a-Haystack setup proposed by Liu et al. (2024). Using the functionality provided in their supplementary material, we embed the single relevant context chunk required to answer a question within a larger body of distractor chunks. The model is then tasked with finding the answer within this extended context.

2. **Babilong-QA** (Kuratov et al., 2024): A suite of synthetic long-context benchmarks designed to assess specific reasoning capabilities. We evaluate models on the Question Answering 2 (QA 2) task across varying context lengths. In this task, a model must integrate two facts, that are hidden within distractor context, to answer a given question.

We provide detailed IDs and sizes of the utilized datasets in Table 8. We use 500 data points from the test set for each dataset per seed. The used seeds are 7, 42, and 47.

Dataset	Huggingface ID	n_{train}	n_{test}
Babilong-QA	RMT-team/babilong-1k-samples	–	20k
Natural-QA	google-research-datasets/natural_questions	10.6k	7.8k

Table 8: Overview of the utilized HuggingFace datasets.

E Implementation

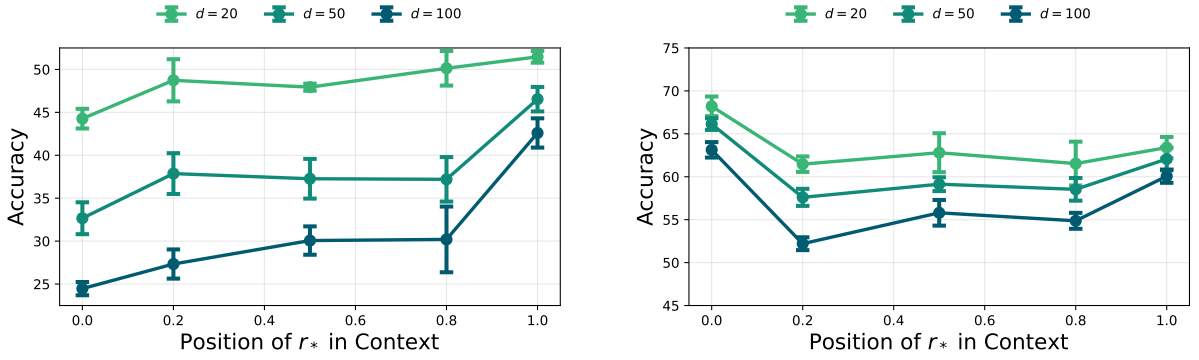
The calibration framework employs a PyTorch hook-based approach to intercept and modify position encodings without altering the underlying model architecture. Retrieved chunk boundaries are tracked by inserting unused tokens from the model's vocabulary, such as `<doc_start>` and `<doc_end>`, into the raw string dataset. These tokens, while semantically meaningless in context, serve as markers to delineate the boundaries of retrieved information.

The implementation attaches a `forward_pre_hook` to the model's embedding layer to identify these boundary tokens during the forward pass. This mechanism maintains a mapping between token positions and their corresponding retrieved chunk indices throughout the sequence, storing retrieved chunk boundaries for subsequent use in position calibration.

Position encoding modification occurs through a `forward_hook` attached to the RoPE positional encoding layer. This hook intercepts the `position_ids` tensor before it enters the encoding function, applies the position remapping $\Phi(t)$ based on the selected calibration algorithm, and returns modified position IDs that implement the desired spacing strategy.

Model	Position of r_* in retrieved information					$\rho_{\delta,a}$
	0%	**20%**	**50%**	**80%**	**100%**	
Llama-3.1-8B						
$d = 20$	44.27 ± 1.15	48.73 ± 2.46	47.93 ± 0.41	50.13 ± 2.02	51.47 ± 0.68	-0.36
$d = 50$	32.67 ± 1.86	37.87 ± 2.39	37.27 ± 2.31	37.20 ± 2.59	46.53 ± 1.43	1.65
$d = 100$	24.47 ± 0.77	27.33 ± 1.70	30.07 ± 1.65	30.20 ± 3.83	42.60 ± 1.70	2.93
Qwen-2.5-14B						
$d = 20$	68.20 ± 1.14	61.47 ± 0.90	62.80 ± 2.28	61.53 ± 2.55	63.40 ± 1.23	2.64
$d = 50$	66.13 ± 0.68	57.60 ± 1.00	59.13 ± 0.81	58.53 ± 1.32	62.07 ± 1.31	4.10
$d = 100$	63.13 ± 0.90	52.20 ± 0.75	55.80 ± 1.50	54.87 ± 0.93	60.07 ± 0.77	4.80

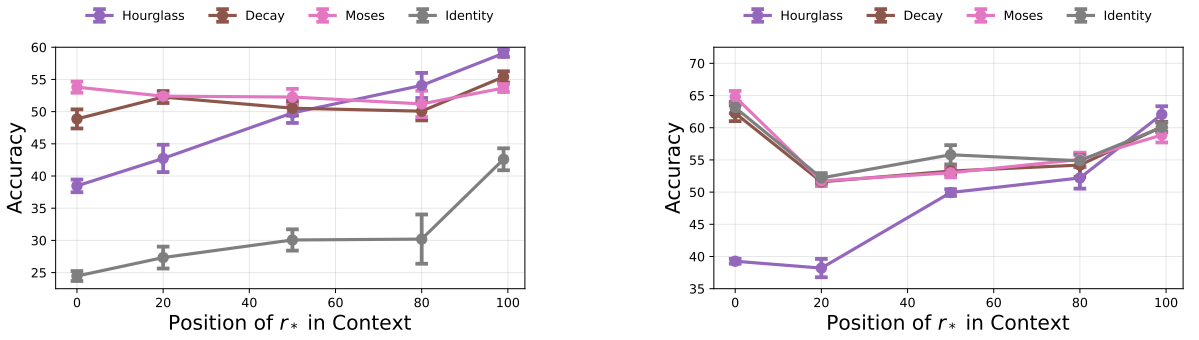
Table 5: Retrieval accuracy (%) measured by substring matching.



(a) Llama-3.1-8B

(b) Qwen-2.5-14B

Figure 4: Comparison of Accuracy across varying context window positions. The x-axis represents the normalized position of the relevant information within the context window. Note that y-axis ranges are adjusted per subplot to emphasize patterns while maintaining consistent scale spans across all plots.



(a) Llama-3.1-8B

(b) Qwen-2.5-14B

Figure 5: Retrieval accuracy measured by Substring Matching using different calibration methods across varying context window positions. Colors indicate different calibration methods; y-axis ranges are adjusted per subplot.

Model	Position of r_* in retrieved information					$\rho_{\delta,a}$
	0%	**20%**	**50%**	**80%**	**100%**	
Llama-3.1-8B						
Hourglass	38.47 ± 0.98	42.73 ± 2.12	49.80 ± 1.56	54.07 ± 1.95	59.07 ± 0.57	-0.78
Decay	48.87 ± 1.48	52.27 ± 0.93	50.53 ± 1.15	50.07 ± 1.43	55.40 ± 0.86	1.15
Moses	53.80 ± 0.86	52.40 ± 0.33	52.27 ± 1.27	51.20 ± 2.05	53.67 ± 0.62	1.18
Identity	24.47 ± 0.77	27.33 ± 1.70	30.07 ± 1.65	30.20 ± 3.83	42.60 ± 1.70	2.93
Qwen-2.5-14B						
Hourglass	39.27 ± 0.38	38.20 ± 1.42	49.93 ± 0.52	52.20 ± 1.66	62.07 ± 1.27	0.87
Decay	62.27 ± 1.25	51.60 ± 0.43	53.27 ± 0.52	54.20 ± 1.85	60.13 ± 0.77	6.31
Moses	64.80 ± 0.91	51.73 ± 0.81	53.00 ± 0.71	55.07 ± 1.06	58.87 ± 1.16	6.99
Identity	63.13 ± 0.90	52.20 ± 0.75	55.80 ± 1.50	54.87 ± 0.93	60.07 ± 0.77	4.80

Table 6: Retrieval accuracy (%) measured by substring matching for the different calibrators.