THE SIRENS' SONG: WHEN PROXIMAL BACKGROUND CONTEXT OVERSHADOWS DISTANT EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-context LLMs focus on retrieving distant evidence from extensive context, yet existing work has largely focused on overcoming distance alone. In this work, we identify the Proximity Trap, insufficient attention to distant evidence often arises less from distance itself than from cumulative competition with abundant, task-irrelevant proximal background. To address the Proximity Trap, we introduce LYRA (**L**ong-context heav**Y**-tailed **R**elevance **A**lignment), a t-distributed directional matching mechanism that reshapes the context retrieval distribution, directing more attention mass toward task-relevant evidence, while preserving the relative positional information encoded. Extensive experiments on LongBench-v2, RULER, and LongBench demonstrate consistent improvements across context lengths and task categories. We further introduce ProxBench, a multi-level fine-grained benchmark for evaluating distant evidence utilization under increasing proximal background interference. Our code is public at `https://anonymous.4open.science/r/LYRA`

1 INTRODUCTION

Recent work on long-context LLMs (Wang et al., 2026b; Zhang et al., 2026; Mudarisov et al., 2025) has focused on improving access to distant evidence. However, access alone does not ensure that such evidence will be used, while proximal, task-irrelevant context may also compete for attention. This raises a less examined question: when a model overlooks distant evidence, is distance alone responsible, or does proximal background also impede its use?

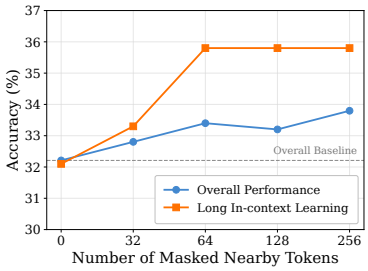


Figure 1: **Less attention to proximal background improves long-context understanding.** On LongBench-v2 (Bai et al., 2025), we progressively mask the context tokens nearest to the query and report accuracy over the full benchmark and the Long In-context Learning subset. The horizontal dashed line denotes the performance without masking, corresponding to zero masked tokens.

Prioritizing recent context is a reasonable inductive bias, as nearby tokens often provide useful information for the current prediction. Accordingly, modern LLMs commonly encode relative positions through RoPE (Su et al., 2024), making QK scores explicitly dependent on relative distance and potentially attenuating the scores of distant tokens. While this positional bias supports local context modeling, positional proximity does not necessarily indicate task relevance: essential evidence may lie far from the query, whereas nearby content may provide little information for the required prediction. This mismatch raises uncertainty about how attention is allocated when distant evidence competes with proximal background.

In this work, we present a counter-intuitive observation: reducing attention to proximal background can improve long-context understanding. As shown in Fig. 1, masking proximal background tokens consistently improves the overall accuracy on LongBench-v2 compared with the unmasked baseline, with a more pronounced improvement on Long In-context Learning. This result is striking because removing accessible context improves performance rather than degrading it. It suggests that the difficulty of

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.

using distant evidence cannot be attributed to distance alone; competition from proximal background also plays a critical role. We refer to this phenomenon as the *Proximity Trap*.

This counter-intuitive observation is closely related to how attention combines positional distance with contextual competition. Long-context inputs are inherently non-uniform: the evidence required for a prediction is often sparse and distant, whereas much of the nearby context serves only as background. Relative positional encoding makes QK matching dependent on distance and can weaken the directional agreement between the query and distant evidence. Under softmax normalization, however, this weakened evidence competes not with a single nearby token, but with all proximal background tokens simultaneously. Although each background token may have limited task relevance, their contributions accumulate in the softmax denominator and can collectively draw attention away from distant evidence. This interaction between positional attenuation and cumulative background competition underlies the Proximity Trap and explains why masking proximal background can improve evidence utilization.

Consequently, we introduce LYRA, a simple t-distributed directional matching mechanism. Preserving distant evidence does not require explicitly favoring distant positions or manually suppressing proximal context. Instead, LYRA measures the directional agreement between the RoPE-transformed query and keys, and then applies a t-distributed transformation to reshape their score differences before softmax normalization. This design reduces misleading advantages held by weakly related proximal background while making strongly matched evidence more distinguishable. Crucially, LYRA replaces only the conventional QK scoring function, leaving RoPE, causal masking, softmax normalization, and value aggregation unchanged. As a result, LYRA preserves distant evidence according to relevance rather than position, allowing it to resist cumulative background competition without indiscriminately promoting all distant tokens.

In summary, our main contributions are as follows:

- We characterize the Proximity Trap, a counter-intuitive phenomenon in which proximal, task-irrelevant background can collectively overshadow distant but relevant evidence.
- LYRA is introduced as a t-distributed directional matching mechanism that reshapes the context retrieval distribution to redistribute attention mass toward task-relevant evidence while preserving the relative positional information encoded.
- Extensive experiments are conducted on various benchmarks. The results demonstrate consistent improvements across different context lengths and task categories, supporting the effectiveness and generalization of LYRA.
- We introduce ProxBench, a multi-level controlled benchmark for evaluating distant evidence utilization under increasing proximal background interference.

2 The Proximity Trap

In this section, we first introduce the preliminaries in Section 2.1. Subsequently, we establish our central observation that proximal background context overshadows distant evidence in Section 2.2. Moreover, we provide an explanation of its underlying mechanism through cumulative attention competition in Section 2.3.

2.1 Preliminaries

Given a query at the current decoding position t, let $\mathbf{q}_t \in \mathbb{R}^d$ denote the query representation and let $(\mathbf{k}_i, \mathbf{v}_i)$ denote the key–value pair at each historical position $i < t$. Standard attention first computes the QK score between the query and each key, followed by softmax normalization:

$$s_{t,i} = \frac{\mathbf{q}_t^\top \mathbf{k}_i}{\sqrt{d}}, \qquad a_{t,i} = \frac{\exp(s_{t,i})}{\sum_{j<t} \exp(s_{t,j})}. \tag{1}$$

The resulting attention output is $\mathbf{o}_t = \sum_{i<t} a_{t,i}\mathbf{v}_i$. Since $a_{t,i} \geq 0$ and $\sum_{i<t} a_{t,i} = 1$, all historical positions compete for a finite amount of attention weight. Our analysis therefore focuses on how the QK scores assigned to different historical positions determine this competition when the model generates the token at position t.

In particular, the QK score defined in Eq. (1) becomes explicitly position-dependent when RoPE (Su et al., 2024) is applied to the query and key representations. Let $\mathbf{R}_m \in \mathbb{R}^{d\times d}$ denote the rotation matrix associated with position m. RoPE transforms the query and key as $\widetilde{\mathbf{q}}_t = \mathbf{R}_t\mathbf{q}_t$ and $\widetilde{\mathbf{k}}_i = \mathbf{R}_i\mathbf{k}_i$, respectively. The resulting QK score is:

$$s_{t,i}^{\mathrm{RoPE}} = \frac{\widetilde{\mathbf{q}}_t^\top \widetilde{\mathbf{k}}_i}{\sqrt{d}} = \frac{\mathbf{q}_t^\top \mathbf{R}_t^\top \mathbf{R}_i \mathbf{k}_i}{\sqrt{d}} = \frac{\mathbf{q}_t^\top \mathbf{R}_{-(t-i)} \mathbf{k}_i}{\sqrt{d}}, \tag{2}$$

where $\mathbf{R}_t^\top \mathbf{R}_i = \mathbf{R}_{-(t-i)}$ depends only on the relative offset $t-i$, where $\Delta = t - i \geq 0$ denotes the relative distance. Consequently, RoPE makes the QK score jointly dependent on the content representations $\mathbf{q}_t$ and $\mathbf{k}_i$ and their relative position.

Building on the relative-position-dependent scoring formulation above, we denote task-relevant information far from the query as Distant Evidence $\mathcal{E}$ and nearby content not required for the prediction as Proximal Background $\mathcal{B}$. Importantly, $\mathcal{B}$ does not need to be adversarial, corrupted, or linguistically unnatural. It may consist of ordinary contextual or grammatical content that is coherent within the input but uninformative for the required prediction.

Remark 2.1. *Counterintuitively, distance does not determine task relevance: $\mathcal{E}$ can be distant but essential, whereas $\mathcal{B}$ can be nearby but uninformative. Nevertheless, tokens in both spans enter the same softmax normalization and therefore compete for attention weight.*

2.2 Proximal Background Context Overshadows Distant Evidence

Building on the preceding definitions, we conduct two complementary experiments to examine the competition between distant evidence $\mathcal{E}$ and proximal background $\mathcal{B}$. The first determines whether pretrained LLMs fail to locate distant evidence or instead retrieve it but have its signal diluted by proximal background. The second compares moving the evidence closer with attenuating proximal background to identify which intervention more effectively restores evidence attention and answer confidence.

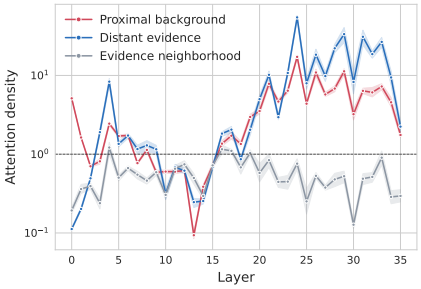


Figure 2: **Distant Evidence is selectively retrieved but still diluted by cumulative competition from proximal background.** We report layer-wise attention density for distant evidence, its neighboring tokens, and proximal background near the query.

First, to distinguish whether the model fails to locate distant evidence or loses it under attention competition, we compare the layer-wise attention assigned to three regions: distant evidence, its neighboring tokens, and proximal background near the query. For each layer, we report the mean per-token attention density normalized by uniform attention, where a value of 1 denotes the uniform baseline. These statistics are computed using Qwen3-8B (Bai et al., 2025) on LongBench-v2 (Bai et al., 2025).

As illustrated in Fig. 2, the three regions are not consistently separated in the early and middle layers. From approximately layer 19 onward, however, attention to distant evidence increases sharply, whereas its neighborhood remains largely below the uniform baseline. This contrast shows that the model can precisely identify distant evidence among ordinary tokens at the same distant location. However, proximal background also receives substantial attention. Although its per-token attention is generally lower than that of distant evidence, its abundance produces sufficient cumulative attention mass to overshadow the evidence. The central issue is therefore not whether the model can locate distant evidence, but whether that evidence can withstand competition from abundant proximal background after being identified.

Subsequently, building on this observation, we examine whether evidence utilization in Qwen3-8B is constrained primarily by evidence distance or by competition from proximal background. Figure 3 compares two interventions, moving distant evidence closer and attenuating proximal background, and reveals their distinct effects on evidence attention and answer confidence. For each evidence-dependent question, we compare moving distant evidence closer with proximal background attenuation, where the latter suppresses QK scores of nearby task-irrelevant context without changing

the input tokens or their positions. We report the resulting changes in Evidence Attention and correct-answer confidence relative to the original input.

Accordingly, as Fig. 3 illustrates, the blue points span all four quadrants, indicating that moving distant evidence closer does not consistently improve the performance. In contrast, the orange points are concentrated toward increased evidence attention, with many also showing improved confidence. This divergence suggests that shortening the evidence distance does not remove competition from abundant proximal background. Attenuating proximal background instead directly weakens this competition, allowing more attention to be reassigned to relevant evidence. Therefore, compared with moving distant evidence closer, proximal background attenuation produces a more consistent improvement in evidence utilization.

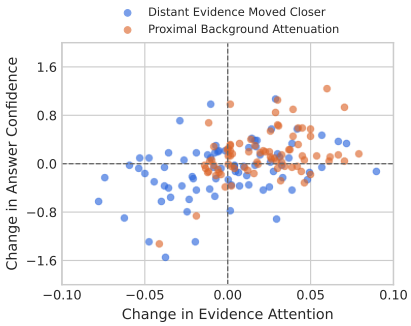


Figure 3: **Proximal background limits its evidence utilization in Qwen3-8B.** Points show paired changes in evidence attention and answer confidence. Blue moves evidence closer, while orange attenuates proximal background scores with positions fixed.

Taken together, these two experiments reveal the following counter-intuitive finding:

Finding 2.2 (The Proximity Trap). *Insufficient attention to distant evidence often arises less from distance itself. Instead, it is more from cumulative competition with abundant task-irrelevant proximal background context. In this sense, proximal background acts as the Sirens' Song by drawing attention toward nearby but uninformative content and away from the distant evidence that matters.*

2.3 Mechanism: Proximal Background Overshadowing

To better understand why task-relevant distant evidence can be overshadowed by task-irrelevant proximal background, we examine how relative position alters QK scores and how softmax converts these changes into attention competition.

Decomposing Eq. (2) into $d/2$ two-dimensional components, the QK score in the r-th component is

$$\mathbf{q}_r^\top \mathbf{R}(-\Delta\omega_r)\mathbf{k}_r = a_r \cos(\Delta\omega_r) + b_r \sin(\Delta\omega_r), \tag{3}$$

where the content-dependent coefficients are defined as $a_r = q_{r,1}k_{r,1} + q_{r,2}k_{r,2}$ and $b_r = q_{r,1}k_{r,2} - q_{r,2}k_{r,1}$, and ω_r denotes the rotation frequency. Summing Eq. (3) across all components gives

$$s(\Delta) = \frac{1}{\sqrt{d}} \sum_{r=1}^{d/2} \left[a_r \cos(\Delta\omega_r) + b_r \sin(\Delta\omega_r)\right]. \tag{4}$$

Because RoPE rotations are orthogonal, they preserve the query and key norms, namely $\|\widetilde{\mathbf{q}}_t\|_2 = \|\mathbf{q}_t\|_2$ and $\|\widetilde{\mathbf{k}}_i\|_2 = \|\mathbf{k}_i\|_2$. Thus, distance changes the directional match between the query and key across frequencies rather than reducing their vector lengths.

To isolate the positional effect in its simplest form, we consider a maximally aligned query–key pair with $\mathbf{q}_r = \mathbf{k}_r$. This serves only as an illustrative reference case and is not assumed to hold in pretrained LLMs. In this case, $b_r = 0$ and $a_r = \|\mathbf{q}r\|_2^2$, reducing Eq. (4) to

$$s(\Delta) = \frac{1}{\sqrt{d}} \sum_{r=1}^{d/2} \|\mathbf{q}_r\|_2^2 \cos(\Delta\omega_r). \tag{5}$$

For $|\Delta\omega_r| \ll 1$, a second-order expansion gives:

$$s(\Delta) \approx s(0) - \frac{\Delta^2}{2\sqrt{d}} \sum_{r=1}^{d/2} \|\mathbf{q}_r\|_2^2 \omega_r^2. \tag{6}$$

Crucially, score attenuation alone does not determine the evidence attention. Consider one distant evidence token with QK score $s_{\mathcal{E}}$ and M proximal background tokens, each with QK score $s_{\mathcal{B}}$. Their softmax competition gives

$$\alpha_{\mathcal{E}} = \frac{1}{1 + Me^{s_{\mathcal{B}} - s_{\mathcal{E}}}}, \qquad \alpha_{\mathcal{E}} < \frac{1}{2} \iff s_{\mathcal{E}} - s_{\mathcal{B}} < \log M. \tag{7}$$

Accordingly, Eq. (7) explains both experiments in Section 2.2. Fig. 2 shows that although the model attends to distant evidence, abundant proximal background can collectively overshadow it through M. Fig. 3 further shows that attenuating proximal background restores evidence attention and answer confidence more consistently than moving the evidence closer. Thus, the primary obstacle is cumulative background competition rather than distance alone.

Together, these results distinguish vulnerability from failure: distance can weaken distant evidence, but insufficient evidence attention arises when abundant proximal background accumulates in the softmax denominator and collectively overshadows the evidence. This mechanism motivates an attention formulation that limits cumulative background competition while preserving access to distant evidence, leading to the LYRA formulation introduced next.

3 LYRA FOR DISTANT EVIDENCE

The findings in Section 2 reveal that distant evidence can be overshadowed by proximal background context, even when the latter contributes little task-relevant information. In this section, we introduce LYRA (**L**ong-context heav**Y**-tailed **R**elevance **A**lignment), a t-distributed directional matching method designed to preserve distant yet relevant evidence in Section 3.1, and analyze how it mitigates background-induced suppression in Section 3.2.

3.1 LYRA: T-DISTRIBUTED QK MATCHING

As established in Section 2, Eq.(7) exhibits that even a moderate reduction in $s_{\mathcal{E}}$ can be exponentially amplified by softmax normalization, while the cumulative suppression from proximal background grows with M. A suitable mechanism should therefore preserve distant evidence under moderate score attenuation, distinguish evidence from background according to directional alignment with the query, and avoid indiscriminately increasing the weights of all distant tokens.

Motivated by these requirements, we introduce LYRA, a heavy-tailed directional matching mechanism that preserves attenuated yet relevant evidence without indiscriminately promoting distant context. Given the RoPE-transformed query $\widetilde{\mathbf{q}}_t$ and key $\widetilde{\mathbf{k}}_i$, we first measure their directional agreement using the normalized similarity $c_{t,i} = \frac{\widetilde{\mathbf{q}}_t^\top \widetilde{\mathbf{k}}_i}{\|\widetilde{\mathbf{q}}_t\|_2 \|\widetilde{\mathbf{k}}_i\|_2} \in [-1, 1]$. We then reshape this similarity using the t-distributed transformation:

$$\phi_\kappa(c_{t,i}) = \frac{1 + c_{t,i}}{1 + \kappa(1 - c_{t,i})} - 1, \qquad \kappa \geq 0, \tag{8}$$

where κ controls the angular concentration parameter and $\phi_0(c_{t,i}) = c_{t,i}$ recovers the original cosine similarity. The resulting attention weight is defined as $\alpha^{LYRA} = \mathrm{softmax}_i\left(\beta \phi_\kappa(c_{t,i})\right)$, where $\beta > 0$ is an inverse-temperature parameter. This formulation increases the contrast between highly aligned evidence and weakly aligned background without introducing an explicit position-dependent increase for distant tokens.

Operationally, LYRA modifies only the query-key scoring stage of attention. Instead of directly using the conventional QK score, it computes the directional similarity between the RoPE-transformed query and key and maps it to the score $\beta\phi_\kappa(c_{t,i})$ through Eq. (8). RoPE, causal masking, softmax normalization, and value aggregation remain unchanged. LYRA can therefore be integrated into an existing attention layer by replacing its QK scoring function without modifying the remaining architecture. The detailed implementation is provided in Appendix C.

3.2 PRESERVING DISTANT EVIDENCE UNDER BACKGROUND COMPETITION

To characterize how LYRA locally reshapes directional similarity, we consider the derivative of the transformation in Eq. (8) and the critical similarity at which this derivative equals one:

$$\phi'_\kappa(c) = \frac{1+2\kappa}{[1+\kappa(1-c)]^2} > 0, \qquad c^*_\kappa = 1 - \frac{2}{\sqrt{1+2\kappa}+1}. \tag{9}$$

The positive derivative shows that ϕ_κ is strictly increasing and therefore preserves the original similarity ordering. For $\kappa > 0$, c^*_κ divides the similarity domain into a compression regime with $c < c^*_\kappa$ and $\phi'_\kappa(c) < 1$, and an amplification regime with $c > c^*_\kappa$ and $\phi'_\kappa(c) > 1$. The transformation therefore reduces similarity differences in the former regime and enlarges them in the latter.

To extend the local characterization above to the finite difference between distant evidence and proximal background, we define the similarity-gap scaling factor for $c_\mathcal{B} \neq c_\mathcal{E}$ as

$$G_\kappa(c_\mathcal{B}, c_\mathcal{E}) = \frac{\phi_\kappa(c_\mathcal{B}) - \phi_\kappa(c_\mathcal{E})}{c_\mathcal{B} - c_\mathcal{E}} = \frac{1+2\kappa}{[1+\kappa(1-c_\mathcal{B})][1+\kappa(1-c_\mathcal{E})]}. \tag{10}$$

Accordingly, the transformed gap satisfies $\phi_\kappa(c_\mathcal{B}) - \phi_\kappa(c_\mathcal{E}) = G_\kappa(c_\mathcal{B}, c_\mathcal{E})(c_\mathcal{B} - c_\mathcal{E})$. As the secant slope of ϕ_κ between the two similarities, $G_\kappa < 1$ indicates gap compression, whereas $G_\kappa > 1$ indicates gap amplification. Compared with cosine-softmax under the same inverse temperature β, the relative change in the evidence-to-background attention ratio is

$$\frac{\alpha^{LYRA}_\mathcal{E} / \alpha^{LYRA}_\mathcal{B}}{\alpha^{\cos}_\mathcal{E} / \alpha^{\cos}_\mathcal{B}} = \exp\left[\beta(1 - G_\kappa)(c_\mathcal{B} - c_\mathcal{E})\right]. \tag{11}$$

The relative weight of evidence is therefore improved when $(1 - G_\kappa)(c_\mathcal{B} - c_\mathcal{E}) > 0$. When proximal background has a higher similarity than distant evidence, $G_\kappa < 1$ compresses the background advantage. When evidence retains a higher similarity, $G_\kappa > 1$ enlarges the evidence advantage against cumulative background competition. Consequently, LYRA reduces the influence of weakly related background tokens while making strongly matched evidence more distinguishable, thereby directing attention toward task-relevant information.

Beyond competition from a single proximal background token, we extend the analysis to M proximal background tokens, which does not change above improvement condition. The factor M increases the cumulative competing mass, while G_κ determines whether the score gap is adjusted in favor of distant evidence. Therefore, LYRA mitigates the Proximity Trap by narrowing a misleading background advantage or strengthening an existing evidence advantage before competition accumulates across background tokens.

3.3 PROXBENCH: FINE-GRAINED PROXIMAL PERTURBATIONS FOR LONG-CONTEXT

Existing long-context benchmarks primarily evaluate whether models can retrieve and integrate information as the input length increases. However, they provide limited control over the proximal background context that competes with distant evidence. As a result, it remains difficult to determine whether a model succeeds by robustly identifying task-relevant evidence or fails because semantically confusable information near the query captures excessive attention. To address this limitation, we introduce PROXBENCH, a controlled benchmark designed to evaluate the robustness of long-context models against proximal background interference. PROXBENCH places sparse evidence at a distant position and introduces task-irrelevant but increasingly confusable background context near the query, thereby directly measuring the Proximity Trap under controlled conditions.

PROXBENCH organizes proximal perturbations into four progressively challenging levels. Level 1 introduces style-matched background that follows a similar syntactic structure but differs in entity, topic, and answer relation. Level 2 creates crossed bindings by placing the target entity and an access-code cue in the same sentence while explicitly associating the candidate value with another entity. Level 3 further increases ambiguity by mixing the same relation for different entities with different attributes of the target entity. Level 4 introduces fine-grained perturbations involving subtypes, attributes, semantic roles, and value formats, producing background context that closely resembles the required evidence while remaining logically irrelevant to the answer. This hierarchy enables a fine-grained assessment of how model performance changes as proximal background becomes increasingly difficult to distinguish from distant evidence. Further details on data construction, quality control, and dataset statistics are provided in Appendix B.

4 EXPERIMENTS

We organize our evaluation around three complementary aspects. First, in Section 4.1, we evaluate LYRA on LongBench-v2 (Bai et al., 2025) and RULER (Hsieh et al., 2024), examining its ability to retrieve and utilize distant evidence under increasingly long inputs. Second, in Section 4.2, we report results across the task categories of LongBench to assess whether the benefits of LYRA generalize beyond specific long-context settings. Third, in Section 4.3, we evaluate LYRA on our proposed ProxBench, where progressively challenging perturbation levels are used to directly measure its robustness to different forms of proximal background interference. Additional ablation studies are provided in the Appendix D to examine the contribution and sensitivity of the key design in LYRA.

Experimental Settings. We adopt Qwen3-8B (Yang et al., 2025) as the base model and replace the standard attention mechanism only in its final Transformer block with LYRA, leaving the remaining architecture unchanged. We then fine-tune the model on LongAlign for one epoch, updating only the final Transformer block while freezing all preceding layers, token embeddings, and the language-modeling head. Training is performed in bfloat16 using AdamW, with a learning rate of 2×10^{-5}, and a maximum sequence length of 16,384. We evaluate long-context retrieval and understanding on RULER and LongBench-v2, and assess robustness to proximity trap on ProxBench. The detailed derivation process is provided in Appendix C.

4.1 MAIN RESULTS

First, we conduct experiments on LongBench-v2 (Bai et al., 2025), examining whether our LYRA improves long-context understanding across different input lengths. As exhibited in Table 1, LYRA achieves the best overall performance and consistently ranks among the strongest methods across all context-length splits. The leading performance on the Short split indicates that the improvement does not come at the cost of local context modeling, while the clear advantage on the Long split demonstrates a stronger ability to preserve and retrieve relevant evidence over extended contexts. This consistent behavior can be attributed to the combination of directional QK matching and the t-distributed transformation. Directional matching distinguishes semantically relevant evidence from weakly aligned background context, while the t-distributed transformation prevents relevant evidence from being excessively suppressed when its original QK score is weakened by distance. As a result, LYRA reduces the competition from proximal background tokens without indiscriminately increasing the scores of all distant tokens. The results therefore demonstrate that LYRA provides a robust balance between local context modeling and distant evidence utilization, supporting its effectiveness in mitigating the Proximity Trap.

Table 1: **Evaluation results of different homogeneous methods on LongBench-v2 (Bai et al., 2025),** following the evaluation setting of ProxyAttn (Wang et al., 2026b). We compare our method against the baseline, MInference (Jiang et al., 2024), FlexPrefill (Lai et al., 2025), XAttention (Xu et al., 2025), ProxyAttn (Wang et al., 2026b), S2O (Zhang et al., 2026) and PBS-Attn (Wang et al., 2026a). We report accuracy on the full evaluation set and across three context-length splits defined by word count: Short (<32K), Medium (32K–128K), and Long (>128K). The overall score is computed over the complete evaluation set. The best and second-best results in each column are highlighted in red and blue, respectively.

Methods	Venue	Short	Medium	Long	Overall
Baseline		36.67	29.30	30.56	32.21
Minference	NeurIPS'24	40.60	26.50	28.70	32.02
FlexPrefill	ICLR'25	41.70	30.20	27.80	33.80
Xattention	ICML'25	41.70	27.40	32.40	33.59
ProxyAttn	ICLR'26	43.90	28.40	27.80	33.82
S2O	ACL'26	-	-	-	32.60
PBS-Attn	ICML'26	-	-	-	34.39
Ours		**47.22**	**29.63**	**33.33**	**36.72**

Moreover, we evaluate LYRA on RULER (Hsieh et al., 2024) to provide a more fine-grained analysis of robustness to context length. Unlike LongBench-v2, which groups examples into broad context-length splits, RULER evaluates the same set of controlled tasks at explicitly specified input lengths, allowing the performance variation of each method to be examined as the context expands.

As illustrated in Table 2, LYRA achieves the best performance at every evaluated context length and establishes the strongest average result. The consistent advantage from 8K to 128K indicates

Table 2: **Evaluation results of different homogeneous methods on RULER (Hsieh et al., 2024)**, following the evaluation setting of S2O (Zhang et al., 2026). We compare our method against FlexPrefill (Lai et al., 2025), XAttention (Xu et al., 2025), ProxyAttn (Wang et al., 2026b), S2O (Zhang et al., 2026), and PBS-Attn (Wang et al., 2026a). We report accuracy at five explicitly controlled context lengths, namely 8K, 16K, 32K, 64K, and 128K, providing a more fine-grained evaluation of length robustness. The average score is computed across these five context lengths. The best and second-best results in each column are highlighted in red and blue, respectively.

	Venue	8K	16K	32K	64K	128K	Avg.
FlexPrefill	ICLR'25	71.65	73.89	75.38	72.65	68.51	72.42
Xattention	ICML'25	85.63	82.25	81.60	73.18	69.91	78.51
ProxyAttn	ICLR'26	92.07	89.75	86.68	83.57	77.09	85.83
S2O	ACL'26	85.80	82.73	80.34	73.95	69.97	78.56
PBS-Attn	ICML'26	85.56	79.34	80.95	70.70	67.90	76.89
Ours		**96.33**	**94.19**	**93.14**	**85.39**	**77.56**	**89.32**

that the effectiveness of LYRA is not restricted to a particular context range. In particular, LYRA delivers substantial improvements at intermediate and long context lengths while retaining the leading performance at 128K, where relevant evidence faces increasingly strong competition from background tokens. These results suggest that directional QK matching and the t-distributed transformation remain effective as the amount of competing context increases. Together with the results on LongBench-v2, the consistent performance on RULER demonstrates that LYRA improves both practical long-context understanding and fine-grained length robustness.

4.2 Generalization Across Diverse Tasks

Subsequently, we evaluate LYRA on LongBench (Bai et al., 2024b) to examine whether the improvements generalize across tasks with different evidence structures and reasoning requirements. In contrast to the preceding experiments, which focus on robustness across context lengths, this evaluation covers six task categories and therefore provides a broader assessment of task-level generalization.

Table 3: **Evaluation results of different homogeneous methods on LongBench (Bai et al., 2024b).** We compare our method against MInference (Jiang et al., 2024), FlexPrefill (Lai et al., 2025), XAttention (Xu et al., 2025), PBS-Attn (Wang et al., 2026a), Stem (Niu et al., 2026), and Kascade (Deshmukh et al., 2026). We report the official LongBench score across six task categories: single-document question answering (SQA), multi-document question answering (MQA), summarization, few-shot learning, code completion, and synthetic tasks. The best and second-best results in each column are highlighted in red and blue, respectively.

Method	Venue	SQA	MQA	Summ.	Few-shot	Code	Synthetic	Avg.
MInference	NeurIPS'24	46.92	37.11	16.53	34.15	2.98	66.17	33.98
FlexPrefill	ICLR'25	46.08	37.29	16.53	33.63	2.52	49.00	30.84
XAttention	ICML'25	45.66	37.77	16.66	36.20	2.02	64.67	33.83
PBS-Attn	ICML'26	47.04	37.17	16.66	33.88	2.80	66.33	33.98
STEM	ICML'26	-	14.97	20.21	61.84	19.43	62.19	35.73
Kascade	Arxiv'26	44.87	42.34	23.74	61.99	62.71	34.50	45.03
Ours		**46.01**	**43.35**	**24.73**	**62.17**	**58.41**	**65.67**	**50.06**

As presented in Table 3, LYRA achieves the best average performance, with leading results on multi-document question answering, summarization, and few-shot learning, as well as competitive performance on the remaining tasks. The improvements are particularly clear for tasks that require relevant information to be identified and integrated from multiple locations. In multi-document question answering and summarization, evidence is often distributed across long contexts and competes with substantial background information, making these tasks especially susceptible to the

Proximity Trap. Similarly, few-shot learning requires the model to identify useful demonstrations despite interference from other examples, while code completion often depends on distant definitions and dependencies surrounded by locally similar code. The strong performance across these tasks suggests that directional QK matching effectively distinguishes relevant evidence from weakly aligned background context, while the t-distributed transformation prevents useful information from being excessively suppressed by positional distance. Meanwhile, the competitive results on single-document question answering and synthetic tasks indicate that this mechanism does not compromise tasks dominated by more localized or explicit evidence. Overall, these results demonstrate that LYRA mitigates evidence competition across diverse task structures rather than benefiting only a specific context length or retrieval pattern.

4.3 Alleviating the Proximity Trap on ProxBench

Beyond standard long-context and generalization benchmarks, we evaluate LYRA on our proposed ProxBench to directly examine robustness to the Proximity Trap. The benchmark fixes the query and distant evidence while progressively increasing the semantic similarity and relational ambiguity of proximal background context. We compare LYRA with representative LLMs under the same evaluation protocol across four difficulty levels.

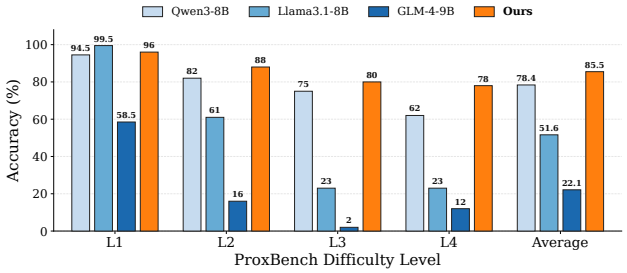


Figure 4: **Evaluation results on ProxBench across four progressively challenging proximal perturbation levels.** We compare Qwen3-8B (Yang et al., 2025), Llama3.1-8B (Grattafiori et al., 2024), GLM-4-9B (Glm et al., 2024), and our LYRA model (8B) on Levels 1–4 and report accuracy (%) at each level. The average score is computed over all four difficulty levels.

As shown in Figure 4, LYRA achieves the best average performance and exhibits substantially greater robustness as the perturbation difficulty increases. Although Llama performs slightly better at the easiest level, LYRA consistently achieves the strongest results at all subsequent levels, with the advantage becoming particularly clear under the most fine-grained perturbations. In contrast, the performance of the baseline models declines sharply when the proximal background shares entities, relations, attributes, or value formats with the target evidence. The smaller performance degradation of LYRA indicates that the model does not rely primarily on positional proximity or surface-level similarity when identifying relevant information. Instead, directional QK matching captures fine-grained relevance between the query and evidence, while the t-distributed transformation prevents distant evidence from being overwhelmed by accumulated competition from proximal background tokens. These results provide direct evidence that LYRA mitigates the Proximity Trap by preserving task-relevant evidence even when nearby context is highly similar but logically irrelevant.

5 Conclusion

In this work, we identify the Proximity Trap, showing that distant evidence can be underutilized not only because of positional distance, but also due to cumulative competition from task-irrelevant proximal background. To address this issue, we introduce LYRA, a t-distributed directional matching mechanism that reshapes the context retrieval distribution to better preserve task-relevant evidence under background competition. Extensive experiments on LongBench-v2, RULER, LongBench, and ProxBench demonstrate consistent improvements across context lengths, task categories, and increasingly challenging proximal perturbations. Overall, our findings suggest that effective long-context modeling should consider not only how far relevant evidence lies from the query, but also what it must compete with along the way.

AI USE STATEMENT

In this work, we used generative AI tools solely to edit and polish the manuscript for improved readability, grammar, and clarity of presentation. We did not use generative AI tools for research ideation, hypothesis formulation, methodology or experimental design, theoretical derivations, implementation of the proposed method, or interpretation of experimental results. All AI-assisted edits were carefully reviewed and revised by the authors to ensure technical correctness and consistency with the original scientific content. We take full responsibility for the final content of this work, including all text, claims, and results.

REPRODUCIBILITY STATEMENT

We release the implementation of LYRA and the corresponding evaluation scripts through our public GitHub repository, and make ProxBench publicly available on Hugging Face. The main paper specifies the model, training, and evaluation protocols used in our experiments, while the appendix provides additional implementation details, hyperparameters, data construction procedures, and theoretical derivations. Together with the released code and benchmark data, these materials are intended to support reproduction of both the main results and the controlled analyses of the Proximity Trap.

REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. Longalign: A recipe for long context alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1376–1395, 2024a.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3119–3137, 2024b.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, 2025.

Dhruv Deshmukh, Saurabh Goyal, Nipun Kwatra, and Ramachandran Ramjee. Kascade: A practical sparse attention method for long-context llm inference, 2026. URL https://arxiv.org/abs/2512.16391.

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.

Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, Yibo Liu, Qianguo Sun, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaxing Zhang. Never lost in the middle: Mastering long-context question answering with position-agnostic decompositional training. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13628–13642, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.736. URL https://aclanthology.org/2024.acl-long.736/.

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.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H Abdi, Dongsheng Li, Chin-Yew Lin, et al. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. *Advances in Neural Information Processing Systems*, 37:52481–52515, 2024.

Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, and Xun Zhou. Flexprefill: A context-aware sparse attention mechanism for efficient long-sequence inference. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 963–989, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/03645743ea35690f30d795d6bac149a5-Paper-Conference.pdf.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Timur Mudarisov, Mikhail Burtsev, Tatiana Petrova, and Radu State. Limitations of normalization in attention. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=16kX08MCav.

Lin Niu, Xin Luo, Linchuan Xie, Yifu Sun, Guanghua Yu, Jianchen Zhu, and S Kevin Zhou. Stem: Rethinking causal information flow in sparse attention. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=8WrMPCd7Kr.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 31932–31951, 2024.

Konrad Staniszewski, Szymon Tworkowski, Sebastian Jaszczur, Yu Zhao, Henryk Michalewski, Łukasz Kuciński, and Piotr Miłoś. Structured packing in llm training improves long context utilization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25201–25209, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Xinghao Wang, Pengyu Wang, Dong Zhang, Chenkun Tan, Shaojun Zhou, Zhaoxiang Liu, Shiguo Lian, Fangxu Liu, Kai Song, and Xipeng Qiu. Sparser block-sparse attention via token permutation. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=5u3Ra6Qf5C.

Yixuan Wang, Huang He, Siqi Bao, hua wu, Haifeng Wang, Qingfu Zhu, and Wanxiang Che. Proxyattn: Guided sparse attention via representative heads. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 18603–18617, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/1fa81061d6d4d7fea88f803d89ae9d6e-Paper-Conference.pdf.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, volume 2025, pp. 62143–62156, 2025a.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67756–67781. PMLR, 13–19 Jul 2025b. URL https://proceedings.mlr.press/v267/wu25ad.html.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Ruyi Xu, Guangxuan Xiao, Haofeng Huang, Junxian Guo, and Song Han. XAttention: Block sparse attention with antidiagonal scoring. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69819–69831. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/xu25ag.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yu Zhang, Songwei Liu, Chenqian Yan, Beichen Ning, Fangmin Chen, Xing Wang, et al. S2o: Early stopping for sparse attention via online permutation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7737–7751, 2026.

Hongbo Zhao, Huibin Wang, Bin Tang, Xianming Hu, Yihong Huang, Yijun Shen, Nuoyi Chen, Ping Li, and Kai Zhang. Adaptive zooming via relevance-informed positional resource allocation for training-free llm context extension. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21067–21083, 2026.

Chuanyang Zheng, Yihang Gao, Guoxuan Chen, Han Shi, Jing Xiong, Xiaozhe Ren, Chao Huang, Zhenguo Li, and Yu Li. Self-adjust softmax. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7827–7847, 2025.

APPENDIX

A RELATED WORKS

Our work is primarily related to two lines of research. The first develops long-context language models through context-window extension, long-context adaptation, and efficient attention computation. The second investigates positional biases and attention normalization, examining how the location of relevant information and competition among contextual tokens affect its influence on model predictions. We review these directions below and distinguish them from our focus on the competition between distant evidence and task-irrelevant proximal background.

A.1 LONG-CONTEXT MODELING

Recent advances in long-context modeling have primarily focused on extending the context window, adapting models to long sequences, and improving the efficiency of long-context inference. One line of work extends the usable context range through positional adaptation. YaRN rescales rotary positional embeddings to support length extrapolation, while LongRoPE employs non-uniform positional interpolation to accommodate substantially longer sequences (Peng et al., 2024; Ding et al., 2024). More recently, RiPRA incorporates semantic relevance into positional resource allocation, assigning finer positional resolution to potentially useful content while compressing less relevant regions (Zhao et al., 2026). These approaches substantially expand the range over which contextual information remains representable, but an extended context window does not necessarily ensure that distant evidence is effectively incorporated into model predictions.

A complementary direction improves long-context capabilities through adaptation and alignment. LongAlign constructs long-form instruction data and develops corresponding training strategies to align language models with long-context tasks (Bai et al., 2024a). Such training-based approaches expose models to longer and more diverse dependencies, improving their ability to operate over extended inputs. Nevertheless, their primary objective is to establish general long-context competence, rather than to characterize how distant evidence is utilized when it competes with abundant, task-irrelevant context near the query.

Another substantial body of research seeks to reduce the computational and memory costs of processing long sequences. StreamingLLM maintains stable streaming inference by retaining attention sinks and a local context window (Xiao et al., 2024). Dynamic sparse-attention methods selectively allocate computation to important regions: MInference exploits recurring sparse attention patterns, FlexPrefill adapts sparse patterns and computational budgets to individual inputs and attention heads, and XAttention estimates block importance through antidiagonal scoring (Jiang et al., 2024; Lai et al., 2025; Xu et al., 2025). ProxyAttn further leverages cross-head similarity to obtain more fine-grained block estimates with reduced overhead (Wang et al., 2026b). Collectively, these methods make long-context processing increasingly scalable by preserving selected attention regions while avoiding exhaustive computation.

Despite this progress, existing studies largely concern whether distant information remains positionally representable, computationally accessible, or retained under efficient inference. Comparatively less attention has been paid to whether a model can effectively use distant evidence after it becomes accessible. In long inputs, distant evidence must compete with abundant proximal background under normalized attention; consequently, successful context extension or sparse selection alone does not guarantee that the evidence receives sufficient influence over the final prediction. Our work addresses this complementary problem by studying how proximal background systematically suppresses the utilization of distant evidence and by improving the allocation of attention under such contextual competition.

A.2 POSITIONAL BIAS AND ATTENTION NORMALIZATION

A growing body of research has shown that the ability of language models to use long contexts is strongly affected by the position of relevant information. Liu et al. (2024) reveal that models often perform well when relevant information appears near the beginning or end of the input, but struggle to use the same information when it is placed in intermediate positions. Subsequent studies improve

positional robustness through long-context training. Position-agnostic decompositional training encourages models to identify and integrate information across different context positions (He et al., 2024), while structured packing organizes semantically related documents into training sequences to improve the utilization of distributed contextual information (Staniszewski et al., 2025). These studies demonstrate that extending the context window alone does not guarantee positionally robust information use and that targeted training can partially alleviate context underutilization.

Complementary work investigates the mechanisms underlying such positional preferences. Retrieval-head analyses identify a small subset of attention heads that plays a disproportionate role in incorporating contextual evidence into factual predictions (Wu et al., 2025a). From a broader architectural perspective, theoretical analyses attribute position bias to the joint effects of causal masking, positional encoding, and information propagation across layers (Wu et al., 2025b). Collectively, these findings suggest that the influence of contextual evidence is shaped not only by its semantic relevance, but also by where it appears and how attention propagates its information. However, most existing analyses characterize the effect of evidence position itself, leaving the role of the surrounding competing context comparatively underexplored.

Another line of research examines the limitations of attention normalization. Standard softmax couples all contextual tokens through a shared normalization denominator, such that increasing the number of competing tokens can weaken the selectivity of attention. Recent work improves its optimization properties through input-dependent adjustment (Zheng et al., 2025), while theoretical analyses show that the ability of normalized attention to separate informative from non-informative tokens deteriorates as the selected set or context grows (Mudarisov et al., 2025). These studies establish the general limitations of softmax-based selection, but do not explicitly consider the asymmetric competition between distant evidence and abundant proximal background.

Our work connects positional bias with attention normalization through the *Proximity Trap*. Rather than treating the underutilization of distant evidence solely as a consequence of its position, we show that it is jointly driven by distance-dependent matching and the cumulative competition induced by nearby, task-irrelevant context. Accordingly, LYRA strengthens attention to task-relevant distant evidence while suppressing interference from weakly relevant proximal background, without indiscriminately favoring all distant tokens.

B ProxBench Details

B.1 Task Formulation

ProxBench evaluates whether a model can retrieve a target fact when the context contains records that partially overlap with the entity, relation, attribute, semantic role, or value format required by the query. Each instance samples a target entity e from a predefined inventory of ten synthetic facility names and generates an access code v in the format `AA-000`. The relevant evidence and query follow the templates

$$s^+ = \text{“The access code assigned to } e \text{ is } v\text{.”}, \tag{12}$$

$$q = \text{“What is the access code assigned to } e\text{?”}. \tag{13}$$

The expected output is the code v without additional explanation.

Each instance contains four proximal background records. These records are constructed to share selected surface or semantic cues with s^+ and q, while remaining logically insufficient for answering the query. In particular, the correct value never appears in the background, and no background record establishes the complete binding between the target entity, the queried access-code attribute, and the candidate value.

B.2 Perturbation Levels

The four levels control which components of the target fact are preserved in the background. Let the target fact be represented as the tuple

$$(e,\ a_{\text{access}},\ r_{\text{assigned}},\ v),$$

where e is the target entity, a_{access} is the access-code attribute, r_{assigned} denotes the assignment relation and recipient role, and v is the correct value.

Level 1: style-matched background. L1 controls for superficial similarity. Its records use the same declarative style and assignment-oriented syntax as the evidence, but change the entity, topic, and answer relation. Example attributes include *maintenance schedule*, *inspection timetable*, *freight allocation*, and *renovation plan*. Because these records contain neither the target entity nor a code-shaped candidate, they can be rejected using relatively coarse lexical and semantic cues.

Level 2: crossed bindings. L2 creates a local binding conflict by placing the target entity, an access-code phrase, another entity, and a candidate code in the same sentence. A typical template is:

> The [target entity] cross-reference lists [candidate code] as the access code assigned to [other entity], not to [target entity].

Although most query-relevant tokens appear together, the semantic recipient of the candidate value is the other entity. Correct prediction therefore requires resolving the argument structure rather than copying the closest code-shaped span.

Level 3: mixed entity–relation interference. L3 alternates between two complementary perturbation types. The first preserves the full access-code relation but replaces the entity:

> The access code assigned to [other entity] is [candidate code].

The second preserves the target entity and assignment relation but changes the queried attribute and value format:

> The emergency contact number assigned to [target entity] is [phone number].

With four background records, each subtype appears twice. Consequently, the background contains both relation-matching candidates and entity-matching candidates, requiring the model to jointly resolve entity and attribute bindings.

Level 4: fine-grained binding characterization. L4 provides a finer diagnostic representation of the mixed binding condition. Each background record is assigned a semantic subtype describing the precise source of overlap, such as preserving the relation while replacing the entity or preserving the entity while replacing the attribute. The annotations additionally record entity, relation, attribute, and semantic-role overlap, as well as whether a candidate shares the expected code or phone-number format. This representation supports subtype-level error analysis and separates failures caused by entity confusion, attribute confusion, role reversal, and value-format attraction.

B.3 Answer-Neutrality and Quality Control

We apply the following automatic constraints during generation:

- **No answer leakage.** The gold access code is prohibited from appearing in any background record.
- **Distinct candidate values.** Distractor codes are different from the gold answer and from other distractor codes in the same instance.
- **Entity separation.** Whenever a template requires another entity, it is sampled from the entity inventory after excluding the target entity.
- **Binding validity.** No background record asserts that a distractor value is the access code assigned to the target entity.
- **Exact background count.** Every instance contains four perturbation records, and all four spans must be recoverable after tokenization.
- **Length control.** Contexts are constructed under the target tokenizer so that the tokenized input plus the reserved generation budget exactly matches the requested sequence length.
- **Span verification.** Evidence, query, and background spans are located after serialization and re-tokenization. Instances that fail round-trip alignment are regenerated using a deterministic retry seed.

- **Reproducibility.** Instance-level seeds are derived from the global seed, context length, perturbation level, and example index using SHA-256. Dataset shards are additionally recorded with byte counts and SHA-256 checksums.

C Implementation and Computational Overhead

Implementation. For Qwen3-8B, we apply LYRA only to the final Transformer block (block 35 under zero-based indexing). We optimize the complete final block while freezing the preceding 35 blocks, token embeddings, final normalization layer, and language-modeling head. This results in approximately 193M trainable parameters, corresponding to 2.36% of the 8.19B-parameter model. The head-specific LYRA parameters themselves introduce only $2H = 64$ additional scalars for $H = 32$ query heads. Training is performed for one epoch in bfloat16 using AdamW, with a maximum sequence length of $16{,}384$. We use a per-device batch size of one on two devices, four gradient-accumulation steps, a learning rate of 2×10^{-5}, a 3% warmup ratio, zero weight decay, and random seed 42.

Computational Overhead. Consider dense attention with batch size B, sequence length n, H query heads, and head dimension d_h. Standard attention requires $\mathcal{O}(BHn^2d_h)$ arithmetic for query–key matching and value aggregation. LYRA replaces the query–key product with a product between normalized queries and keys and additionally performs $\mathcal{O}(BHnd_h)$ normalization and $\mathcal{O}(BHn^2)$ element-wise transformations. Its overall prefill complexity therefore remains $\mathcal{O}(BHn^2d_h)$, identical to dense attention asymptotically. Autoregressive decoding likewise retains $\mathcal{O}(BHnd_h)$ time per generated token, and LYRA does not alter the $\mathcal{O}(BH_{\mathrm{kv}}nd_h)$ KV-cache size.

Table 4: Analytical computation of Qwen3-8B and the additional FLOPs introduced by LYRA. The relative overhead remains below 0.04% across all evaluated context lengths.

Context	Qwen3-8B (TFLOPs)	Additional LYRA (GFLOPs)	Relative Overhead
8K	153.382	17.382	0.0113%
16K	385.929	69.122	0.0179%
32K	1,088.517	275.683	0.0253%
64K	3,443.670	1101	0.0320%
128K	11,953.890	4401	0.0368%

Table 4 reports the analytical computation of Qwen3-8B and the additional FLOPs introduced by LYRA. Across context lengths from 8K to 128K, the relative overhead increases only from 0.0113% to 0.0368% and remains below 0.04% in all settings. This small overhead arises because LYRA modifies only the final Transformer block and adds primarily element-wise normalization and score-transformation operations, whose cost is minor relative to the matrix multiplications in dense attention. Consequently, LYRA preserves the asymptotic computational complexity of the base model while introducing negligible additional arithmetic.

D Ablation Studies

D.1 Ablation Studies on Concentration Parameter κ

Table 5: **Ablation studies on different κ on LongBench-v2 (Bai et al., 2025).** We vary κ, which controls the transformation strength of the t-distributed mapping, while keeping all other settings fixed. Accuracy is reported.

κ	Short	Medium	Long	Overall
2	41.11	27.91	31.48	33.40
4	47.22	29.63	33.33	36.72
8	44.44	28.84	30.63	34.81
16	38.33	24.19	33.33	31.21

We conduct an ablation study to examine the effect of κ, which controls the strength of the t-distributed transformation. We vary κ, and evaluate each configuration across the context-length splits of LongBench-v2. As shown in Table 5, the results exhibit a clear non-monotonic trend, with a moderate value of κ achieving the strongest and most consistent performance across all context lengths. A smaller value provides insufficient separation between relevant evidence and weakly aligned background context, whereas an excessively large value applies overly aggressive score transformation and may suppress moderately aligned keys that still contain useful information. The consistent preference for the same intermediate setting across all splits indicates that κ does not require length-specific tuning. These results confirm that effective mitigation of the Proximity Trap requires a balanced transformation that suppresses proximal background competition while preserving task-relevant evidence.

D.2 Ablation Studies on Different Backbone

Table 6: **Ablation Studies on Different Backbones.** RULER accuracy (%) using Llama-3.1-8B across context lengths from 8K to 128K. Avg. denotes the mean across the five lengths. Red and blue indicate the best and second-best results, respectively.

	Venue	8k	16k	32k	64k	128k	Avg.
Minference	NeurIPS'24	94.22	93.92	88.82	84.64	70.74	86.47
FlexPrefill	ICLR'25	86.89	86.51	81.86	76.35	72.28	80.78
Xattention	ICML'25	91.10	88.62	87.90	84.67	73.15	85.09
ProxyAttn	ICLR'26	94.57	94.01	91.55	86.63	78.25	89.00
S2O	ACL'26	93.03	90.33	87.91	85.29	72.40	85.79
PBS-Attn	ICML'26	91.06	88.46	85.69	78.36	67.01	82.12
Ours		**97.22**	**95.79**	**91.63**	**87.42**	**78.79**	**90.17**

We investigate whether the improvement of Lyra depends on the choice of backbone. The main RULER results in Table 2 use Qwen3-8B; here, we additionally evaluate Llama-3.1-8B to test whether the method remains effective across model families. We compare Lyra with the competing methods on RULER at context lengths ranging from 8K to 128K, and report accuracy at each length together with the average.

As shown in Table 6, Lyra achieves the best result at every evaluated length. Its advantage persists as the context grows, while its performance at shorter lengths remains strong. This consistent pattern suggests that the improvement observed with Qwen3-8B is not specific to that backbone. Rather, reshaping attention according to directional relevance also benefits long-context retrieval when applied to Llama-3.1-8B.

D.3 Ablation Study on Layer Placement

To examine how the placement of Lyra affects long-context performance, we apply it to one Transformer layer at a time in Qwen3-8B, selecting layers 15, 24, 33, and 35 while keeping the remaining layers frozen. These positions span the transition from relatively weak evidence attention to the later layers where distant evidence receives substantially greater attention, as shown in Fig. 2. All variants use the same training and evaluation settings, allowing us to isolate the effect of layer placement.

As shown in Table 7, applying Lyra to later layers generally yields stronger overall performance, with the final layer producing the best result. The improvement at layers 24 and 33 is consistent with the attention analysis: these

Table 7: **Ablation study on the placement of Lyra across Transformer layers.** We replace and fine-tune the attention mechanism in one indicated layer of Qwen3-8B while keeping all other layers frozen. Results are LongBench-v2 accuracy (%) for the Short, Medium, and Long context splits and the overall benchmark. Layer indices start from 0.

Layer	Short	Medium	Long	Overall
15	40.56	24.19	28.70	31.02
24	37.22	27.91	33.33	32.41
33	39.44	28.37	32.41	33.20
35	47.22	29.63	33.33	36.72

layers attend strongly to distant evidence, yet
proximal background continues to compete for attention. Notably, the gains are not uniform across context lengths, so we interpret this experiment as evidence for the importance of layer placement rather than a monotonic relationship between layer depth and performance.