

PHAROS: Navigated Hierarchical Attention with Adaptive Budgets, Cross-Level Guidance, Tidal Momentum, and Uncertainty-Aware Fallback

 Research paper





Abstract

Long-context transformer training remains constrained by the quadratic compute and memory cost of dense scaled dot-product attention. Lighthouse Attention reduces this cost through symmetric hierarchical pooling and gradient-free topk token selection, but its selection process remains limited by uniform per-head budgets, independent pyramid-level scoring, and stateless selection across transformer depth. We present PHAROS, a composable augmentation of Lighthouse-style hierarchical attention that improves long-context token selection while preserving compatibility with efficient attention pipelines. PHAROS introduces four coordinated mechanisms:

Depth-Sounding Budget Allocation, Range Light Cross-Level Guidance, Tidal Momentum Selection, and Fog Gate Confidence Fallback. Together, these mechanisms define an adaptive, guided, persistent, and uncertainty-aware framework for hierarchical attention without requiring changes to the attention kernel or the two-stage training recipe.

Introduction

Scaling language model context from tens of thousands to million-token ranges remains primarily bottlenecked by the quadratic compute and memory cost of dense scaled dotproduct attention [Vaswani et al.(2017), Dao et al.(2022)]. Efficient attention methods reduce this burden by selecting, pooling, routing, or approximating the context processed by each attention head [Yuan et al.(2025), Lu et al.(2024), Zhang et al.(2024)]. Lighthouse Attention introduced a training-time hierarchical alternative in which pooled query, key, and value pyramids are scored, top-k pooled tokens are gathered, and standard FlashAttention is applied to the resulting dense subsequence [Peng, Ghosh, and Quesnelle(2026)].

Despite these gains, Lighthouse-style selection contains three rigidities. First, the same top-k budget is used for every head, even though attention heads differ in entropy and focus width [Voita et al.(2019), Clark et al.(2019)]. Second, pyramid levels are scored independently, even though coarse levels provide useful global priors. Third, selection is recomputed independently across transformer depth, even though important tokens often recur across layers [Geva et al.(2022), Yu et al.(2023)]. PHAROS addresses these limitations while remaining a thin wrapper over efficient attention pipelines.

Background: Hierarchical Attention Selection

Let a sequence have length N and each attention head have hidden dimension d . For head h , denote queries, keys, and values as $Q_h, K_h, V_h \in \mathbb{R}^{N \times d}$. A Lighthouse-style selector constructs an L -level pyramid by average pooling along the sequence dimension:

$$\bar{X}_h^{(\ell)} = \text{AvgPool}_{p\ell}(X_h), \quad X \in \{Q, K, V\}.$$

A scorer computes logits over pooled tokens. A fixed topk operation then selects pooled indices, gathers the corresponding key/value subsequence, and applies FlashAttention to the gathered dense subsequence [Dao et al.(2022), Peng, Ghosh, and Quesnelle(2026)]. This preserves compatibility with efficient attention kernels, but the selection policy remains uniform, independent across levels, and stateless across depth.

PHAROS Method

PHAROS contributes four coordinated mechanisms. Depth-Sounding Budget Allocation. DBA assigns perhead token budgets according to selection entropy while preserving a global budget. Let $p_{h,t} = \text{softmax}(s_{h,t})$ and let $\hat{H}_h$ be the average entropy of head h . DBA computes normalized inverse-entropy weights:

$$W_h = \frac{\exp(-\hat{H}_h/\tau)}{\sum_h \exp(-\hat{H}_h/\tau)}$$

and assigns $k_h = \max(k_{\min}, \text{round}(k_{\text{total}} w_h))$. Low entropy heads can use smaller budgets, while diffuse heads receive more capacity.

Range Light Cross-Level Guidance. RCG injects coarse pyramid priors into finer levels. Coarse logits are upsampled and added as a learned bias:

$$\hat{s}_h^{(\ell)} = s_h^{(\ell)} + \gamma_h^\ell \text{Upsample}(\hat{s}_h^{(\ell+1)})$$

A zero initialization recovers the baseline at the start of training.

Tidal Momentum Selection. TMS maintains a lightweight exponential moving average of selection logits across transformer depth:

$$\hat{s}_h^{(\ell)} = s_h^{(\ell)} + \beta_h m_h^{(\ell-1)}, m_h^{(\ell)} = p m_h^{(\ell-1)} + (1-p) \hat{s}_h^{(\ell)}$$

This propagates selection evidence without adding a recurrent hidden state, differentiating it from heavier recurrent or statespace mechanisms [Bulatov, Kuratov, and Burtsev(2022), Hutchins et al.(2022), Gu and Dao(2023)].

Fog Gate Confidence Fallback. FCF widens the selected beam only when selector confidence is low. With confidence $ch,t = \max_j \text{softmax}(s_{h,t})_j$, fallback is triggered if $ch,t < \theta_h$, and the local budget expands from k_h to $\min(2k_h, P)$. This uses selection confidence as a lightweight uncertainty proxy, related in spirit to uncertainty-aware attention and adaptive computation [Fan et al.(2021), Schuster et al.(2022), Raposo et al.(2024), Ainslie et al.(2023)].





Algorithmic Composition

PHAROS composes the four mechanisms in a fixed order. First, pooled pyramids are constructed for each head. Second, pooled scores are computed. Third, RCG propagates coarse priors to finer levels. Fourth, logits are concatenated and TMS adds the previous-layer tidal prior. Fifth, DBA assigns the current per-head budget. Sixth, FCF adjusts the beam under low confidence. Finally, selected keys and values are gathered and passed into FlashAttention. The method introduces only small per-head scalar parameters for guidance, momentum, and fallback.

Theoretical Properties

DBA is cost-neutral under exact allocation without rounding or lower-bound correction: the sum of per-head budgets equals the global budget. Intuitively, entropy-aware reallocation can improve expected recall when some heads saturate early while others benefit from additional selected tokens. TMS reduces layer-to-layer selection fluctuation under bounded logit variance because the exponential moving average smooths selection evidence across depth. FCF increases computation only for low-confidence positions, so expected cost remains close to the baseline when fog activation is rare.

Evaluation Protocol

This submission reports no unsupported quantitative claims. PHAROS should be evaluated using the same architecture family, tokenizer, dataset mixture, and training recipe as the Lighthouse baseline [Peng, Ghosh, and Quesnelle(2026)]. Recommended baselines include dense SDPA, Lighthouse with the norm scorer, Lighthouse with the dilated scorer, and PHAROS with individual mechanisms ablated. Metrics should include training loss, resume loss, validation perplexity, long-context retrieval accuracy, throughput, memory usage, and wall-clock latency. The ablation study should isolate DBA, RCG, TMS, and FCF while holding all other training choices fixed.



Related Work

Sparse attention methods route attention over subsets of tokens, blocks, or learned patterns [Yuan et al.(2025), Lu et al.(2024), Zhang et al.(2024)]. Adaptive computation methods allocate compute across positions, layers, or submodules [Ainslie et al.(2023), Raposo et al.(2024), Schuster et al.(2022)]. Crosslayer state methods propagate information across sequence or depth [Bulatov, Kuratov, and Burtsev(2022), Hutchins et al.(2022), Gu and Dao(2023)]. PHAROS is complementary to these directions: it targets training-time hierarchical selection, preserves dense-kernel compatibility, and augments selection with adaptive budgets, cross-level guidance, persistence, and fallback.

Limitations

The main limitation is the absence of verified quantitative experiments in the current draft. PHAROS is therefore presented as a method and reproducible evaluation plan, not as an empirically validated improvement. Future work must validate PHAROS across model scales, context lengths, tokenizers, and datasets, and must report variance across runs. The theoretical analysis also relies on assumptions about entropy, recall curvature, and cross-layer selection stability that should be tested empirically.

Conclusion

PHAROS is a four-mechanism augmentation of Lighthousestyle hierarchical attention. DBA adapts budgets across heads, RCG propagates coarse selection priors, TMS maintains crosslayer selection evidence, and FCF widens selection under uncertainty. Together, these mechanisms provide an adaptive, guided, persistent, and uncertainty-aware framework for longcontext hierarchical attention while preserving compatibility with efficient attention pipelines.



Authors:



Saipriyadarshan Cindula,
Zensar Technologies



Nikhil Gajanan Bajaj,
Zensar Technologies

References

- [Ainslie et al.(2023)] Ainslie, J.; et al. 2023. CoLT5: Faster Long-Range Transformers with Conditional Computation. In Proceedings of EMNLP.
- [Bulatov, Kuratov, and Burtsev(2022)] Bulatov, A.; Kuratov, Y.; and Burtsev, M. 2022. Recurrent Memory Transformer. In Advances in Neural Information Processing Systems.
- [Chen et al.(2024)] Chen, T.; et al. 2024. DeepSeek Sparse Attention. Technical report, DeepSeek AI.
- [Clark et al.(2019)] Clark, K.; et al. 2019. What Does BERT Look At? An Analysis of BERT's Attention. In BlackboxNLP.
- [Dao et al.(2022)] Dao, T.; et al. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IOAwareness. In Advances in Neural Information Processing Systems.
- [Elhage et al.(2021)] Elhage, N.; et al. 2021. A Mathematical Framework for Transformer Circuits. Transformer Circuits Thread.
- [Fan et al.(2021)] Fan, A.; et al. 2021. Bayesian Attention Modules. In Advances in Neural Information Processing Systems.
- [Geva et al.(2022)] Geva, M.; et al. 2022. Transformer FeedForward Layers Build Predictions by Promoting Concepts in the Vocabulary Space. In Proceedings of EMNLP.
- [Gu and Dao(2023)] Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv preprint arXiv:2312.00752.
- [Hutchins et al.(2022)] Hutchins, D.; et al. 2022. BlockRecurrent Transformers. In Advances in Neural Information Processing Systems.
- [Lu et al.(2024)] Lu, D.; et al. 2024. MoBA: Mixture of Block Attention for Long-Context LLMs. arXiv preprint.
- [Peng, Ghosh, and Quesnelle(2026)] Peng, B.; Ghosh, S.; and Quesnelle, J. 2026. Long Context Pre-Training with Lighthouse Attention. arXiv preprint.
- [Raposo et al.(2024)] Raposo, D.; et al. 2024. Mixture-ofDepths: Dynamically Allocating Compute in Transformers. arXiv preprint arXiv:2404.02258.
- [Schuster et al.(2022)] Schuster, T.; et al. 2022. Confident Adaptive Language Modeling. In Advances in Neural Information Processing Systems.
- [Vaswani et al.(2017)] Vaswani, A.; et al. 2017. Attention Is All You Need. In Advances in Neural Information Processing Systems.
- [Voita et al.(2019)] Voita, E.; et al. 2019. Analyzing MultiHead Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. In Proceedings of ACL.
- [Yu et al.(2023)] Yu, L.; et al. 2023. Characterising the Role of Pre-Training Data in Large Language Models. arXiv preprint arXiv:2302.03728.
- [Yuan et al.(2025)] Yuan, Z.; et al. 2025. Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention. arXiv preprint.
- [Zhang et al.(2024)] Zhang, Q.; et al. 2024. HISA: Hierarchical Indexing for Sparse Attention. In Proceedings of ICML.



zensar

An  **RPG** Company

At Zensar, we're 'experience-led everything.' We are committed to conceptualizing, designing, engineering, marketing, and managing digital solutions and experiences for over 145 leading enterprises. Using our 3Es of experience, engineering, and engagement, we harness the power of technology, creativity, and insight to deliver impact.

Part of the \$4.8 billion RPG Group, we are headquartered in Pune, India. Our 10,000+ employees work across 30+ locations worldwide, including Milpitas, Seattle, Princeton, Cape Town, London, Zurich, Singapore, and Mexico City.

For more information, please contact: info@zensar.com | www.zensar.com