WHY ATTENTION PATTERNS EXIST: A UNIFYING TEMPORAL PERSPECTIVE ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention patterns play a crucial role in both training and inference of large language models (LLMs). Prior works have identified individual patterns—such as retrieval heads, sink heads, and diagonal traces—but these observations remain fragmented and lack a unifying explanation. To bridge this gap, we provide a unifying framework to explain the existence of diverse attention patterns by analyzing their underlying mathematical formulations with a temporal continuous perspective. Our work can both deepen the understanding of attention behavior and guide inference acceleration approaches. Specifically, this framework characterizes attention patterns as either predictable patterns, characterized by clear regularities, or unpredictable ones that appear random. Our analysis further reveals that the distinction between them can be explained by variations in query self-similarity across the temporal dimension. Focusing on the predictable patterns, we further provide a detailed mathematical analysis of three representative predictable patterns in terms of the joint effect of queries, keys, and Rotary Positional Embeddings. To validate the framework, we apply it to KV cache compression and LLM pruning tasks. In these experiments, a simple metric inspired by our theory consistently improves performance over baseline methods.

1 INTRODUCTION

Attention patterns matter for both LLM training and inference (Xiao et al., 2023; 2024; Jiang et al., 2024; Li et al., 2025; Yang et al., 2025). Prior studies have shown that attention heads exhibit structured and reusable forms, such as streaming heads, retrieval heads, sink heads, and diagonal-like patterns. Understanding why such patterns emerge is critical for a deeper conceptual understanding of the attention mechanism and can directly inform the design of architectures and inference strategies that improve efficiency and robustness, for example, cache compression, long-context streaming, and pruning.

A substantial body of recent research has investigated the architecture of transformer attention mechanisms. Prior analyses typically focus on a single phenomenon, for example, the attention sink at the first token (Gu et al., 2024) or diagonal traces linked to high-frequency components of RoPE (Barbero et al., 2025). Other studies categorize heads by functional roles, such as retrieval and streaming (Xiao et al., 2023; 2024). Despite these advances, it remains unclear what factors determine which attention pattern a head will adopt under the same attention formulation. Our goal is to uncover a unifying underlying mechanism that explains the emergence of these diverse patterns.

To address this gap, we adopt a temporal view of auto-regressive inference and analyze how attention evolves over time. During inference, a transformer LLM generates each token from the previously generated sequence, so the hidden states and attention scores across positions can be regarded as temporal series. We then isolate the source of temporal variation in attention along the time axis. The attention weight from a current position to a past token is computed as the dot product between the current query and the corresponding key after being rotated by Rotary Positional Embedding (RoPE). For each fixed past position, both the key and its RoPE rotation are fixed, whereas the query varies with the current position. Therefore, the evolution of attention is essentially governed by the query. In this interaction, a few embedding channels may dominate the inner product, which determines the shape of the attention pattern. Figure 1 provides an illustration of how changes in queries and dominant embedding channels reshape the attention pattern.

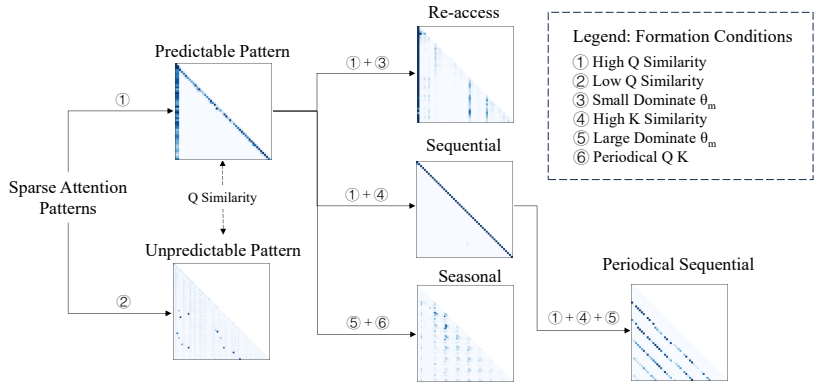


Figure 1: This framework explains the formation of sparse attention patterns from a temporal continuity perspective. We first establish the fundamental Predictable and Unpredictable patterns in Sec. 4. We then detail the conditions that form the Re-access (Sec. 5.1), Sequential (Sec. 5.2), Seasonal (Sec. 5.4), and Periodical Sequential (Sec. 5.3) patterns in their dedicated sections.

Guided by the analysis, we propose a unified framework that interprets attention patterns through the temporal behavior of the queries and the response of the RoPE channels. We view the sequence of query vectors and the associated attention distributions as a time series and characterize them using the notion of continuity. We mathematically show that temporal continuity of queries—measured by their self-similarity—is the key factor distinguishing *predictable patterns*, characterized by clear regularities, and *unpredictable patterns* that appear random. Within the predictable regime, we further provide theoretical conditions for three representative patterns with the joint effect of queries, keys, and RoPE. **Re-access patterns**, where an attention head repeatedly focuses on a small set of tokens, require high query self-similarity and a favorable initial query-key geometry. **Sequential patterns**, which appear as diagonals, are driven by high self-similarity in both queries and keys. In this case, we prove that continuity alone is sufficient to create diagonal-like traces and that special reliance on high-frequency RoPE components is not necessary, which refines and generalizes the conclusions of prior empirical work such as (Barbero et al., 2025). **Seasonal patterns** arise when input periodicity combines with the periodic nature of dominant embedding channels. Since the computing attention from queries, keys, and RoPE is a common design in transformer-based models, our framework both unifies diverse attention patterns and is broadly applicable across LLMs.

To validate our framework, we evaluate it on downstream tasks. Prior works have shown that attention patterns are closely linked to a model's representational capacity (Li et al., 2025; Xiao et al., 2024) and can guide compression. Building on this view, we focus on two complementary compression settings: KV-cache compression for stored states and LLM pruning for model weights. In both cases, a simple metric derived from pattern stability and query similarity consistently outperforms baselines, demonstrating that these principles are practically useful.

In summary, our contributions are as follows: (1) We provide the first systematic analysis of the shapes of attention patterns from a unifying temporal perspective, analyzing random patterns alongside three stable types: re-access, sequential, and seasonal. (2) Theoretically, we demonstrate that stable patterns emerge from the continuity of queries and keys combined with the RoPE mechanism. (3) We identify periodic sequential diagonals and explain them as a consequence of the RoPE rotation period of the dominant channel. (4) We apply our insights to downstream tasks, including KV cache compression and LLM pruning, achieving accuracy improvements.

2 Related Work

2.1 Attention Patterns

The sparse nature of attention mechanisms in Large Language Models (LLMs) is well-documented, giving rise to distinct, recurring patterns. Prior work has largely focused on identifying these patterns and using them for inference optimization. For instance, one widely discussed pattern is the *attention sink*, where high attention scores are consistently assigned to the initial tokens (Xiao et al., 2023),

attracting significant research interest and analysis from various perspectives (Gu et al., 2024; Yu et al., 2024; Cancedda, 2024). Xiao et al. (2023) also highlighted the importance of attention to *recent tokens*, which form a distinct diagonal trace in the attention map. The structured nature of these patterns has been widely exploited for KV cache compression and inference optimization by various methods, such as Minference (Jiang et al., 2024), H2O (Zhang et al., 2024), SnapKV (Li et al., 2024), DuoAttention (Xiao et al., 2024), and KVTuner (Li et al., 2025). Alongside these structured patterns, other works have identified *retrieval heads* (Wu et al., 2024; Xiao et al., 2024). These heads appear to scan the entire context for semantically relevant information, resulting in seemingly random attention maps that are crucial for long-context reasoning and factuality (Xiao et al., 2024). However, these observations have remained largely fragmented, lacking a unifying theory to explain the co-existence and emergence of these diverse patterns.

2.2 The Role of Positional Encoding

A growing body of work has sought a mechanistic explanation for these patterns by examining the role of Rotary Positional Embeddings (RoPE) (Su et al., 2024). Research has shown a direct link between RoPE's frequency components and specific pattern shapes. For instance, high-frequency components in RoPE have been demonstrated to be responsible for the formation of diagonal or previous-token patterns (Barbero et al., 2025). Conversely, other studies suggest that low-frequency components, or specific "outlier" channels with large magnitudes, may contribute to the emergence of attention sinks by creating a rotational offset that favors certain positions (Jonasson, 2025). While these studies provide crucial insights into how positional encoding shapes attention, they often analyze RoPE's effects in isolation, without fully modeling its interaction with the dynamic content of the query and key vectors.

2.3 The Influence of Input Dynamics

A parallel line of research investigates how the properties of the input tokens themselves influence attention patterns. AttentionPredictor (Yang et al., 2025) proposed that the temporal continuity of queries is a key driver for pattern formation, though it did not provide a deep mathematical analysis or consider the interplay with RoPE. Other works have corroborated the importance of input features, suggesting that attention sinks may arise from specific query-key angular relationships that are independent of position (Gu et al., 2024). Similarly, the continuity of queries between layers and constant massive channels of keys has also been noted (Lee et al., 2024; Liu et al., 2024), hinting at the inherent temporal consistency within the model. However, this line of inquiry has yet to be formally connected with the rotational effects of RoPE to provide a complete picture.

In this work, we bridge the gap between these latter two perspectives. We propose a unifying theoretical framework that explains how input dynamics and positional encoding together influence attention patterns. Specifically, we demonstrate that variations in query self-similarity over time, when coupled with the rotational mechanics of RoPE, can mathematically account for the diverse patterns observed in prior works.

3 Background

Attention Mechanism. At the decoding step t, let the query be $q_t \in \mathbb{R}^d$, the key matrix $K = [k_1, \ldots, k_T]^\top \in \mathbb{R}^{T\times d}$ with $k_j \in \mathbb{R}^d$, and the unnormalized logits

$$a_{t,j} = q_t^\top R_{t-j} k_j, \quad a_t \in \mathbb{R}^T, \tag{1}$$

where R_{t-j} is the *Rotary Positional Embedding* (RoPE) operator that rotates vector k_j by a relative phase proportional to $(t-j)$.

The attention distribution is then

$$A_t = \mathrm{softmax}(a_t). \tag{2}$$

Since the softmax function is monotonic with respect to the logits, it preserves their relative order across positions. Therefore, for clarity, we focus our discussion on the logits a_t , and the resulting conclusions directly extend to the final attention distribution A_t.

RoPE. RoPE encodes relative position information by applying channel-wise 2D rotations to pairs of embedding dimensions. For feature pair $(m, m + d//2)$ at position m, the rotation is

$$R_{n,m} = \begin{pmatrix} \cos(n\theta_m) & -\sin(n\theta_m) \\ \sin(n\theta_m) & \cos(n\theta_m) \end{pmatrix}, \tag{3}$$

where $\theta_m = c^{-2m/d}$ is the frequency of the m-th channel, d is the hidden dimension, and c is a hyperparameter. While the original RoPE paper (Su et al., 2024) proposed pairing adjacent dimensions, this half-split pairing scheme is adopted by large-scale models like Llama and Qwen2 for greater computational efficiency.

Thus, for a query q_t and key k_i, the RoPE-augmented attention score on channel m is

$$a_{t,i}^{(m)} = q_t^{(m)\top} R_{t-i,m}\, k_i^{(m)}, \tag{4}$$

where $q_t^{(m)} = (q_{t,2m}, q_{t,2m+1})^\top$.

Decomposition View of Attention. Using the RoPE formulation, the attention logits $a_{t,i}$ between q_t and k_i can be decomposed channel-wise. Let $q_t = \bigoplus_{m=1}^{M} q_t^{(m)}$ and $k_i = \bigoplus_{m=1}^{M} k_i^{(m)}$, where each pair $q_t^{(m)}, k_i^{(m)} \in \mathbb{R}^2$ corresponds to a frequency channel with angular frequency $\theta_m = c^{-2m/d}$. Then

$$a_{t,i} \;=\; \sum_{m=1}^{M} \|q_t^{(m)}\|\, \|k_i^{(m)}\| \cos\Big(\phi_{t,i}^{(m)} + (i-t)\theta_m\Big), \tag{5}$$

where $\phi_{t,i}^{(m)}$ denotes the angle between $q_t^{(m)}$ and $k_i^{(m)}$. This decomposition highlights how each frequency channel contributes additively to the overall attention score, and how temporal shifts $(i - t)$ are modulated by channel-dependent phases θ_m.

RoPE Key Property. RoPE satisfies a relative-position identity:

$$R_m^\top R_n = R_{m-n}, \tag{6}$$

which ensures that attention depends only on the relative distance $(t - i)$, not absolute positions.

Attention Patterns. It is well-established that the attention mechanism is sparse and shows various patterns. In this work, we focus on these sparse attention patterns, especially in Llama-3.1-8B (Dubey et al., 2024) and Qwen-2.5-7B (Yang et al., 2024) with GSM8K (Cobbe et al., 2021) and AIGC (SoftAge-AI, 2024) datasets.

4 Why Predictable and Unpredictable Attention Patterns Exist

Previous works mainly analyze and utilize attention patterns, including retrieval/streaming heads or A-shape/vertical-slash/block-sparse patterns, from the functionality or geometric morphology view. In contrast, we provide a new and unifying time-series analysis perspective to theoretically understand the existence of diverse attention patterns, utilizing the underlying attention mechanisms. We classify attention patterns into two temporal categories: **predictable** and **unpredictable**. Predictable patterns exhibit temporal continuity across decoding steps or the temporal dimension, where the indices of high attention evolve smoothly over time. Unpredictable patterns, in contrast, display irregular jumps with little temporal consistency. This distinction matters because temporal stability enables inference optimization: stable patterns can be anticipated and efficiently compressed in the KV cache, while unpredictable ones resist such treatment.

Empirically, retrieval attention heads exemplify the unpredictable case. Their attention often jumps across the entire context in a seemingly random fashion (Wu et al., 2024; Xiao et al., 2024; Li et al., 2025), which is crucial for retrieving semantically relevant information but undermines predictability. Predictable patterns, by contrast, correspond to heads that consistently attend to locally structured or repeatedly accessed tokens, reflecting stable model behaviors that are exploitable for compression and acceleration.

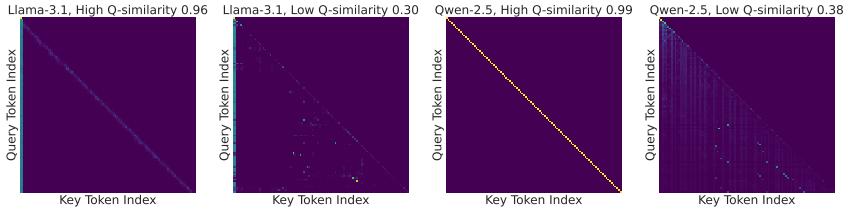


Figure 2: Attention patterns at high and low Query similarity on the Llama and Qwen models. Stable patterns emerge under high similarity, whereas low similarity results in random patterns. There are random bright dots of critical keys in the second and fourth figures.

We argue that the key differentiator behind these two regimes is **query self-similarity**. *When successive queries remain close in representation space, attention indices change smoothly, producing predictable attention maps*. When queries drift strongly, the inequalities that define structured patterns are violated, and even with RoPE's relative rotations, attention jumps unpredictably. To capture this distinction, we introduce a quantitative measure of query continuity, termed *q-similarity*. In Appendix F.3, we further study the distribution of q-similarity across layers, heads, models, and datasets, and show that high-continuity heads are common but not universal. High q-similarity correlates with stable, predictable heads, while low q-similarity leads to retrieval-like, unpredictable behavior. Figure 2 shows the attention patterns of the two models with high and low q-similarity scores. It can be seen that patterns with high q-similarity are more stable, while patterns with low q-similarity are more random.

Proposition 4.1. *Let $q_t, q_{t+1} \in \mathbb{R}^d$ be consecutive queries, $K = [k_1, \ldots, k_T]^\top$ the key matrix, and define the logits*

$$a_{t,j} = q_t^\top R_{t-j} k_j, \qquad a_{t+1,j} = q_{t+1}^\top R_{t+1-j} k_j.$$

If $q_{t+1} - q_t$ has a large norm and is not orthogonal to all rotated keys $\{R_{t+1-j}k_j\}$, then the difference between the logit vectors a_t and a_{t+1} is necessarily large. In particular, there exist constants $c_1, c_2 > 0$ such that

$$\|a_{t+1} - a_t\|_\infty \geq c_1 \|q_{t+1} - q_t\| - c_2.$$

Proposition 4.1 demonstrates that while low q-similarity leads to more random patterns, high q-similarity is a necessary condition for predictable ones. In summary, q-similarity provides a quantitative indicator of whether an attention head behaves in a predictable or unpredictable manner. In the following sections, our theoretical analysis focuses on the predictable heads.

5 Predictable Attention Patterns

In this section, we provide a temporal perspective analysis on predictable attention patterns, which rely on the temporal continuity of queries. The **re-access** pattern occurs when queries are highly self-similar, with low-frequency RoPE components helping to maintain alignment with fixed keys. We also discuss how our analysis relates to the conditions described in prior work (Gu et al., 2024). **Sequential** patterns arise from the combination of high query and key similarity and the relative-position property of RoPE. In some cases, **periodic sequential** patterns appear. We provide a clear calculation for the spacing between adjacent periods and verify it experimentally by varying the location of the dominant RoPE channel and the RoPE base parameter. Finally, we analyze a **seasonal** pattern with periodical queries and keys.

These predictable patterns are useful to LLM inference acceleration. Methods that exploit such temporal regularities (e.g., Minference (Jiang et al., 2024), H2O (Zhang et al., 2024), SnapKV (Li et al., 2024)) can compress the KV cache with little loss in LLM performance, which empirically supports the claim that temporal stability is a important signal for effective KV compression (Jiang et al., 2024; Zhang et al., 2024; Li et al., 2024).

5.1 Re-access Pattern

The re-access pattern describes repeated attention to a small set of key tokens, appearing as vertical lines in the attention map and often referred to as attention sink (Xiao et al., 2023). Prior work has

attributed this phenomenon to query continuity (Yang et al., 2025) or to the small angle between the first key and all queries (Gu et al., 2024), while others observed its correlation with low-frequency RoPE rotations (Jonasson, 2025). However, these explanations are partial.

We propose that the stability of reaccess pattern relies on two factors: (1) high self-similarity of consecutive queries, which prevents attention scores from drifting, and (2) the low-frequency components of RoPE, which preserve alignment between queries and fixed keys even as time t increases.

Theorem 5.1 (Vertical Stability of Attention)**.** *Suppose the channel-wise decomposition (Background, Eq. 5) holds for the attention logits $a_{t,i}$. Assume that the queries evolve continuously in the sense that $\|q_{t+1} - q_t\| \le \varepsilon$, while all keys k_i remain fixed between steps t and $t+1$. Further assume the existence of a dominant low-frequency channel $m^\star$ whose weight $w_{m^\star}$ dominates the other channels, and whose RoPE frequency $\theta_{m^\star}$ is small. Then the per-key differences $a_{t+1,i} - a_{t,i}$ are uniformly small, and the attention logits are vertically stable.*

When queries vary little over time or decoding steps, the only source of temporal change in equation 5 is the RoPE-induced phase $(i-t)\theta_m$. If a dominant channel with small θ_m controls the sum, then shifting $t \mapsto t+1$ changes the cosine term only marginally, hence $a_{t+1,i} \approx a_{t,i}$. This yields vertically aligned attention weights.

Connection to Attention Sink in the First Token. A well-known empirical phenomenon is the *attention sink*, which typically appears at the first token position. Prior work Gu et al. (2024) observed that queries and keys at the initial position tend to have a very small angle, and attributed this alignment as the cause of the sink. Our analysis provides a complementary explanation: from the decomposition in Equation 5, when the angle $\phi_{t,i}^{(m)}$ between $q_t^{(m)}$ and $k_i^{(m)}$ is small, the cosine term $\cos(\phi_{t,i}^{(m)} + (i-t)\theta_m)$ is close to 1. Consequently, the logit contribution from that channel approaches its maximum possible value $\|q_t^{(m)}\|\|k_i^{(m)}\|$, making the overall attention score $a_{t,i}$ large. This alignment effect explains why high attention scores often emerge at positions where q and k are nearly aligned, particularly at the first token.

5.2 Sequential Pattern

Sequential patterns exhibit a shifting focus across tokens, typically progressing step by step along the sequence. The diagonal slash often observed near the main diagonal is commonly attributed to positional heads, which attend to tokens at fixed relative offsets. We argue that the sequential pattern arises from the combined effect of both high q-similarity and k-similarity and the relative-position property of RoPE.

Theorem 5.2 (Sequential Patterns under High Self-similarity)**.** *Under the RoPE relative-position encoding, suppose queries and keys both exhibit high self-similarity, in the sense that*

$$\|q_{t+1} - q_t\| \le \varepsilon, \qquad \|k_{i+1} - k_i\| \le \varepsilon$$

for sufficiently small $\varepsilon > 0$. Then the attention logits satisfy

$$|a_{t+1,i+1} - a_{t,i}| \le C\varepsilon,$$

for some constant $C > 0$. Consequently, the attention logits exhibit approximate shift-invariance along the $(+1,+1)$ diagonal, giving rise to sequential patterns in the attention map.

RoPE encodes relative positions through rotations. When queries and keys vary little across steps, this rotation structure preserves their interactions under a simultaneous shift. As a result, attention scores propagate along the $(+1,+1)$ diagonal, producing sequential (slash-like) patterns.

Empirical Results. High self-similarity in both query and key representations is a sufficient condition for the emergence of Sequential patterns. Figure 3 illustrates the patterns of heads with high query similarity and high key similarity, all of which clearly exhibit diagonal structures.

5.3 Periodicity of Sequential Patterns

Empirically, we sometimes observe multiple parallel diagonal lines in attention maps, with a roughly constant spacing between adjacent lines (*periodic sequential pattern*). We attribute this periodicity to the rotation angle of the dominant RoPE channel.

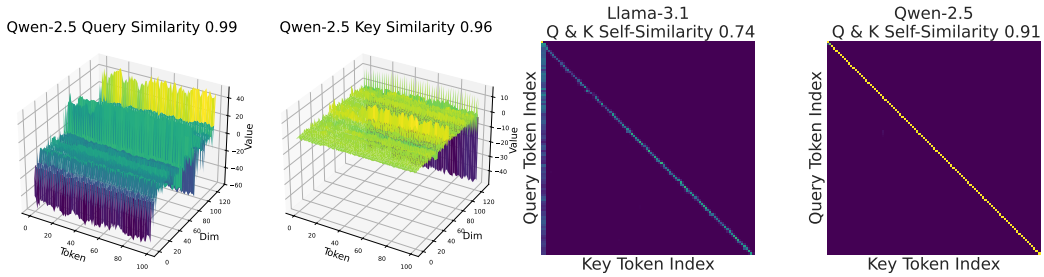


Figure 3: High self-similarity in Query (Q) and Key (K) matrices results in sequential attention patterns. An example from a Qwen-2.5 head (left) with high Q and K self-similarity (0.99 and 0.96) produces a strong diagonal pattern in the attention map (far right). This phenomenon is also observed in Llama-3.1 (center right).

Theorem 5.3 (Periodic Sequential Pattern from a Dominant RoPE Channel)**.** *If a sequential pattern arises and the corresponding key exhibits a massive channel at index $m^\star$, then the spacing between adjacent diagonals is determined by the rotation frequency of that channel:*

$$T = \frac{2\pi}{\theta_{m^\star}} = 2\pi\, c^{2m^\star/d}. \tag{7}$$

Intuition. When the massive channel is located at index $m^\star$, the attention score is dominated by that component:

$$a_{t,j} \approx \|q_t^{(m^\star)}\|\,\|k_j^{(m^\star)}\|\cos\Big(\phi_{t,j}^{(m^\star)} + (j-t)\theta_{m^\star}\Big).$$

This term is a cosine function of the relative offset $(j-t)$ with angular frequency $\theta_{m^\star}$. Consequently, the diagonal lines in the attention map exhibit a regular repetition with period $T = 2\pi/\theta_{m^\star}$, as given in equation 7. Since $\theta_{m^\star} = c^{-2m^\star/d}$, higher channel indices $m^\star$ correspond to lower angular frequencies and therefore to greater spacing between adjacent diagonals.

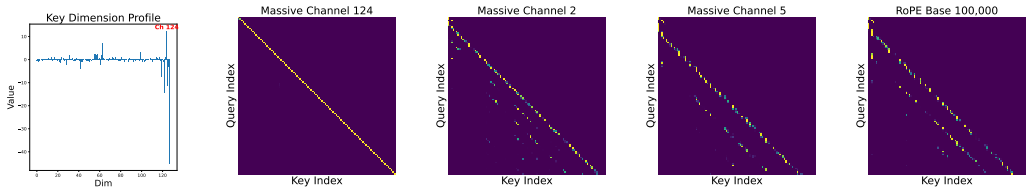
We validate the theoretical mechanism with controlled manipulations on learned key vectors. Our analysis separates two axes of intervention: (i) relocating the massive channel across different indices, and (ii) varying the RoPE base hyperparameter c.

Relocating the massive channel. We first analyze a key vector k_j whose attention map exhibits a single diagonal, as shown in Figure 4 (b). We identify its massive channel at index $m^\star = 124$ as shown in Figure 4 (a). Given the Qwen2.5 RoPE hyperparameters (base $c = 1,000,000$, dimension $d = 128$), this high-index channel corresponds to an extremely low angular frequency. Its theoretical period is $T = 2\pi c^{2m^\star/d} \approx 2.4 \times 10^6$, a value so large that no repetition can be observed within a practical context window.

To demonstrate the relationship between channel frequency and periodicity, we experimentally relocate this massive channel to different target indices m, recomputing the RoPE-augmented attention for each case. The resulting attention maps with $m = 2$ and $m = 3$, visualized in Figure 4 (c) and (d), show that periodic diagonals emerge as the massive channel is moved to lower-index, higher-frequency positions. Specifically, as the channel index m decreases, the angular frequency θ_m increases, shortening the period T and making the diagonals denser. This confirms our first finding: **observable periodic diagonals require the key's massive channel to reside in a high-frequency (low-index) position.**

Furthermore, we observe that even for high-frequency channels, the diagonal patterns fade over long distances. This occurs because the self-similarity between queries and keys naturally diminishes as their relative distance increases, which disrupts the continuity required to sustain the pattern.

Varying the RoPE base c. Independent of the channel index, the choice of RoPE base also controls the periodicity. To isolate this effect, we keep the same dominant channel $m^\star = 5$ and repeat the above procedure for different values of the base (e.g., $c = 1,000,000$ and 100,000 in Figure 4 (d) and (e)). Since the channel frequency is given by $\theta_m = c^{-2m/d}$, decreasing c directly increases θ_m and hence reduces the diagonal period $T = 2\pi/\theta_m$.



(a) (b) (c) (d) (e)

Figure 4: An illustration of how RoPE configuration affects attention patterns. (a) and (b) show a sequential pattern with a dominant channel at $m = 124$. In (c) and (d), we manually change the dominant channel to higher frequencies ($m = 2$ and $m = 5$), which causes periodic diagonals to emerge. In (e), we change the RoPE base from $c = 1,000,000$ to $c = 100,000$ with $m = 5$.

5.4 Seasonal Pattern

Seasonal patterns arise when attention maps repeat with a fixed periodicity. This periodicity can manifest along either the temporal axis or the spatial axis. Due to the periodicity of the hidden states, the periodicity of queries and keys is often aligned, so temporal and spatial repetitions typically occur simultaneously and share the same period. We argue that the underlying cause of the seasonal pattern is that queries and keys exhibit periodicity, which is preserved and sometimes amplified by RoPE through its relative-position encoding. Although the query condition does not exhibit temporal continuity, the pattern remains predictable over time and is therefore a predictable pattern.

Theorem 5.4 (Seasonal Attention Pattern from Periodic Keys and Dominant RoPE Channel)**.** *Suppose the query and key vectors are approximately periodic with interval L, in the sense that*

$$\|q_{t+L} - q_t\| \le \varepsilon_q, \qquad \|k_{i+L} - k_i\| \le \varepsilon_k$$

for sufficiently small $\varepsilon_q, \varepsilon_k > 0$, and that this interval is in near resonance with the dominant RoPE frequency, i.e.,

$$\left|L\,\theta_{m^\star} - 2k\pi\right| \le \delta$$

for some positive integer k and sufficiently small $\delta > 0$. Then the attention logits satisfy

$$|a_{t+L,i} - a_{t,i}| \le C_1(\varepsilon_q + \varepsilon_k) + C_2\delta, \quad |a_{t,i+L} - a_{t,i}| \le C_3(\varepsilon_q + \varepsilon_k) + C_4\delta$$

for some constants $C_1, C_2, C_3, C_4 > 0$, and therefore exhibit a seasonal pattern *with period L along both query and key dimensions.*

The seasonal pattern arises from two combined effects. First, the approximate periodicity of the input queries and keys induces a corresponding periodicity in the attention map. This type of periodicity is common in structured data, such as looking at corresponding elements in consecutive lines of code or data records. Second, when the interval L is in resonance with the dominant RoPE frequency, the relative-position rotations align with the input periodicity, reinforcing the repetition and producing a stronger, more regular seasonal pattern. This dual condition—periodic keys amplified by RoPE resonance—explains the emergence of clean, regularly spaced diagonal slashes in the attention pattern. The observed interval L is therefore determined primarily by the period of the input data itself.

6 Downstream Task

6.1 KV Cache Compression

To demonstrate the practical value of our findings, we apply query similarity to the **KV cache compression** task, which aims to reduce the memory footprint of key-value caches during large language model inference while maintaining model accuracy. Based on our observations, a lower query similarity indicates a higher likelihood of the emergence of retrieval patterns. Since retrieval patterns attend to scattered and unpredictable key positions, they generally require a larger cache budget to preserve critical information (Xiao et al., 2024; Li et al., 2025). Therefore, we leverage query similarity as a proxy signal to dynamically guide the per-layer cache budget allocation under limited memory resources, thereby improving inference efficiency while maintaining model accuracy. We provide the experiment details in Appendix G.1

Table 1: The evaluation results on the LongBench dataset across 512, 1024, and 2048 KV cache budgets. Ours denotes CAKE enhanced with the proposed q-similarity scores.

Budget	Method	Single-DocumentQA			Multi-DocumentQA			Summary			Few-shot Learning			Synthetic		Code		Average↑
		NrtvQA	Qasper	MF-en	HotpotQA	2WikiMQA	Musique	GovReport	QMSum	MultiNews	TREC	TriviaQA	SAMSum	PCount	PRe	Lcc	RB-P	
Llama-3.1-8B																		
Full	**Full**	31.06	45.43	53.78	55.04	47.14	31.29	34.87	25.33	27.49	72.50	91.25	43.81	6.00	99.50	63.36	56.65	49.06
512	**StreamingLLM**	25.64	27.48	33.30	47.36	40.06	24.80	23.16	20.80	22.85	57.50	87.60	42.08	**6.50**	97.00	60.51	51.28	41.75
	H2O	27.76	29.01	44.75	52.78	44.31	29.22	24.71	23.11	24.56	54.50	91.38	42.10	6.36	99.00	62.30	54.33	44.39
	SnapKV	30.76	42.03	52.13	54.15	46.14	30.51	24.98	24.24	24.65	**64.00**	92.05	42.04	6.08	**99.50**	62.62	54.90	46.92
	PyramidKV	30.47	42.15	**52.17**	**54.67**	45.25	30.60	25.00	24.33	24.51	62.50	91.24	41.67	5.95	**99.50**	61.58	53.89	46.59
	CAKE	**31.82**	**42.99**	51.65	54.37	**46.89**	30.73	**26.36**	**24.94**	**25.27**	63.50	91.54	**42.52**	6.33	**99.50**	62.30	54.30	47.19
	Ours	29.47	42.66	51.63	54.53	46.64	**30.81**	25.48	24.57	24.71	62.50	**92.35**	42.42	6.25	**99.50**	**64.56**	**57.35**	**47.21**
1024	**StreamingLLM**	26.64	30.77	35.59	47.31	42.03	24.17	25.81	21.31	25.66	63.50	88.84	42.76	**6.50**	88.00	61.31	53.47	42.73
	H2O	29.57	36.15	45.94	54.43	44.81	29.04	27.64	23.31	26.47	62.00	91.43	**43.14**	6.36	99.00	62.24	55.74	46.11
	SnapKV	**30.95**	44.74	52.58	55.09	46.83	30.37	27.87	24.57	25.99	68.50	92.03	42.60	**6.50**	**99.50**	63.00	56.50	47.95
	PyramidKV	30.54	43.64	**52.73**	55.29	46.29	**31.28**	27.53	24.50	26.00	68.00	**92.09**	41.75	6.05	**99.50**	62.35	55.44	47.69
	CAKE	30.88	**44.95**	52.38	**55.49**	**46.99**	30.82	28.68	**24.91**	26.39	69.00	91.94	42.60	6.00	**99.50**	62.65	56.89	48.13
	Ours	30.77	44.94	52.14	55.43	**46.99**	31.16	**28.72**	24.90	**26.65**	**69.50**	91.95	42.38	6.00	**99.50**	**64.99**	**58.84**	**48.43**
2048	**StreamingLLM**	27.40	36.91	37.85	49.23	44.66	24.31	28.57	21.67	27.12	67.50	90.98	42.49	**6.12**	87.00	63.06	55.32	44.39
	H2O	29.65	39.53	48.64	54.23	46.50	29.28	29.97	23.68	27.21	68.50	91.48	43.06	6.11	**99.50**	63.06	56.91	47.30
	SnapKV	30.99	45.06	53.15	55.25	46.56	30.78	30.24	24.63	27.32	70.50	91.48	42.37	6.00	**99.50**	63.28	56.86	48.32
	PyramidKV	**31.13**	45.06	**53.80**	**55.78**	46.59	30.89	30.25	**24.82**	**27.35**	**71.00**	**91.65**	42.62	6.00	**99.50**	63.27	56.44	48.51
	CAKE	30.79	**45.83**	53.57	55.50	46.60	30.47	**31.12**	24.67	27.16	70.50	91.48	**43.48**	6.00	**99.50**	63.23	56.64	48.43
	Ours	30.70	45.69	53.06	55.49	**46.68**	**30.94**	30.54	24.65	27.12	**71.00**	**91.65**	43.00	6.00	**99.50**	**64.93**	**58.80**	**48.73**
Qwen2.5-7B																		
Full	**Full**	29.05	43.34	52.52	57.59	47.05	30.24	31.78	23.64	23.96	72.50	89.47	45.61	8.50	100.00	59.61	67.12	48.87
512	**StreamingLLM**	19.82	25.40	35.57	43.24	39.18	18.59	25.45	19.07	**22.33**	58.50	71.13	32.29	**8.00**	23.00	46.18	49.01	33.55
	H2O	26.83	34.17	41.43	50.80	41.83	22.82	**25.57**	21.35	22.03	60.50	84.67	**45.86**	**8.00**	95.50	59.11	**64.66**	44.07
	SnapKV	28.94	**40.70**	50.40	**55.80**	44.21	27.83	24.42	22.74	21.07	**66.50**	86.56	44.14	**8.00**	99.50	**59.17**	64.22	45.51
	PyramidKV	27.33	38.04	50.38	55.73	44.28	27.12	22.24	21.86	19.54	66.00	86.36	43.69	**8.00**	99.00	57.59	62.09	45.58
	CAKE	**28.97**	39.46	50.40	54.80	**44.70**	**28.02**	23.90	22.35	20.74	55.00	86.91	44.92	**8.00**	99.50	57.06	64.26	45.56
	Ours	**28.97**	39.40	**50.46**	55.48	44.47	**28.02**	23.99	**22.87**	20.72	55.00	**87.02**	44.66	**8.00**	**100.00**	59.04	64.39	**45.78**
1024	**StreamingLLM**	22.72	29.42	31.47	43.57	38.18	17.99	24.33	19.47	22.46	61.00	87.53	43.79	**8.50**	34.00	55.17	58.43	37.38
	H2O	26.45	34.94	40.49	48.63	42.02	22.27	25.67	20.90	22.41	59.00	87.83	45.07	**8.50**	98.48	59.77	63.88	44.14
	SnapKV	29.24	41.61	50.93	**57.60**	45.50	29.39	25.63	23.06	22.26	65.50	88.92	44.65	**8.50**	**100.00**	58.16	65.30	47.27
	PyramidKV	29.34	38.60	50.17	55.67	45.12	27.82	23.26	22.16	20.55	62.50	86.85	43.26	**8.50**	**100.00**	57.76	61.99	45.85
	CAKE	29.47	42.71	**52.12**	56.11	**46.41**	29.13	**26.86**	**23.12**	22.72	**67.50**	89.23	**45.46**	**8.50**	**100.00**	59.11	64.79	47.70
	Ours	**29.64**	**43.59**	51.53	56.90	45.86	**29.43**	26.64	22.57	**23.00**	65.50	**89.48**	45.24	8.00	**100.00**	**60.04**	**66.06**	**47.72**
2048	**StreamingLLM**	23.18	36.93	45.64	45.30	40.10	19.74	28.49	20.57	23.50	68.00	74.41	33.06	**8.00**	18.00	54.50	53.73	37.07
	H2O	28.55	40.20	47.45	53.49	44.44	27.00	28.93	22.66	**23.79**	63.50	88.50	**46.08**	**8.00**	**100.00**	61.06	67.50	46.95
	SnapKV	29.11	41.53	52.05	57.17	**46.26**	**30.69**	**29.49**	23.23	23.64	**71.50**	89.17	45.49	**8.00**	**100.00**	60.92	67.94	48.12
	PyramidKV	28.39	43.36	51.83	56.75	45.60	30.50	26.90	23.03	23.38	71.00	88.22	45.01	**8.00**	**100.00**	**61.36**	67.37	48.17
	CAKE	29.08	43.35	51.92	57.20	45.77	30.26	29.35	**23.46**	23.59	69.00	89.37	45.37	**8.00**	**100.00**	59.35	67.88	48.31
	Ours	**29.18**	**44.03**	**52.24**	**57.36**	45.77	30.19	29.14	23.31	23.62	69.00	**89.47**	44.99	**8.00**	**100.00**	60.95	**68.06**	**48.46**

Results. As shown in Table 1, our approach consistently outperforms CAKE and the other four baselines across three different budget settings. These results confirm that query similarity effectively reflects the likelihood of the emergence of retrieval patterns, and by allocating more cache budget to layers exhibiting higher query similarity, we are able to preserve critical information more effectively, thereby enabling efficient KV cache compression.

6.2 LLM Pruning

To reduce the parameter size of LLMs and accelerate inference, structured pruning, which removes entire components such as layers, has emerged as a promising approach. Our specific goal is to design more effective proxy metrics to guide whole-layer pruning, so as to achieve higher accuracy under the same compression ratio. Based on our previous analysis, higher query similarity indicates more stable and predictable patterns. Such stability implies that the layer extracts less novel information, making it more dispensable. Consequently, layers with higher query similarity can be pruned with less impact on model performance, while low-similarity layers—which are more likely to host retrieval-like and task-critical behaviors—are preserved. We provide the experiment details in Appendix G.2

Results. As shown in Table 2, our method consistently outperforms ShortGPT across different pruning ratios and models, validating the effectiveness of combining Block Influence with *q-similarity* as a proxy signal for structured layer pruning. These evaluation results on LLM pruning validate our hypothesis regarding the connection between query similarity and stable, predictable patterns. Layers with higher query similarity exhibit greater redundancy due to their stability, and can therefore be pruned with minimal impact on overall model performance.

7 Conclusion

In this work, we introduced a unifying framework to systematically analyze the diverse attention patterns within large language models. We demonstrated that the distinction between predictable

Table 2: Comparison of our proposed method with ShortGPT under the some pruning ratios.

Model	Method	Pruned	Piqa	Hellaswag	Winogrande	Arc Easy	Average (%)↑
Llama-2-7B	ShortGPT	31%	63.33	45.94	61.40	**47.26**	54.48
	∼ with q-similarity (ours)	31%	**63.87**	**50.83**	**63.54**	45.03	**55.82**
	ShortGPT	34%	60.83	42.11	60.38	**44.15**	51.87
	∼ with q-similarity (ours)	34%	60.45	**48.53**	**62.43**	42.55	**53.49**
Llama-3.1-8B	ShortGPT	28%	**66.65**	42.41	58.72	46.25	53.51
	∼ with q-similarity (ours)	28%	64.69	**55.09**	**63.77**	**52.90**	**59.11**
	ShortGPT	31%	64.96	37.69	58.41	42.76	50.96
	∼ with q-similarity (ours)	31%	**65.51**	**42.22**	**62.51**	**46.59**	**54.21**
Qwen-2.5-7B	ShortGPT	39%	**63.17**	**41.83**	50.59	44.32	49.98
	∼ with q-similarity (ours)	39%	62.89	41.80	**51.93**	**45.03**	**50.42**
	ShortGPT	43%	60.83	36.13	47.43	39.77	46.04
	∼ with q-similarity (ours)	43%	**60.88**	**39.87**	**49.72**	**43.94**	**48.60**

and unpredictable patterns can be explained by the temporal self-similarity of queries. Our theoretical analysis further elucidated that stable, predictable patterns arise from the combined effects of query-key continuity and Rotary Positional Embeddings (RoPE), providing a clear explanation for phenomena like periodic sequential diagonals. The practical value of this framework is confirmed by applying its insights to downstream tasks. A simple metric inspired by our theory successfully improved performance in both KV cache compression and LLM pruning, validating our approach.

ETHICS STATEMENT

This research does not involve any personally identifiable information. All datasets used are publicly available and widely adopted in the community, and we have verified that their licenses permit research use. In accordance with the ICLR Code of Ethics, we ensure that our work adheres to principles of fairness, transparency, and responsible AI research. We also disclose that LLMs were used for text polishing, while all conceptual contributions and validation remain the responsibility of the authors in Appendix I.

REPRODUCIBILITY STATEMENT

We will provide open access to all source code, configuration files, and preprocessing scripts, together with detailed instructions to reproduce the main experimental results. All datasets employed are publicly available, and we specify the exact versions and preprocessing steps. Collectively, these resources and specifications enable reliable and faithful reproduction of our results.

REFERENCES

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

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *The Thirteenth International Conference on Learning Representations*, 2025.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019. URL https://arxiv.org/abs/1911.11641.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Wen Xiao. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling, 2025.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4792–4808, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781*, 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. *arXiv preprint arXiv:2407.02490*, 2024.

André Jonasson. Rotary outliers and rotary offset features in large language models. *arXiv preprint arXiv:2503.01832*, 2025.

Wonbeom Lee, Jungi Lee, Junghwan Seo, and Jaewoong Sim. {InfiniGen}: Efficient generative inference of large language models with dynamic {KV} cache management. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pp. 155–172, 2024.

Xing Li, Zeyu Xing, Yiming Li, Linping Qu, Hui-Ling Zhen, Wulong Liu, Yiwu Yao, Sinno Jialin Pan, and Mingxuan Yuan. Kvtuner: Sensitivity-aware layer-wise mixed precision kv cache quantization for efficient and nearly lossless llm inference, 2025. URL https://arxiv.org/abs/2502.04420.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *arXiv preprint arXiv:2404.14469*, 2024.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. *arXiv preprint arXiv:2402.02750*, 2024.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. ShortGPT: Layers in large language models are more redundant than you expect. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20192–20204, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Ziran Qin, Yuchen Cao, Mingbao Lin, Wen Hu, Shixuan Fan, Ke Cheng, Weiyao Lin, and Jianguo Li. CAKE: Cascading and adaptive KV cache eviction with layer preferences. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, Chloe Hillier, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint*, 2019.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019. URL https://arxiv.org/abs/1907.10641.

SoftAge-AI, 2024. URL https://huggingface.co/datasets/SoftAge-AI/multi-turn_dataset.

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

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

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality, 2024. URL https://arxiv.org/abs/2404.15574.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. Duoattention: Efficient long-context llm inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Qingyue Yang, Jie Wang, Xing Li, Zhihai Wang, Chen Chen, Lei Chen, Xianzhi Yu, Wulong Liu, Jianye Hao, Mingxuan Yuan, et al. Attentionpredictor: Temporal pattern matters for efficient llm inference. *arXiv preprint arXiv:2502.04077*, 2025.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. In *International Conference on Machine Learning*, pp. 57659–57677. PMLR, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

A Proof of Unpredictable Pattern

Proposition 4.1. *Let $q_t, q_{t+1} \in \mathbb{R}^d$ be consecutive queries, $K = [k_1, \ldots, k_T]^\top$ the key matrix, and define the logits*

$$a_{t,j} = q_t^\top R_{t-j} k_j, \qquad a_{t+1,j} = q_{t+1}^\top R_{t+1-j} k_j.$$

If $q_{t+1} - q_t$ has a large norm and is not orthogonal to all rotated keys $\{R_{t+1-j}k_j\}$, then the difference between the logit vectors a_t and a_{t+1} is necessarily large. In particular, there exist constants $c_1, c_2 > 0$ such that

$$\|a_{t+1} - a_t\|_\infty \geq c_1 \|q_{t+1} - q_t\| - c_2.$$

Proof. Let $\Delta q := q_{t+1} - q_t$. For each position j, the change in the logit is

$$\Delta a_j = a_{t+1,j} - a_{t,j} = (\Delta q)^\top R_{t+1-j} k_j + q_t^\top \big(R_{t+1-j} - R_{t-j}\big) k_j.$$

Denote the first term by $T_{1,j}$ and the second by $T_{2,j}$.

Step 1: Bounding the RoPE difference term. Since R_m is an orthogonal rotation, its operator norm is 1, so by the triangle inequality, $\|R_{t+1-j} - R_{t-j}\|_{\text{op}} \leq \|R_{t+1-j}\|_{\text{op}} + \| - R_{t-j}\|_{\text{op}} \leq 2$. If we assume the keys are bounded such that $\|k_j\| \leq B_K$ for all j, then

$$|T_{2,j}| \leq \|q_t\| \, \|R_{t+1-j} - R_{t-j}\|_{\text{op}} \, \|k_j\| \leq 2\|q_t\| B_K.$$

Step 2: Lower bounding the query difference term. The first term can be written as

$$|T_{1,j}| = \|\Delta q\| \cdot \big|\langle \tfrac{\Delta q}{\|\Delta q\|}, \, R_{t+1-j} k_j \rangle\big|.$$

The condition that Δq is not orthogonal to all rotated keys implies that the inner product is not always zero. We formalize this by assuming there exists an index j^* and a constant $\alpha > 0$ such that the normalized vectors have a significant projection:

$$\big|\langle \tfrac{\Delta q}{\|\Delta q\|}, \, R_{t+1-j^*} k_{j^*} / \|k_{j^*}\| \rangle\big| \;\geq\; \alpha.$$

This condition essentially states that the direction of the query change aligns with at least one rotated key. Under this condition, and assuming a minimum key norm $\|k_{j^*}\| \geq B_{k,\min}$, we get

$$|T_{1,j^*}| \;\geq\; \alpha B_{k,\min} \|\Delta q\|.$$

Step 3: Combining both terms. Using the bounds for the two terms at index j^*, the reverse triangle inequality gives

$$|\Delta a_{j^*}| \;\geq\; |T_{1,j^*}| - |T_{2,j^*}| \;\geq\; \alpha B_{k,\min} \|\Delta q\| - 2\|q_t\| B_K.$$

Since the infinity norm of a vector is the maximum of the absolute values of its components, we have

$$\|a_{t+1} - a_t\|_\infty = \max_j |\Delta a_j| \geq |\Delta a_{j^*}| \geq \alpha B_{k,\min} \|\Delta q\| - 2\|q_t\| B_K.$$

This establishes the proposition with constants $c_1 = \alpha B_{k,\min}$ and $c_2 = 2\|q_t\| B_K$. This completes the proof. □

B Proof of Re-access Pattern

Theorem 5.1(Vertical Stability of Attention): *Suppose the channel-wise decomposition (Eq. equation 5) holds for the attention logits $a_{t,i}$. Assume that the queries evolve continuously in the sense that $\|q_{t+1} - q_t\| \leq \varepsilon$, while all keys k_i remain fixed between steps t and $t+1$. Further assume the existence of a dominant low-frequency channel $m^\star$ whose weight $w_{m^\star}$ dominates the other channels, and whose RoPE frequency $\theta_{m^\star}$ is small. Then the per-key differences $a_{t+1,i} - a_{t,i}$ are uniformly small, and the attention logits are vertically stable.*

Proof. We derive an explicit uniform bound for the per-key logit difference and show how it depends on the query increment and channel parameters.

Using the channel decomposition from Eq. equation 5, write for each channel m

$$w_m := \|q_t^{(m)}\| \, \|k_i^{(m)}\|, \qquad w'_m := \|q_{t+1}^{(m)}\| \, \|k_i^{(m)}\|,$$

and

$$\psi_m := \phi_{t,i}^{(m)} + (i-t)\theta_m, \qquad \psi'_m := \phi_{t+1,i}^{(m)} + (i-(t+1))\theta_m.$$

Define the logit difference

$$\Delta_{t,i} := a_{t+1,i} - a_{t,i}.$$

Direct subtraction yields the exact identity

$$\Delta_{t,i} = \sum_{m=1}^{M} (w'_m - w_m) \cos \psi'_m \; + \; \sum_{m=1}^{M} w_m \big(\cos \psi'_m - \cos \psi_m \big). \tag{8}$$

We bound the two sums on the right-hand side separately. Let

$$\varepsilon := \|q_{t+1} - q_t\|.$$

First sum. By the triangle inequality and the definition of w_m,

$$\Big| \sum_{m=1}^{M} (w'_m - w_m) \cos \psi'_m \Big| \le \sum_{m=1}^{M} |w'_m - w_m| = \sum_{m=1}^{M} \|k_i^{(m)}\| \, \big| \|q_{t+1}^{(m)}\| - \|q_t^{(m)}\| \big|.$$

Since the Euclidean norm is 1-Lipschitz,

$$\big| \|q_{t+1}^{(m)}\| - \|q_t^{(m)}\| \big| \le \|q_{t+1}^{(m)} - q_t^{(m)}\| \le \|q_{t+1} - q_t\| = \varepsilon.$$

Hence

$$\Big| \sum_{m=1}^{M} (w'_m - w_m) \cos \psi'_m \Big| \le \varepsilon \sum_{m=1}^{M} \|k_i^{(m)}\|. \tag{9}$$

Second sum. Use the inequality $|\cos u - \cos v| \le |u - v|$, so

$$\big| \cos \psi'_m - \cos \psi_m \big| \le |\psi'_m - \psi_m| = \big| \phi_{t+1,i}^{(m)} - \phi_{t,i}^{(m)} - \theta_m \big| \le |\phi_{t+1,i}^{(m)} - \phi_{t,i}^{(m)}| + |\theta_m|.$$

To control the angular difference, let $r_m := \min\{\|q_t^{(m)}\|, \|q_{t+1}^{(m)}\|\}$ and assume $r_m > 0$, and denote $\varepsilon^{(m)} := \|q_{t+1}^{(m)} - q_t^{(m)}\|$. In the 2D RoPE subspace, write $q_t^{(m)} = r_t u_t$ and $q_{t+1}^{(m)} = r_{t+1} u_{t+1}$ with $\|u_t\| = \|u_{t+1}\| = 1$ and let $\Delta\phi^{(m)} := \phi_{t+1,i}^{(m)} - \phi_{t,i}^{(m)}$ be the angle between $q_t^{(m)}$ and $q_{t+1}^{(m)}$. Projecting both vectors onto the circle of radius r_m can only decrease their Euclidean distance while preserving the angle, so by elementary planar geometry we have

$$2 r_m \sin\left(\frac{|\Delta\phi^{(m)}|}{2} \right) \; \le \; \varepsilon^{(m)}.$$

Therefore

$$|\phi_{t+1,i}^{(m)} - \phi_{t,i}^{(m)}| = |\Delta\phi^{(m)}| \le 2 \arcsin\left(\frac{\varepsilon^{(m)}}{2 r_m} \right) \le \frac{\pi}{2} \frac{\varepsilon^{(m)}}{r_m} \le \frac{\pi}{2} \frac{\varepsilon}{r_m},$$

where we used $\varepsilon^{(m)} \le \|q_{t+1} - q_t\| = \varepsilon$ in the last inequality.

Therefore

$$\big| w_m \big(\cos \psi'_m - \cos \psi_m \big) \big| \le w_m \Big(\frac{\pi}{2} \frac{\varepsilon}{r_m} + |\theta_m| \Big).$$

Summing over m yields

$$\Big| \sum_{m=1}^{M} w_m \big(\cos \psi'_m - \cos \psi_m \big) \Big| \le \frac{\pi}{2} \varepsilon \sum_{m=1}^{M} \frac{w_m}{r_m} \; + \; \sum_{m=1}^{M} w_m |\theta_m|. \tag{10}$$

Combine bounds. Inserting equation 9 and equation 10 into equation 8 gives the explicit uniform bound

$$|\Delta_{t,i}| \le \varepsilon \sum_{m=1}^{M} \|k_i^{(m)}\| \; + \; \frac{\pi}{2}\,\varepsilon \sum_{m=1}^{M} \frac{w_m}{r_m} \; + \; \sum_{m=1}^{M} w_m |\theta_m|. \tag{11}$$

Define

$$\delta := \varepsilon \sum_{m=1}^{M} \|k_i^{(m)}\| \; + \; \frac{\pi}{2}\,\varepsilon \sum_{m=1}^{M} \frac{w_m}{r_m} \; + \; \sum_{m=1}^{M} w_m |\theta_m|.$$

Thus $|\Delta_{t,i}| \le \delta$ for every token index i.

Conclusion and asymptotics. Under the theorem hypotheses the keys are bounded and there exists a dominant channel $m^\star$ with $w_{m^\star}$ much larger than the remaining $\{w_m\}_{m \neq m^\star}$, while $r_{m^\star}$ is bounded away from zero and $|\theta_{m^\star}|$ is small. In that regime the two terms proportional to ε in δ vanish as $\varepsilon \to 0$, and the last term is small because the dominant channel's frequency $|\theta_{m^\star}|$ is small and the remaining channels carry only a small total weight. Consequently δ can be made arbitrarily small by taking $\varepsilon \to 0$, $|\theta_{m^\star}| \to 0$, and by increasing the dominance of $w_{m^\star}$ over other channel weights. Therefore the per-key differences $\Delta_{t,i} = a_{t+1,i} - a_{t,i}$ are uniformly small, which proves vertical stability. □

C Proof of Sequential Pattern

Theorem 5.2(Sequential Patterns under High Self-similarity): *Under the RoPE relative-position encoding, suppose queries and keys both exhibit high self-similarity, in the sense that*

$$\|q_{t+1} - q_t\| \le \varepsilon, \qquad \|k_{i+1} - k_i\| \le \varepsilon$$

for sufficiently small $\varepsilon > 0$. Then the attention logits satisfy

$$|a_{t+1,i+1} - a_{t,i}| \le C\varepsilon,$$

for some constant $C > 0$. Consequently, the attention logits exhibit approximate shift-invariance along the $(+1, +1)$ diagonal, giving rise to sequential patterns in the attention map.

Proof. Recall the attention logit

$$a_{t,i} \coloneqq q_t^\top R_{t-i} k_i,$$

where R_Δ is the RoPE rotation for relative offset Δ. By the RoPE identity we have $R_{(t+1)-(i+1)} = R_{t-i}$, hence

$$a_{t+1,i+1} = q_{t+1}^\top R_{t-i} k_{i+1}.$$

Therefore the difference can be written as

$$a_{t+1,i+1} - a_{t,i} = (q_{t+1} - q_t)^\top R_{t-i} k_{i+1} + q_t^\top R_{t-i}(k_{i+1} - k_i).$$

Taking absolute values and applying the Cauchy–Schwarz inequality gives

$$\begin{aligned} \big|a_{t+1,i+1} - a_{t,i}\big| &\le \|q_{t+1} - q_t\|\,\|R_{t-i}k_{i+1}\| \; + \; \|q_t\|\,\|R_{t-i}(k_{i+1} - k_i)\| \\ &= \|q_{t+1} - q_t\|\,\|k_{i+1}\| \; + \; \|q_t\|\,\|k_{i+1} - k_i\|, \end{aligned}$$

where the last equality uses that each R_Δ is orthogonal (rotation), hence $\|R_\Delta v\| = \|v\|$.

Now impose the high self-similarity hypothesis in the rigorous form

$$\|q_{t+1} - q_t\| \le \varepsilon, \qquad \|k_{i+1} - k_i\| \le \varepsilon$$

for some $\varepsilon > 0$. Further assume the query/key vectors are uniformly norm-bounded, i.e. there exist constants $Q, K > 0$ with $\|q_t\| \le Q$ and $\|k_i\| \le K$ for all relevant t, i. Then

$$\big|a_{t+1,i+1} - a_{t,i}\big| \le \varepsilon\,\|k_{i+1}\| + \|q_t\|\,\varepsilon \le \varepsilon(K + Q).$$

Setting $C \coloneqq K + Q$ yields the claimed bound

$$\big|a_{t+1,i+1} - a_{t,i}\big| \le C\,\varepsilon.$$

Thus, under the stated assumptions, the attention logits are approximately shift-invariant along the $(+1, +1)$ diagonal (with error at most $C\varepsilon$), which produces the sequential diagonal structure in the logit map. □

D Proof of Periodic Sequential Pattern

Theorem 5.3 (Periodic Sequential Pattern from a Dominant RoPE Channel): *If a sequential pattern arises and the corresponding key exhibits a massive channel at index $m^\star$, then the spacing between adjacent diagonals is determined by the rotation frequency of that channel:*

$$T \;=\; \frac{2\pi}{\theta_{m^\star}} \;=\; 2\pi\, c^{\,2m^\star/d}. \tag{12}$$

Proof. From the decomposition view of attention, the attention logits can be written as a sum over channels:

$$a_{t,i} = \sum_{m=1}^{M} \|q_t^{(m)}\|\, \|k_i^{(m)}\| \cos\big(\phi_{t,i}^{(m)} + (i-t)\theta_m\big).$$

By assumption, channel $m^\star$ is massive, meaning its contribution to $a_{t,i}$ dominates all other channels:

$$\|q_t^{(m^\star)}\|\, \|k_i^{(m^\star)}\| \gg \|q_t^{(m)}\|\, \|k_i^{(m)}\| \quad \text{for all } m \neq m^\star.$$

Hence, the logits are approximately

$$a_{t,i} \approx \|q_t^{(m^\star)}\|\, \|k_i^{(m^\star)}\| \cos\big(\phi_{t,i}^{(m^\star)} + (i-t)\theta_{m^\star}\big).$$

Consider positions i and $i+T$. Assuming that the magnitudes $\|k_i^{(m^\star)}\|$ and angles $\phi_{t,i}^{(m^\star)}$ vary slowly across consecutive tokens forming the sequential pattern, the attention pattern repeats whenever

$$(i-t)\theta_{m^\star} \;\equiv\; (i+T-t)\theta_{m^\star} \pmod{2\pi},$$

which yields

$$T = \frac{2\pi}{\theta_{m^\star}}.$$

By the definition of RoPE, $\theta_m = c^{-2m/d}$, and substituting $m = m^\star$ gives

$$T = \frac{2\pi}{\theta_{m^\star}} = 2\pi\, c^{2m^\star/d}.$$

Therefore, the interval between adjacent diagonals in the attention map is exactly determined by the rotation frequency of the dominant channel, as claimed. □

E Proof of Seasonal Pattern

Theorem 5.4 (Seasonal Attention Pattern from Periodic Keys and Dominant RoPE Channel): *Suppose the query and key vectors are approximately periodic with interval L, in the sense that*

$$\|q_{t+L} - q_t\| \le \varepsilon_q, \qquad \|k_{i+L} - k_i\| \le \varepsilon_k$$

for sufficiently small $\varepsilon_q, \varepsilon_k > 0$, and that this interval is in near resonance with the dominant RoPE frequency, i.e.,

$$\big|L\,\theta_{m^\star} - 2k\pi\big| \le \delta$$

for some positive integer k and sufficiently small $\delta > 0$. Then the attention logits satisfy

$$|a_{t+L,i} - a_{t,i}| \le C_1(\varepsilon_q + \varepsilon_k) + C_2\delta, \quad |a_{t,i+L} - a_{t,i}| \le C_3(\varepsilon_q + \varepsilon_k) + C_4\delta$$

for some constants $C_1, C_2, C_3, C_4 > 0$, and therefore exhibit a seasonal pattern *with period L along both query and key dimensions.*

Proof. We again use the channel-wise RoPE decomposition. For each channel m, let $R_t^{(m)}$ and $R_i^{(m)}$ denote the 2×2 rotation matrices induced by RoPE at positions t and i with angular frequency θ_m. We define the post-RoPE query and key components as

$$\tilde{q}_t^{(m)} := R_t^{(m)} q_t^{(m)}, \qquad \tilde{k}_i^{(m)} := R_i^{(m)} k_i^{(m)}.$$

By construction, RoPE is an orthogonal transformation, so $\|\tilde{q}_t^{(m)}\| = \|q_t^{(m)}\|$ and $\|\tilde{k}_i^{(m)}\| = \|k_i^{(m)}\|$. The logit contributed by channel m can be written as a dot product

$$a_{t,i}^{(m)} = \langle \tilde{q}_t^{(m)}, \tilde{k}_i^{(m)} \rangle, \quad a_{t,i} = \sum_{m=1}^{M} a_{t,i}^{(m)}.$$

We first bound the variation of the *dominant* channel $m^\star$ along the query dimension. For arbitrary vectors u, u', v, v' we use the standard dot-product inequality

$$\left| u^\top v - u'^\top v' \right| \le \|v\| \, \|u - u'\| + \|u'\| \, \|v - v'\|. \tag{$\star$}$$

Applying ($\star$) with $u = \tilde{q}_{t+L}^{(m^\star)}$, $u' = \tilde{q}_t^{(m^\star)}$ and $v = v' = \tilde{k}_i^{(m^\star)}$ gives

$$\begin{aligned} |a_{t+L,i}^{(m^\star)} - a_{t,i}^{(m^\star)}| &= \left| \langle \tilde{q}_{t+L}^{(m^\star)}, \tilde{k}_i^{(m^\star)} \rangle - \langle \tilde{q}_t^{(m^\star)}, \tilde{k}_i^{(m^\star)} \rangle \right| \\ &\le \|\tilde{k}_i^{(m^\star)}\| \, \|\tilde{q}_{t+L}^{(m^\star)} - \tilde{q}_t^{(m^\star)}\|. \end{aligned} \tag{13}$$

It remains to control $\|\tilde{q}_{t+L}^{(m^\star)} - \tilde{q}_t^{(m^\star)}\|$. Using the definition of $\tilde{q}_t^{(m^\star)}$ we have

$$\begin{aligned} \tilde{q}_{t+L}^{(m^\star)} - \tilde{q}_t^{(m^\star)} &= R_{t+L}^{(m^\star)} q_{t+L}^{(m^\star)} - R_t^{(m^\star)} q_t^{(m^\star)} \\ &= R_{t+L}^{(m^\star)} \big(q_{t+L}^{(m^\star)} - q_t^{(m^\star)} \big) + \big(R_{t+L}^{(m^\star)} - R_t^{(m^\star)} \big) q_t^{(m^\star)}. \end{aligned} \tag{14}$$

Taking norms and using orthogonality of $R_{t+L}^{(m^\star)}$ yields

$$\|\tilde{q}_{t+L}^{(m^\star)} - \tilde{q}_t^{(m^\star)}\| \le \|q_{t+L}^{(m^\star)} - q_t^{(m^\star)}\| + \big\| \big(R_{t+L}^{(m^\star)} - R_t^{(m^\star)} \big) q_t^{(m^\star)} \big\|. \tag{15}$$

The first term is controlled by the assumed L-periodicity of the queries:

$$\|q_{t+L}^{(m^\star)} - q_t^{(m^\star)}\| \le \varepsilon_q.$$

For the second term, we use the near-resonance condition. By definition of RoPE, $R_{t+L}^{(m^\star)} = R_t^{(m^\star)} R_L^{(m^\star)}$, where $R_L^{(m^\star)}$ is a rotation by angle $L\theta_{m^\star}$ in the channel-$m^\star$ plane. The hypothesis $|L\theta_{m^\star} - 2k\pi| \le \delta$ means that $R_L^{(m^\star)}$ is in fact a rotation by angle of magnitude at most δ around the identity. For a planar rotation by angle γ we have $\|R(\gamma) - I\| = 2|\sin(\gamma/2)| \le |\gamma|$, so

$$\begin{aligned} \big\| \big(R_{t+L}^{(m^\star)} - R_t^{(m^\star)} \big) q_t^{(m^\star)} \big\| &= \big\| R_t^{(m^\star)} \big(R_L^{(m^\star)} - I \big) q_t^{(m^\star)} \big\| \\ &\le \|R_L^{(m^\star)} - I\| \, \|q_t^{(m^\star)}\| \le \delta \, \|q_t^{(m^\star)}\|. \end{aligned} \tag{16}$$

Combining equation 15 and equation 16 gives

$$\|\tilde{q}_{t+L}^{(m^\star)} - \tilde{q}_t^{(m^\star)}\| \le \varepsilon_q + \delta \, \|q_t^{(m^\star)}\|.$$

Substituting this into equation 13 and recalling $\|\tilde{k}_i^{(m^\star)}\| = \|k_i^{(m^\star)}\|$ yields

$$|a_{t+L,i}^{(m^\star)} - a_{t,i}^{(m^\star)}| \le \|k_i^{(m^\star)}\| \varepsilon_q + \|k_i^{(m^\star)}\| \|q_t^{(m^\star)}\| \delta =: C_1^{(\star)} \varepsilon_q + C_2^{(\star)} \delta.$$

An entirely symmetric argument, exchanging the roles of t and i and using the L-periodicity of the keys $\|k_{i+L}^{(m^\star)} - k_i^{(m^\star)}\| \le \varepsilon_k$, shows that

$$|a_{t,i+L}^{(m^\star)} - a_{t,i}^{(m^\star)}| \le C_3^{(\star)} \varepsilon_k + C_4^{(\star)} \delta$$

for some constants $C_3^{(\star)}, C_4^{(\star)} > 0$ depending only on the norms of $q_t^{(m^\star)}$ and $k_i^{(m^\star)}$.

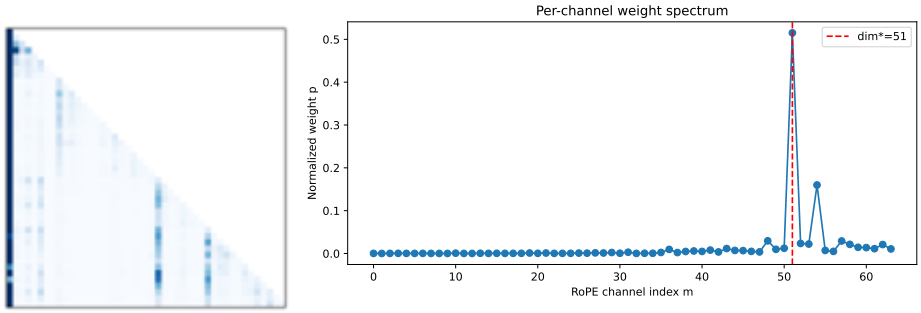
Finally, recall that channel $m^\star$ is assumed to be *massive*: its contribution $\|q_t^{(m^\star)}\| \|k_i^{(m^\star)}\|$ dominates the contributions of all other channels. The residual variation coming from non-dominant channels $\{m \neq m^\star\}$ is therefore uniformly bounded and can be absorbed into the constants $C_1, \ldots, C_4$. Renaming the constants and noting that $\varepsilon_q + \varepsilon_k \ge \varepsilon_q$ and $\varepsilon_q + \varepsilon_k \ge \varepsilon_k$, we obtain the bounds stated in Theorem 5.4:

$$|a_{t+L,i} - a_{t,i}| \le C_1(\varepsilon_q + \varepsilon_k) + C_2\delta, \quad |a_{t,i+L} - a_{t,i}| \le C_3(\varepsilon_q + \varepsilon_k) + C_4\delta.$$

This shows that the dominant component of the attention logits approximately repeats every L steps along both query and key dimensions, giving rise to a seasonal pattern with period L. □

F EMPIRICAL SUPPORT

F.1 EMPIRICAL VALIDATION OF THE DOMINANT-CHANNEL ASSUMPTION OF RE-ACCESS PATTERN



(a) (b)

Figure 5: Empirical validation of the dominant-channel assumption for a re-access head. (a) is an attention heatmap with re-access pattern. (b) plots the RoPE-channel weights of attention at the sink position (dark vertical stripe), showing that a single low-frequency channel m^* accounts for most of the total weight.

Theorem 5.1 assumes that the attention logits of re-access heads are dominated by a single low-frequency channel. To directly examine this assumption, we perform a simple spectrum analysis on a head whose attention map exhibits a clear re-access pattern in Figure 5(a).

For this head, we focus on the key position corresponding to the re-access stripe (the attention "sink"). We decompose the query and key vectors into $M = D/2$ RoPE channels, where each channel m groups the two feature dimensions that share the same RoPE frequency. For every channel m, we aggregate its contribution over the decoding steps and then normalize the resulting values so that they sum to 1. This gives a one-dimensional spectrum $\{p_m\}_{m=0}^{M-1}$.

Figure 5(b) plots the weight of each attention channel. The horizontal axis is the RoPE channel index m ($0 \le m < M$), and the vertical axis is the *normalized channel weight* p_m, i.e., the relative contribution of each channel to the attention logits at the sink position. We observe a highly concentrated pattern: a single channel m^* carries about $p_{m^*} \approx 51\%$ of the total mass, while the remaining channels form a long tail with much smaller weights. The dominant channel m^* lies in the low-frequency half of the RoPE spectrum, consistent with the "dominant low-frequency channel" assumption used in Theorem 5.1.

Together, these observations provide direct empirical evidence that, for the re-access heads we analyse, the attention logits are indeed governed by a single low-frequency channel.

F.2 DISENTANGLING QUERY DYNAMICS AND ROPE IN SEQUENTIAL PATTERN.

To empirically separate the roles of input dynamics and RoPE, we conduct a controlled ablation on a single attention head that exhibits a clear sequential pattern. For this head, the average cosine similarity between consecutive queries is approximately 0.99, and the full model (with RoPE enabled) produces an almost perfectly smooth diagonal attention pattern.

We construct three variants using the same head and the same input sequence (Figure 6):

1. **High q-similarity with RoPE (full model).** In the original model, both the queries and keys have high temporal self-similarity, and RoPE is applied as usual. The resulting attention map shows a clean, nearly translation-invariant diagonal stripe: as t increases, the high-attention region shifts along the $(+1, +1)$ direction with very little distortion. This behavior is consistent with our theoretical analysis, which predicts that when both q_t and k_i vary smoothly in time, RoPE induces approximate shift-invariance along the main diagonal.

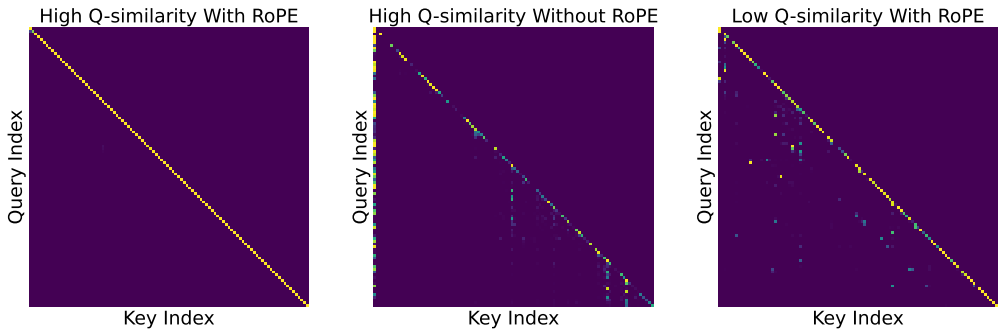


Figure 6: Ablation of query dynamics and RoPE on a head with a strong sequential pattern. **Left**: original head with high q-similarity and RoPE enabled. **Middle**: high q-similarity without RoPE, which retains a rough, broken diagonal with additional vertical streaks. **Right**: RoPE with perturbed q, where the diagonal tendency is overlaid with scattered, unpredictable activation spikes.

2. **High q-similarity without RoPE.** In the second variant, we disable RoPE for this head by replacing the rotation matrices with identity, while keeping the original queries and keys unchanged. The attention map still exhibits a diagonal bias, reflecting the strong local similarity in the queries and keys. However, the diagonal becomes noticeably rough: it is broken into segments and is superposed with vertical streaks. This indicates that high q-similarity alone is sufficient to encourage local, near-diagonal attention, but it does not guarantee the smooth, globally shift-invariant diagonal pattern observed in the full model.

3. **Perturbed q-dynamics with RoPE.** In the third variant, we keep RoPE enabled but mildly perturb the temporal dynamics of the queries by randomly resampling their time indices within the same sequence. This reduces the average cosine similarity between consecutive queries from 0.99 to 0.97, while leaving the keys and RoPE parameters unchanged. The resulting attention map still contains a visible diagonal tendency, but it is now overlaid with many scattered, seemingly random activation spots. In other words, the attention pattern becomes a mixture of a predictable diagonal component and unpredictable spikes.

Across these three conditions, we observe that: (i) high q-similarity without RoPE yields a coarse, locally diagonal pattern, (ii) RoPE with perturbed q-dynamics produces partially diagonal but noticeably more unpredictable attention, and (iii) only when smooth q-dynamics and RoPE are both present do we obtain the clean, stable sequential pattern seen in the full model. This ablation supports our view that sequential attention patterns arise from the *joint* effect of smooth input dynamics and RoPE, and that these two factors play complementary roles: input dynamics control whether the pattern is predictable or unpredictable, while RoPE shapes the predictable component into a regular, shift-invariant diagonal structure.

F.3 Q-SIMILARITY DISTRIBUTION

To better understand the behavior of q-similarity, we compute per-head q-similarity scores across all layers for two models (Llama-3.1 and Qwen-2.5) on two representative datasets (GSM8K and AIGC). As shown in Figure 7, we have following observations:

Overall high q-similarity supporting temporal continuity. Across all heads and layers, the average q-similarity is high for both models (around 0.80 for Llama-3.1 and 0.86 for Qwen-2.5). This empirically supports our working assumption that queries tend to evolve in a temporally continuous manner in a large portion of the network.

Model-specific but layer-structured distributions. Each model exhibits its own characteristic distribution of q-similarity values, indicating that the q-similarity distribution reflects model-specific properties and thus naturally calls for per-model calibration. At the same time, within a given model we observe a clear and consistent structure: heads in the *same* layer have very similar q-similarity scores (forming tight clusters), whereas the average q-similarity differs significantly *across* layers.

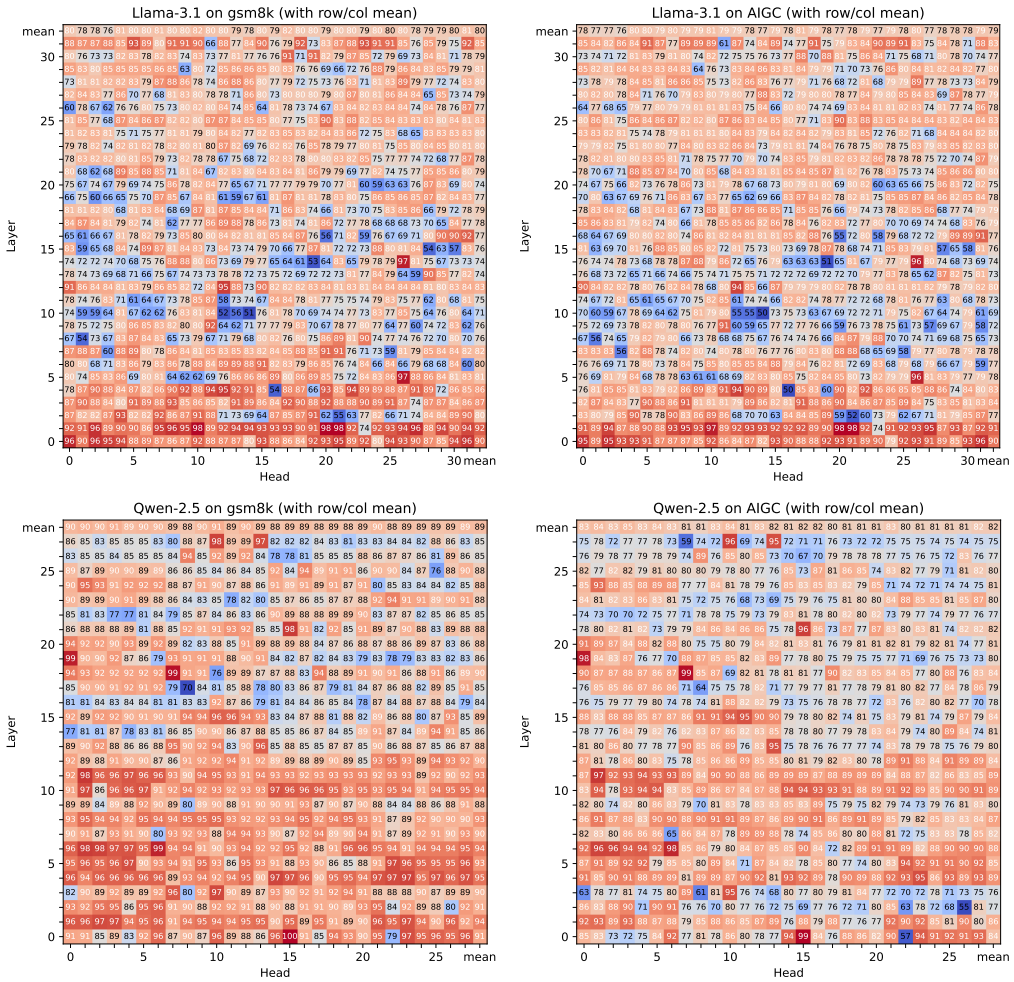


Figure 7: Head-wise q-similarity heatmaps for Llama-3.1 and Qwen-2.5 on GSM8K and AIGC. For readability, we show only the two decimal digits of each q-similarity value (e.g., “83” denotes a q-similarity of 0.83).

This justifies our design choice of operating at the layer level (e.g., using layer-wise averages) when building downstream metrics and policies.

Stable across datasets for the same model, enabling lightweight calibration. For a fixed model, the q-similarity distribution is highly consistent across datasets. For Llama-3.1, the average q-similarity on GSM8K and AIGC differs by only about 0.01. For Qwen-2.5, the absolute mean difference between the two datasets is about 0.07, but the overall shape and ranking of layers/heads are very similar. In particular, the relative ordering of heads is largely preserved, so percentile-based selection strategies (e.g., "top $x\%$ most continuous heads") are unaffected. This indicates that q-similarity has good stability and generalization across datasets, and that only a small amount of data is needed to calibrate q-similarity for a given model, without requiring separate tuning for each task.

G Experiment Details

G.1 Detials for KV Cache Compression

Implementation details. Following CAKE (Qin et al., 2025), we introduce an adjusted per-layer performance score that incorporates query similarity:

$$P_l' = P_l + \alpha(1 - S_l), \tag{17}$$

where P_l denotes the original layer preference score based on entropy and variance of attention patterns (as defined in Equation (6) of CAKE), S_l is the cosine similarity among queries within a recent window, and α is a hyperparameter controlling the contribution of query similarity. Formally,

$$S_l = \text{sim}(Q_{[-Sw:]}), \tag{18}$$

The intuition is that lower query similarity indicates a more random and dispersed attention pattern, which generally requires allocating a larger budget. By adjusting P_l with $(1 - S_l)$, we bias the score toward layers exhibiting retrieval-like behaviors.

Finally, following the allocation rule in CAKE, we normalize the adjusted scores to distribute the total budget across layers:

$$B_l = \frac{P_l'}{\sum_{k=0}^{L-1} P_k'} \cdot B_{\text{total}}. \tag{19}$$

LLMs, benchmark and baselines. We evaluate our method on Llama-3.1-8B (Dubey et al., 2024) and Qwen2.5-7B (Yang et al., 2024), using the **LongBench** (Bai et al., 2024) benchmark, which covers 16 long-context understanding tasks.

G.1.1 Baselines of KV Cache Compression

Baselines include StreamingLLM (Xiao et al., 2023), H2O (Zhang et al., 2024), SnapKV (Li et al., 2024), PyramidKV (Cai et al., 2025), and CAKE (Qin et al., 2025). We provide detailed descriptions of these baselines in Appendix G.1.1. In the KV cache compression task, we evaluate our method against five representative baselines. Based on whether the budget allocation across layers is uniform, these baselines can be categorized into Uniform Allocation, represented by *StreamingLLM* (Xiao et al., 2023), *H2O* (Zhang et al., 2024), and *SnapKV* (Li et al., 2024), and Non-Uniform Allocation, represented by *PyramidKV* (Cai et al., 2025) and *CAKE* (Qin et al., 2025).

- *StreamingLLM*: retains the first and most recent tokens.
- *H2O*: prioritizes tokens with high cumulative attention.
- *SnapKV*: leverages an observation window at the end of the input to cluster and preserve important KV positions for each head.
- *PyramidKV*: allocates larger budgets to lower layers and smaller ones to higher layers with SnapKV's eviction indicator.
- *CAKE*: introduces a preference-prioritized adaptive allocation strategy, dynamically adjusting budgets across layers.

G.2 Detials for LLM Pruning

Implementation details. Building on the Block Influence (BI) metric proposed by ShortGPT (Men et al., 2025), we design an adjusted proxy score:

$$BI' = BI + \beta(1 - q), \tag{20}$$

where β is a hyperparameter and $1 - q$ is an importance score derived from query similarity q.

Following ShortGPT's pruning pipeline, we use the PG19 dataset (Rae et al., 2019) as a calibration set. First, we collect hidden states and queries from each layer while running inference on the calibration data. Next, we compute the proxy scores for all layers based on the adjusted BI score. Finally, we sort the layers in the ascending order of scores and remove those with the lowest scores. The number of pruned layers can be adjusted to balance efficiency gains and accuracy preservation.

LLMs, benchmark, and baselines. We evaluate our method on Llama-2-7B (Touvron et al., 2023), Llama-3.1-8B (Dubey et al., 2024) and Qwen-2.5-7B (Yang et al., 2024). Using the procedure described above, we first evaluate how redundant each layer is and decide which layers are to be pruned. Then we perform zero-shot task classification on common sense reasoning datasets: PIQA (Bisk et al., 2019), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2019) and ARC-easy (Clark et al., 2018) at different pruning ratios. In all our experiments, we compare our method with ShortGPT as a baseline. We list the removed layers in Table 3 of Appendix G.2.1.

G.2.1 List of Removed Layers

In the LLM Pruning downstream task, we evaluated our pruning method on different LLMs and pruning ratios. we list the removed layers in Table 3.

Table 3: Removed layers for different benchmark models, using PG19 as calibration dataset.

Model	Method	Pruning Ratio	Removed Layers
Llama-2-7B	ShortGPT	31%	21, 22, 23, 24, 25, 26, 27, 28, 29, 30
	Ours	31%	19, 21, 22, 23, 24, 25, 26, 27, 28, 29
	ShortGPT	34%	19, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30
	Ours	34%	19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29
Llama-3.1-8B	ShortGPT	28%	20, 22, 23, 24, 25, 26, 27, 28, 29
	Ours	28%	21, 22, 23, 24, 25, 26, 27, 28, 29
	ShortGPT	31%	20, 21, 22, 23, 24, 25, 26, 27, 28, 29
	Ours	31%	19, 21, 22, 23, 24, 25, 26, 27, 28, 29
Qwen-2.5-7B	ShortGPT	39%	4, 5, 6, 7, 8, 10, 11, 12, 14, 15, 20
	Ours	39%	4, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20
	ShortGPT	43%	11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23
	Ours	43%	4, 5, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20

H Comparison with DuoAttention

In this section, we provide a detailed comparison between our proposed method and DuoAttention (Xiao et al., 2024), a recent baseline that explicitly distinguishes retrieval heads and streaming heads for KV cache compression.

H.1 Baselines and Methodology Adaptation

DuoAttention is an optimization-based method that explicitly identifies retrieval heads via training. It assigns a learnable scalar, which we denote as α_{duo}, to each attention head to represent its retrieval importance.

To conduct a direct comparison between our q-similarity metric and DuoAttention's learned importance for the layer-wise budget allocation task, we adapted their scoring mechanism into our

framework. Specifically, we calculate the importance score for each layer l by averaging the α_{duo} values across all heads in that layer. We then compute the allocated budget B_l for layer l using a formulation analogous to Eq. 19:

$$B_l = \frac{\bar{\alpha}_{\text{duo}}^{(l)}}{\sum_{k=0}^{L-1} \bar{\alpha}_{\text{duo}}^{(k)}} \cdot B_{\text{total}}, \tag{21}$$

where $\bar{\alpha}_{\text{duo}}^{(l)}$ is the average score of layer l. This setup allows us to fairly evaluate the effectiveness of the two metrics in identifying layers that require higher KV cache budgets.

H.2 Potential for Higher Compression Ratio

It is crucial to highlight the fundamental difference in how the two methods categorize attention patterns and the resulting impact on the compression scope. DuoAttention operates on a binary premise where it differentiates *Streaming Heads* that necessitate only sink and recent tokens from *Retrieval Heads* requiring full history retention. Consequently, its compression efforts primarily focus on heads exhibiting streaming behavior.

In contrast, our method provides a more detailed categorization. Our q-similarity metric distinguishes complex *Retrieval* patterns from a variety of regular attention patterns, including Re-access, Sequential, and Seasonal patterns. Crucially, our framework identifies these regular patterns as compressible. This effectively expands the scope of compressible heads beyond just streaming heads. By compressing these additional heads that might otherwise be preserved, our method could achieve a higher compression ratio while maintaining model performance.

H.3 Experimental Setup

We conducted experiments to compare the performance of our method against DuoAttention under strict KV cache budget constraints. We evaluated both methods at budget levels of 512 and 1024 tokens.

H.4 Results and Analysis

The quantitative results are presented in Table 4. Our method demonstrates consistent superiority or comparable performance to DuoAttention across different budgets. As shown in Table 4, our method achieves higher average accuracy at both budget levels (64.52% vs. 64.46% at budget 512, and 64.80% vs. 64.68% at budget 1024). Notably, on challenging multi-hop reasoning tasks such as **HotpotQA**, our method significantly outperforms DuoAttention (e.g., 55.45% vs. 54.58% at budget 1024), indicating that our temporal pattern-based approach is more robust in preserving critical information for complex reasoning.

Table 4: Performance comparison with DuoAttention.

Budget	Method	MF-en	HotpotQA	QMSum	TriviaQA	Pre	Lcc	Avg.
-	Full	53.78	55.04	25.33	91.25	99.50	63.36	64.71
512	DuoAttention	**51.70**	54.15	24.50	**92.35**	99.50	64.55	64.46
	Ours	51.63	**54.53**	**24.57**	**92.35**	99.50	**64.56**	**64.52**
1024	DuoAttention	**52.72**	54.58	**24.62**	91.89	99.50	**64.78**	64.68
	Ours	52.63	**55.45**	24.59	**92.04**	99.50	64.58	**64.80**

I The Use of Large Language Models (LLMs)

Large Language Models (LLMs) were employed solely for the purpose of enhancing the linguistic clarity and stylistic refinement of this manuscript.