Demystifying the Slash Pattern in Attention: The Role of RoPE

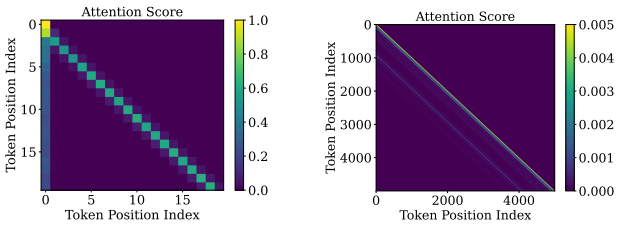
Yuan Cheng [*1] **Fengzhuo Zhang** [†*1] **Yunlong Hou** [*1] **Cunxiao Du** [2] **Chao Du** [2] **Tianyu Pang** [2] **Aixin Sun** [3] **Zhuoran Yang** [4]

Abstract

Large Language Models (LLMs) often exhibit slash attention patterns, where attention scores concentrate along the Δ-th sub-diagonal for some offset Δ. These patterns play a key role in passing information across tokens. But why do they emerge? In this paper, we demystify the emergence of these Slash-Dominant Heads (SDHs) from both empirical and theoretical perspectives. First, by analyzing open-source LLMs, we find that SDHs are intrinsic to models and generalize to out-of-distribution prompts. To explain the intrinsic emergence, we analyze the queries, keys, and Rotary Position Embedding (RoPE), which jointly determine attention scores. Our empirical analysis reveals two characteristic conditions of SDHs: (1) Queries and keys are almost rank-one, and (2) RoPE is dominated by medium- and high-frequency components. Under these conditions, queries and keys are nearly identical across tokens, and interactions between medium- and high-frequency components of RoPE give rise to SDHs. Beyond empirical evidence, we theoretically show that these conditions are sufficient for the emergence of SDHs by formalizing them as modeling assumptions and analyzing the learning dynamics of a shallow Transformer.

1. Introduction

Structured attention patterns are crucial for how large language models (LLMs) organize and transmit information across tokens. Prior work has identified several characteristic attention score patterns, including antidiagonal, block-sparse, vertical and slash patterns (Jiang et al., 2024; Xu et al., 2025). Among these, the *slash pattern*, where the attention score concentrates along the Δ-th sub-diagonal of the attention score matrix for some offset $\Delta \in \mathbb{N}$, is particularly intriguing. We refer to attention heads exhibiting slash patterns as Slash-Dominant Heads (SDHs). SDHs and their slash patterns play important algorithmic roles in LLMs. For example, they enable In-context Learning (ICL) via the induction head circuit (Elhage et al., 2021; Olsson et al., 2022), which is a special case of an SDH with $\Delta = 1$. In addition, another line of work leverages slash patterns to help accelerate long-context inference (Jiang et al., 2024; Xu et al., 2025; Zhao et al., 2025; Lai et al., 2025; Li et al., 2025).



(a) Average attention score matrix of L21H15.

(b) Average attention score matrix of L0H7.

Figure 1. Attention score matrices in Qwen2.5-7B-Instruct with prompts from LongBench. We denote the a-th head at b-th layer as LbHa in this paper. In panel (a), attention concentrates on the sub-diagonals with small offsets 1, respectively. In panel (b), it concentrates on sub-diagonals with large offsets exceeding 500.

More generally, SDHs with diverse values of Δ are prevalent in modern open-source LLMs (see Figure 1). These SDHs enable a token at position i to attend directly to the token at position $i - \Delta$, thereby passing information from earlier tokens to later ones. Their widespread presence and functional importance naturally motivate our central research question:

How do pretrained LLMs implement SDHs using their transformer architectures?

We tackle this question via both thorough empirical studies and rigorous theoretical analysis. In a nutshell, we find that the the emergence of SDHs is intrinsic to the transformer architecture itself, and mainly attributed to the low-rankness of query and key matrices, and Position Embedding (PE),

[*]Equal contribution [†]Project lead [1]National University of Singapore, Singapore. [2]Sea AI Lab, Singapore. [3]Nanyang Technological University, Singapore. [4]Yale University, United States.. Correspondence to: Fengzhuo Zhang <fzzhang@u.nus.edu>, Zhuoran Yang <zhuoran.yang@yale.edu>.

Published as a paper at the 1st FoGen workshop, ICML 2026, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

particularly, Rotary Position Embedding (RoPE) (Su et al., 2024). We show this via thorough empirical studies and rigorous theoretical analysis.

Empirical Studies (Section 3). We focus on open-source LLMs such as Gemma-7B, Qwen2.5-7B-Instruct, and Llama3-8B-Instruct (Gemma Team, 2024; Grattafiori et al., 2024; Qwen Team, 2025), which are decoder-only transformers with RoPE. Our empirical studies consist of three parts, detailed below.

(i) (SDHs are Intrinsic) First, we study *whether SDHs depend on the input prompt or are intrinsic to the model itself.* To this end, we compare the attention scores of the in-distribution prompts to those generated by randomly generated prompts. We observe that the slash dominance pattern persists even in the OOD setting. This implies that SDHs are largely *independent* of the specific input prompt, and they arise mainly due to an *intrinsic algorithmic* mechanism of the transformer model.

Note that attention scores are determined by the interplay of the queries and keys, and the rotational matrices of RoPE. We proceed to examine how each of these components contributes to the emergence of SDHs.

(ii) (Role of Queries and Keys) Perhaps surprisingly, our experiments reveal that the *pre-PE queries and keys are low-rank*, and particularly, *almost rank-one*. As a result, for any token, the pre-PE queries and keys of an SDH point in the same direction, which implies that the semantic contents of queries and keys contribute little to differentiating attention scores. Consequently, the variations in attention scores across tokens has to be *driven almost entirely by RoPE*. Furthermore, we show that these rank-one queries and keys appear because (a) token embeddings lie on a cone and (b) the weight matrices W_K and W_Q project these embeddings to the main axis of the cone.

(iii) (Role of RoPE) As a result of RoPE and rank-one queries and keys, for any $i, j \in \mathbb{N}$, the pre-softmax attention logit from position i attending to position j admits a Fourier-like decomposition. Specifically, we have

$$\texttt{AttnLogit}(i,j) = \sum_{l=1}^{d/2} A_l \cdot \cos\big(\theta_l \cdot (i-j) + \varphi_l\big), \quad (1)$$

where d is the hidden dimension, $\{\theta_l\}_{l=1}^{d/2}$ are the RoPE frequencies. Here $\{A_l\}_{l=1}^{d/2}$ and $\{\varphi_l\}_{l=1}^{d/2}$ are amplitudes and phases, which are almost invariant across tokens thanks to the rank-one structure of the pre-PE queries/keys. Consequently, the slash pattern, peaks of the attention logits in (1) when $i - j = \Delta$, is entirely determined by the frequencies $\{\theta_l\}_{l=1}^{d/2}$. For a more fine-grained understanding, we compare the contributions of individual frequency components in the attention logits. We observe that *high- and medium-frequency* components play a dominant role in forming slash patterns, whereas low frequencies contribute little.

> **Takeaways of Empirical Experiments:** SDHs are an intrinsic architectural effect of the transformer models. They emerge from an almost rank-one structure of the pre-PE queries and keys, which suppresses the variations of semantic information across tokens. Moreover, the high- and medium-frequencies of RoPE interact constructively at a specific offset Δ, producing attention-score peaks at Δ.

We illustrate the mechanism of SDH in Figure 2 and summarize the main takeaways of the empirical studies above.

Theoretical Analysis (Section 4 and Appendix D). To complement the empirical studies, we first propose a concrete sufficient condition on the RoPE frequencies in Section 4, which is validated through empirical experiments. We can then show that, under this condition, a shallow Transformer can learn an SDH from gradient-based training. The details of the theory are deferred to Appendix D.

> **Takeaways of Theoretical Analysis:** When RoPE satisfies a slash-dominance frequency condition and the token embeddings lie on a cone, models trained under these conditions are proven to exhibit SDHs and generalize effectively to OOD input, theoretically demonstrating the *sufficiency* of these conditions for the emergence of SDHs.

2. Preliminaries

In this section, we introduce the causal self-attention (CSA) mechanism with RoPE, which is basic component studied in this paper. We consider decoder-only transformers. Given a prompt P of length $N = N(P)$, let $H = (\mathbf{h}_1, \ldots, \mathbf{h}_N)^\top \in \mathbb{R}^{N \times d}$ denote the hidden states, where d is the hidden dimension. Here H can be either the token embeddings or output of a previous transformer block.

Rotary Position Embedding (RoPE). Assume d is even. For any angle ϕ, define the two-dimensional rotation matrix

$$\rho(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}. \quad (2)$$

Let $\boldsymbol{\vartheta} = (\theta_1, \ldots, \theta_{d/2})$ be the frequency sequence of RoPE. The position $i \in [N]$ is encoded by the block-diagonal rotation matrix

$$R_{\boldsymbol{\vartheta},i} = \operatorname{diag}\big(\rho(i\theta_1), \ldots, \rho(i\theta_{d/2})\big) \in \mathbb{R}^{d \times d}. \quad (3)$$

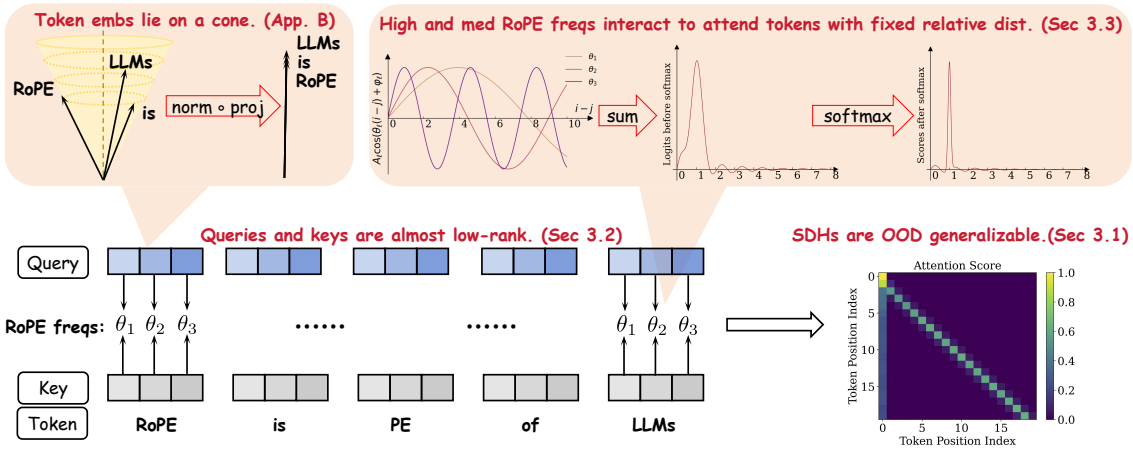


Figure 2. Illustration of the emergence of SDHs. Attention scores are determined by pre-PE queries, keys, and RoPE (left bottom). Because token embeddings lie approximately on a cone, queries/keys are almost rank-one, and nearly identical across tokens (left top), so RoPE primarily governs variation of attention scores across tokens. Then RoPE's high- and medium-frequency components interact constructively at specific lags, producing the attention score peaks at offset Δ (right top). As a result, SDHs emerge and are Out-Of-Distribution (OOD) generalizable (right bottom).

For any $\mathbf{v} \in \mathbb{R}^d$, we define the RoPE operator as

$$\Re_{\boldsymbol{\vartheta},i}(\mathbf{v}) = R_{\boldsymbol{\vartheta},i}\mathbf{v} = \begin{bmatrix} \rho(i\theta_1) & & \\ & \ddots & \\ & & \rho(i\theta_{d/2}) \end{bmatrix} \begin{bmatrix} \mathbf{v}_{1:2} \\ \vdots \\ \mathbf{v}_{d-1:d} \end{bmatrix}.$$

Thus, RoPE rotates the $(2\ell-1)$-th and 2ℓ-th coordinates of $\mathbf{v}$ by angle $i\theta_\ell$ for each $\ell \in [d/2]$. A standard choice is $\theta_\ell = 10000^{-2\ell/d}$ (Su et al., 2024); in this paper, we allow a general frequency sequence $\boldsymbol{\vartheta}$. When there is no ambiguity, we omit the subscripts in $\Re_{\boldsymbol{\vartheta},i}$ and allow it to act row-wise on matrices.

Causal Self-Attention with RoPE. Let $W_Q, W_K, W_V \in \mathbb{R}^{d\times d_1}$ be the query, key, and value matrices. Given hidden states $H \in \mathbb{R}^{N\times d}$, the CSA output is calculated as follows.

(i) CSA first compute queries, keys, and values as $Q = HW_Q = (\mathbf{q}_1, \cdots, \mathbf{q}_N)^\top$, $K = HW_K = (\mathbf{k}_1, \cdots, \mathbf{k}_N)^\top$, and $V = HW_V = (\mathbf{v}_1, \cdots, \mathbf{v}_N)^\top$, respectively[1].

(ii) Then RoPE is applied to queries and keys:

$$\tilde{\mathbf{q}}_i = \Re_{\boldsymbol{\vartheta},i}(\mathbf{q}_i), \qquad \tilde{\mathbf{k}}_i = \Re_{\boldsymbol{\vartheta},i}(\mathbf{k}_i),$$

and we denote the resulting matrices by $\widetilde{Q}$ and $\widetilde{K}$.

(iii) CSA computes attention scores using $\widetilde{Q}$ and $\widetilde{K}$:

$$S = S(P) := \mathrm{softmax}\Big(\mathfrak{M}(\widetilde{Q}\widetilde{K}^\top)\Big) \in \mathbb{R}^{N\times N}, \quad (4)$$

[1]Different from the standard formulation, Qwen2.5 family introduces a special modification where the query vector $\mathbf{q}_i$ (similarly for $\mathbf{k}_i$) includes an additional bias term $\mathbf{b}_Q$, such that $\mathbf{q}_i = W_Q^\top \mathbf{h}_i + \mathbf{b}_Q$.

where $\mathfrak{M}$ is the causal mask, i.e., $\mathfrak{M}(A)_{i,j} = A_{i,j}$ if $i \geq j$ and $-\infty$ otherwise, and the softmax is applied row-wise. By the definition of RoPE, for $i \geq j$, the corresponding pre-softmax logit satisfies $\tilde{\mathbf{q}}_i^\top \tilde{\mathbf{k}}_j = \mathbf{q}_i^\top R_{\boldsymbol{\vartheta},j-i}\mathbf{k}_j$. Finally, the CSA output is

$$\begin{aligned}\mathrm{CSA}\big(H; W_{\{Q,K,V\}}, \boldsymbol{\vartheta}\big) &= S(P)V \\ &= \mathrm{softmax}(\mathfrak{M}(\widetilde{Q}\widetilde{K}^\top))HW_V \in \mathbb{R}^{N\times d}. \quad (5)\end{aligned}$$

By now, we have defined CSA on the hidden states $H \in \mathbb{R}^{N\times d}$ of a given prompt P. In the following sections, the prompt P may be sampled from either the pretraining distribution $\mathcal{D}$ or another OOD distribution $\mathcal{D}'$. For a statistic $f(P)$, we study its expected behavior $\mathbb{E}_{P\sim\bar{\mathcal{D}}}[f(P)]$, where $\bar{\mathcal{D}} \in \{\mathcal{D}, \mathcal{D}'\}$.

3. Empirical Study of Slash-Dominance of Attention

In this section, we empirically study the emergence of slash-dominance of attention in LLMs, which is a special attention pattern observed in various LLMs. As shown in Figure 1, on certain heads, the attention scores have relatively high magnitude on certain sub-diagonals with various values of offsets $\Delta \in \mathbb{N}$. We refer to such a sub-diagonal winning pattern as *slash-dominance*.

To systemically examine the emergence of the slash-dominant pattern, we first define slash-dominance with mathematical rigor, then report the SDHs found on various LLMs empirically. Finally, we investigate why these special heads emerge. In particular, we focus on three models:

Gemma-7B, Llama3-8B-Instruct, and Qwen2.5-7B-Instruct, all of which adopt RoPE as their PE.

Mathematical Definition of Slash-Dominance. Recall that the attention score of an attention head is computed as in (4), which involves the queries, keys, and RoPE. Slash-dominance appears when this attention score matrix has high values on a certain sub-diagonal line with offset Δ.

Definition 3.1 ((κ, Δ)-Slash-Dominance)**.** *Let $\mathcal{D}$ be the distribution of prompts. For any given lag $\Delta \in \mathbb{N}$ and threshold $\kappa \in [0, 1]$, an attention head is said to be κ slash-dominant at lag Δ, if* average slash score :

$$\mathbb{E}_{P\sim\mathcal{D}}\left[\frac{1}{N(P)-\Delta}\sum\nolimits_{i=\Delta+1}^{N(P)} S_{i,i-\Delta}(P)\right] \geq \kappa. \quad (6)$$

Here, $N(P)$ is the length of the prompt P, $S(P)$ is the attention score matrix on this head for the prompt P, and $S_{i,j}(P)$ is the (i, j) component in the attention score matrix $S(P)$, denoting the attention score from position i to j.

We highlight that both $N(P)$ and $S(P)$ depend on the prompt P, which is randomly sampled according to the pretraining data distribution $\mathcal{D}$. In the following, we refer to the left-hand side of (6) as the *average slash score* at lag Δ. Intuitively, it measures the average attention paid to tokens that are Δ positions before the current token. In the attention score matrix, these connections form a slash line parallel to the main diagonal, which motivates the name *slash-dominance*. As shown in this definition, an attention head is an SDH if *the average attention score on some sub-diagonal line with offset Δ is larger than a threshold κ*. It only concerns a single slash line with lag Δ. This definition includes the induction head circuits as a special case. In particular, the first layer of an induction head circuit always copies the information from the previous token, which corresponds to a specialized SDH with $\Delta = 1$ and $\kappa \approx 1$ (Olsson et al., 2022). Furthermore, the choice of κ in (6) is crucial. Setting it too low will identify too many heads as SDHs and thus not specific enough, while setting it too high will identify too few heads and thus not sensitive enough.

Empirical Identification of SDHs. We identify SDHs in three LLMs: Qwen2.5-7B-Instruct (Qwen Team, 2025), Llama3-8B-Instruct (Grattafiori et al., 2024), and Gemma-7B (Gemma Team, 2024). To ensure that the identified SDHs are prominent, we set $\kappa = 0.1$ with a fixed context length of 6000. We let $\mathcal{D}$ denote the distribution of prompts in the `LongBenchV2` benchmark (Bai et al., 2025), which provides a heterogeneous long-context workload and yields stable cross-model statistics. To approximate the expectation in (6), we sample 500 prompts from `LongBenchV2` and truncate each to 6000 tokens, ensuring a consistent context length across models and avoiding context-extension artifacts. Following prior observations that early tokens (including BOS and warm-up positions) exhibit disproportionately large attention scores (Xiao et al., 2023), we exclude attention entries involving positions 1–4 when computing $S_{i,i-\Delta}$ to prevent these anomalous values from dominating the slash-dominance statistics.

We find that all three models exhibit SDHs with $\Delta < 5$, reflecting short-range dependencies in attention. In Fig. 1(a), we plot the attention maps of the identified SDHs in Qwen2.5-7B-Instruct with $\Delta = 1$. In this head, tokens at relative distances of 1 from any target token receive prominent attention. We further remark that when a smaller value of κ is used, as in Fig. 1(b), slash dominance also arises at large offsets $\Delta > 1000$, with much smaller average slash scores on the order of 10^{-3}. Accordingly, in this section, we study the characteristics and mechanisms of SDHs with small offsets, i.e., $\Delta < 5$, while the analysis of long-range SDHs is deferred to Appendix C. Additional experimental details are provided in Appendix F, and the full list of identified SDHs is provided in Appendix H.

3.1. OOD Generalization of SDHs

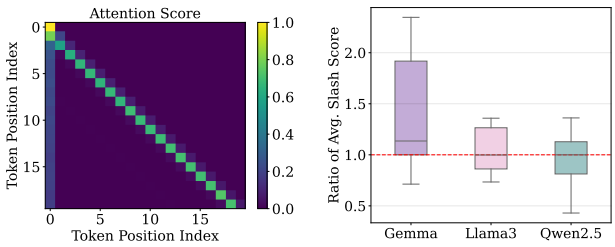
To understand the mechanism of SDHs, we begin by posing the following question:

Are SDHs sensitive to prompt distribution or intrinsic to the model itself?

We answer this question by examining whether SDHs generalize to arbitrary OOD contexts. When we change the prompt distribution from $\mathcal{D}$ to another distribution $\mathcal{D}'$, if the condition in (6) still holds for the same threshold κ, then we say that the SDHs are OOD generalizable. This means that the identified SDHs are independent of the choice of the prompt distribution, and thus intrinsic to the LLM model itself. In other words, regardless of the prompted tokens, the heads identified in Appendix H consistently apply the same mechanism to form the slash pattern.

To test OOD generalization, we evaluate the average slash scores defined in (6) with a special prompt distribution $\mathcal{D}'$, where each token of the prompt is sampled i.i.d. from the uniform distribution over the alphabet. We sample 500 OOD prompts, each with 6000 random tokens. We use these OOD prompts to compute the average slash scores as given in (6) for the identified SDHs.

SDHs can generalize to OOD prompts. We plot average attention scores of identified SDHs under OOD prompts in Fig. 3(a). The SDH is from Qwen2.5-7B-Instruct, and is the same SDHs reported in Fig. 1. Results of more heads can be found in Appendix H. In Fig. 3(a), the same slash pattern persists under OOD prompts. We observe the same results for Llama3-8B-Instruct and Gemma-7B, whose details are deferred to Appendix H. Furthermore, we quantitatively



(a) Avg. attention scores of L21H15 with OOD prompts.

(b) Ratios of average slash scores with OOD and in-distribution prompts.

Figure 3. Results of OOD prompts

assess OOD generalization by computing the average slash scores for all identified SDHs with in-distribution and OOD prompts, and comparing the ratios. That is, we compute

$$\text{ratio} = \frac{\text{average slash score with OOD prompts}}{\text{average slash score with LongBench prompts}}$$

for all the identified SDHs, whose complete list is in Appendix H. We show the box plot of these ratios in Fig. 3(b). We observe that the average slash scores under OOD prompts are generally higher, or at least comparable to those under in-distribution prompts. Detailed head-level results for the OOD case are provided in Table 3 in Appendix H. Furthermore, for the identified SDHs, as long as the average slash scores on OOD prompts remain high, even if they differ from the in-distribution scores, these heads can still retrieve and forward the semantics of preceding tokens to the current token. Indeed, as shown in the box plot, most of the average slash scores on OOD prompts are larger than half of those on the in-distribution prompts and are therefore at least 0.05 (recall that we set $\kappa = 0.1$ when identifying SDHs). As a result, most of the identified SDHs remain valid when $\mathcal{D}$ is replaced by OOD prompts and κ is reduced to 0.05. Therefore, the slash-dominance pattern generalizes beyond the pretraining distribution. Emergence of the slash-dominance pattern is not relevant to the semantic meaning of the prompts, but is intrinsic to the model architecture. The main finding is summarized as follows.

> **Takeaway 1:** The SDHs identified by taking $0 \le \Delta \le 4$ and $\kappa = 0.1$ in (6) are able to generalize to OOD prompts. As a result, the slash-dominance behavior is intrinsic to the model architecture and not due to the semantic meaning of the prompts.

3.2. Approximate Low-Rankness of pre-PE Queries and Keys

Since the emergence of SDHs is intrinsic to the model, it is natural to ask how the transformer architecture contributes to it. Note that attention scores of CSA with RoPE are determined by the pre-PE queries and keys, and RoPE, as shown in (4). In the following, we examine these components in detail. We first focus on the pre-PE queries and keys and study the following question:

How do the pre-PE queries and keys contribute to the emergence of SDHs?

We show that, interestingly, on SDHs, the pre-PE queries and keys make almost no contribution to differentiating attention scores. In particular, as we will show in the following, on SDHs, the pre-PE queries and keys are almost low-rank, particularly rank-one. This means that the semantic contents of queries and keys are nearly identical across tokens and thus contribute little to the emergence of SDHs.

To motivate the exploration, we visualize L21H15 of Qwen2.5-7B-Instruct in Figure 4. As shown in the right two columns, the pre-PE queries and keys are highly similar across tokens, implying that the query and key matrices are almost low-rank, particularly rank-one. In contrast, the hidden states of different tokens exhibit substantial variation. More heads and models are provided in Appendix H, which consistently exhibit the same structural pattern as Figure 4. This preliminary observation motivates us to examine the low-rankness of the pre-PE queries and keys in detail.

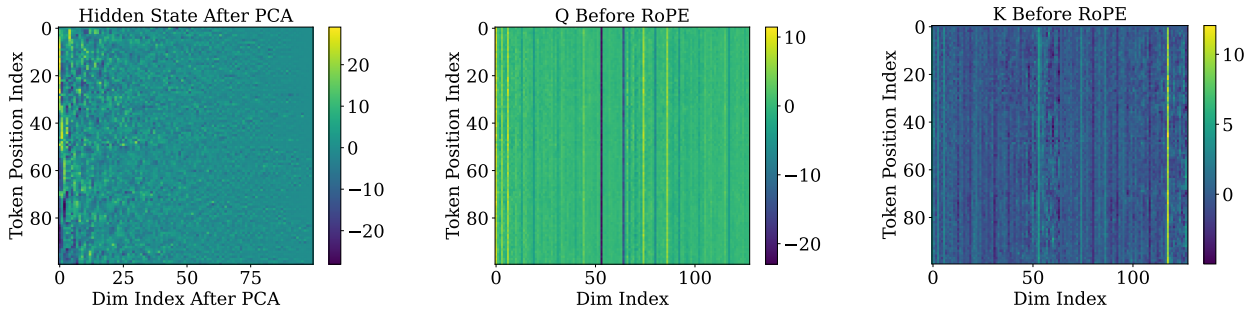
Assessing Low-Rankness of a Matrix. We adopt the following procedure to assess the low-rankness of a matrix $X \in \mathbb{R}^{N\times d}$, which contains d-dimensional features of N tokens. We perform Singular-Value Decomposition (SVD) of X and obtain the d singular values $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_d$. Then we measure the spectral decay of X using the following two quantities:

$$r_\ell(X) = \frac{\sigma_\ell^2}{\sum_{i=1}^d \sigma_i^2}, R_\tau(X) = \min\left\{\ell \in [d] \,\middle|\, \sum_{i=1}^{\ell} r_i(X) \ge \tau\right\}. \tag{7}$$

Here, $r_\ell(X)$ measures the proportion of power captured by the ℓ-th singular value in terms of the squared singular values. Moreover, $R_\tau(X)$ is the minimal subspace dimension required to account for τ of the matrix's total power, which can be viewed as the effective rank.

Thus, we use $r_1(X)$ to measure how close a matrix is to being rank-one. Similarly, we use $R_\tau(X)$ with $\tau \approx 1$ to measure how close a matrix is to being low-rank. In the sequel, we set $\tau = 0.95$ in experiments. When clear from context, we write r_1 and R_τ for brevity.

Pre-PE Queries and Keys are Almost Rank-One. According to (7), we compute r_1 and $R_{0.95}$ for the query and key matrices Q and K on each attention head. In Figure 5,



(a) Hidden state H of L21H15 after PCA.

(b) Queries Q of L21H15.

(c) Keys K of L21H15.

Figure 4. These figures show the queries and keys before the RoPE implementation, as well as the hidden states for Qwen2.5-7B-Instruct. The queries and keys each have dimension 128, and the hidden states are reduced to 128 dimensions for demonstration.

we respectively report the average values of these metrics over all heads and over SDHs only. Head-wise values are provided in Table 3 of Appendix H. Particularly, Figure 5(a) shows that, for each model, at least one of $r_1(Q)$ and $r_1(K)$ strictly exceeds 0.9 on average over SDHs only, and these values are strictly lower on the other heads. Figure 5(b) shows that the effective ranks of Q and K are low only on the SDHs, while these matrices on the other heads have much higher ranks. These results show that the pre-PE queries and keys have substantially lower ranks on SDHs, and at least one of the queries and keys is almost rank-one. This rank-one observation is consistent with Figure 4-(b) and (c). We highlight this finding as follows.

> **Takeaway 2:** On the SDHs identified by taking $0 \leq \Delta \leq 4$ and $\kappa = 0.1$ in (6), the pre-PE queries Q and keys K all have low effective ranks, measured via $R_{0.95}$ defined in (7). Moreover, at least one of Q and K is close to being rank-one.

A natural follow-up question is: *Why do SDHs have approximately rank-one queries and keys?* Due to space limitations, and since this point does not affect our main argument, we summarize only the intuition here and defer the detailed analysis to Appendix B.

> **Takeaway 3:** In the zeroth layer, the query or key matrices of SDHs are approximately rank-one for one of the following two reasons: (i) an overly large bias parameter $\boldsymbol{b}_Q$ or $\boldsymbol{b}_K$, as observed in Qwen2.5-7B-Instruct; or (ii) token embeddings lying approximately on a cone $\mathcal{C}(\mathbf{v}, c) = \{\mathbf{x} \in \mathbb{R}^d \,|\, \mathbf{v}^\top \mathbf{x}/\|\mathbf{x}\| = c\}$ for some axis vector $\mathbf{v}$ and scalar c, as observed in Gemma-7B and Llama3-8B-Instruct.

In this section, we show that at least one of pre-PE queries Q and keys K of SDHs is approximately low-rank and, in particular, rank-one, as indicated by their high r_1 values. This implies that queries and keys across different tokens are largely similar. These shared query and key vectors are then transformed by RoPE, which introduces the position information by rotating different two-dimensional sub-vectors in queries and keys with different frequencies, as in (3). The inner product between the RoPE-transformed queries and keys form the attention logit matrix. Therefore, in the SDHs, the slash pattern, which involves difference in attention scores across different tokens, is primarily determined by the position information introduced by RoPE. In the following subsection, we aim to answer the question:

How does RoPE contribute to SDHs?

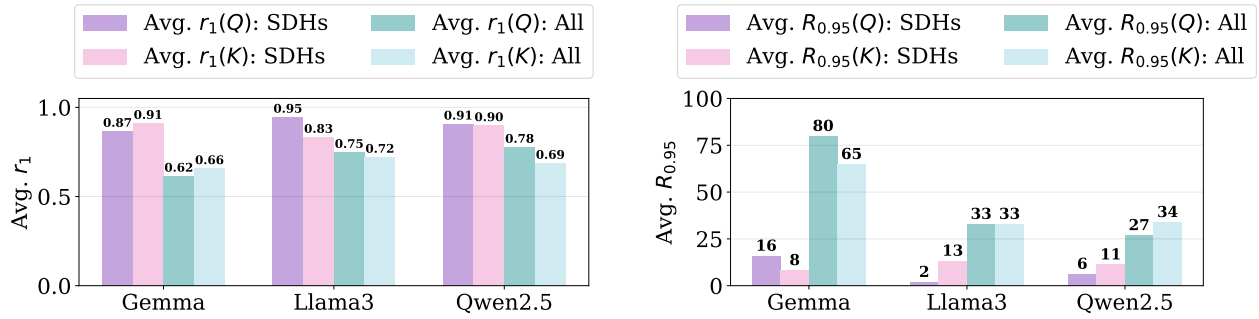
3.3. Collaboration of Frequencies in RoPE Determines Slash Pattern

Recall that we use $\widetilde{\mathbf{q}}_i$ and $\widetilde{\mathbf{k}}_j$ to denote the RoPE-transformed queries and keys, respectively. (See Section 2 for the definition of CSA with RoPE.) We first examine how RoPE gives rise to SDHs when pre-PE queries and keys are approximately rank-one. To this end, we decompose each entry of the attention logit matrix according to the used frequencies in RoPE as

$$\widetilde{\mathbf{q}}_i^\top \widetilde{\mathbf{k}}_j = \sum_{\ell=1}^{d/2} \mathrm{InP}(i, j, \ell), \tag{8}$$

where $\mathrm{InP}(i, j, \ell) = \big[R_{\boldsymbol{\vartheta},i}\mathbf{q}_i\big]_{2\ell-1:2\ell}^\top \big[R_{\boldsymbol{\vartheta},j}\mathbf{k}_j\big]_{2\ell-1:2\ell}$. Here, RoPE multiplies the input with the matrix $R_{\boldsymbol{\vartheta},i}$ (defined in (3)), which is block-diagonal and consists of a sequence of 2×2 rotation matrices. Each 2×2 rotation matrix rotates a two-dimensional sub-vector of $\mathbf{q}_i$ by an angle $i \cdot \theta_\ell$. In this way, the sum above can be expressed as a Fourier-like form with the inner product

$$\mathrm{InP}(i, j, \ell) = A_\ell \cdot \cos\big(\theta_\ell \cdot (i - j) + \varphi_\ell\big), \tag{9}$$



(a) Avg. $r_1(\uparrow)$ of all heads and SDHs with small Δ.

(b) Avg. $R_{0.95}(\downarrow)$ of all heads and SDHs with small Δ.

Figure 5. Comparison of average r_1, $R_{0.95}$ of SDHs with those of all heads.

where A_ℓ and φ_ℓ are almost independent of i and j, and θ_ℓ comes from RoPE. Since Q and K are approximately rank-one, the direction of $\mathbf{q}_i$ is approximately the same for all i (similar for $\mathbf{k}_j$). We further compute the relative variation of the norms of query and key vectors and report their averages over tokens and SDHs in Figure 6. As shown in the figure, the relative variation is at most 0.12, indicating that the norms of $\mathbf{q}_i$ (and similarly $\mathbf{k}_j$) are nearly constant. Consequently, both the directions and magnitudes of the queries and keys are approximately invariant across token pairs. Thus, A_ℓ and φ_ℓ are approximately identical across token pairs, as they only depend on queries and keys. From (9), it follows that for any token pair (i, j), the corresponding logit, and hence the attention score, depends only on the relative distance $i - j$, which leads to the slash pattern.

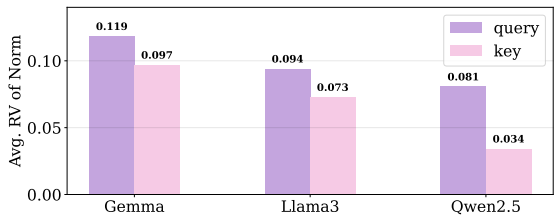


Figure 6. Avg. of the relative variation of the norm of the query vector and key.

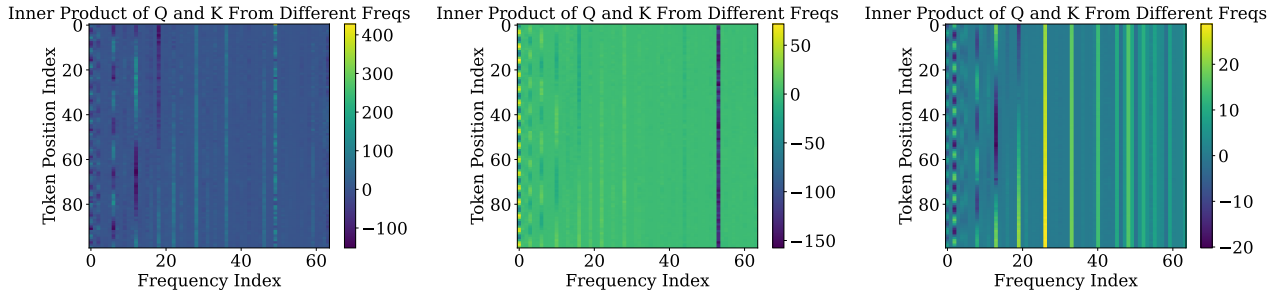
Contributions of RoPE Frequencies. Next, we examine how the different RoPE frequencies collectively contributed to the slash pattern according to (8). Note that $\mathrm{InP}(i, j, \ell)$ only involves the influence of the ℓ-th frequency on the logit from position i to j. We visualize $\mathrm{InP}(i, j, \ell)$ with i fixed to $i = 100$ and j varying within $[1, i]$ in Figure 7, which is based on Qwen2.5-7B-Instruct. Here, each column corresponds to the influence of one frequency on different token positions. We observe clear periodic patterns within columns, which reflect the rotational effect of RoPE on queries and keys. The magnitude of influence is proportional to the variation within each column. Qualitatively, high- and medium-frequency components ($\ell \in [1, 42]$) exhibit greater variation than low-frequency components ($\ell \in [43, 64]$).

For a more quantitative characterization, we evaluate slash-dominance after selectively removing certain frequencies when computing the attention score. Specifically, we split the 64 frequencies into three groups — high ($\ell \in [1, 21]$), medium ($\ell \in [22, 42]$), and low ($\ell \in [43, 64]$). For each group, we remove six frequencies uniformly in each interval, e.g., $\ell = 1, 5, 9, 13, 17, 21$ for $\ell \in [1, 21]$ (10% of the total), whose corresponding sub-vectors are left unrotated, and compute the attention logits still as in (8). Using these modified attention logits, we compute attention scores using the softmax function and average slash score as Definition 3.1. We compute the ratio of the new score to the original score for each identified SDHs, and show the box plot in Figure 8. We observe that, when removing high frequencies, the average slash score decreases in all three models. Removing the medium frequencies leads to a mild decrease, while the low frequencies yield the least drop. Therefore, the high frequencies are the most critical for SDHs, medium frequencies have a moderate impact, and low frequencies contribute the least.

Takeaway 4: The slash pattern appears because the attention logits approximately admit a Fourier-like decomposition as in (8) and (9). The high- and medium-frequency components in RoPE are more important than the low-frequency components.

4. Theoretical Study of Shallow Transformers

In the previous section, we observed that only a subset of RoPE frequencies contributes substantially to the slash pattern. Motivated by this observation, in this section, we propose a concrete sufficient condition on the RoPE fre-



(a) $\mathrm{InP}(100, j, \ell)$ of L18H7. (b) $\mathrm{InP}(100, j, \ell)$ of L21H15. (c) $\mathrm{InP}(100, j, \ell)$ of L1H3.

Figure 7. Heatmaps of $\mathrm{InP}(i, j, \ell)$ on various SDHs of Qwen2.5-7B-Instruct. We set $i = 100$ and let j vary. Qwen2.5-7B-Instruct has 64 frequencies ($1 \le \ell \le 64$). The variation within each column shows the influence of each frequency.

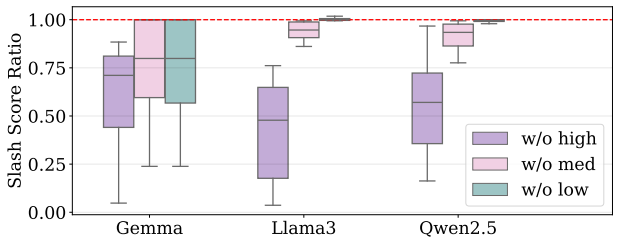


Figure 8. The figure quantifies the effect of low-, medium-, and high-frequency components on SDHs.

quencies, which is validated through empirical experiments. We can then analyze the training dynamics of a shallow transformer and show that, under this condition, a shallow Transformer can learn an SDH. Due to the page limit, we defer the detailed theoretical analysis to Appendix D.

Assumption 4.1 (Slash-Dominance Frequency sequence)**.** *Let N be the fixed prompt length and d be the embedding dimension. Let $\delta_0(x) := \mathbb{1}\{x = 0\}$. We assume that the frequency sequence $\boldsymbol{\vartheta} = (\theta_1, \ldots, \theta_{d/2})$ satisfies the following pulse-approximation condition: there exist constants $C_1 \ge 0$, $C_2 \in \mathbb{R}$, and a sufficiently small tolerance ϵ_{FN} such that, for any $x \in \mathbb{Z}$ with $|x| \le N$,*

$$\left| \sum_{s=1}^{d} \cos(\theta_s x) - C_1 \delta_0(x) - C_2 \right| \le \epsilon_{\mathrm{FN}}. \tag{10}$$

This assumption states that if the selected frequency components have identical amplitudes 1, then the resulting sum of sinusoidal components approximates a pulse, which is represented by the Kronecker delta function $\delta_0(x)$ as in (10). Here, the uniform amplitudes are intrinsic to the Fourier expansion of the approximation target $\delta_0(x)$, which distributes equal amplitude across all basis frequencies.

Assumption 4.1 further verifies that not all frequencies in RoPE are required for the SDHs. In particular, the low frequencies do not affect the validity of Assumption 4.1, consistent with Takeaway 4. On the contrary, if medium-frequencies or high-frequencies are set to zero, Assumption 4.1 may be violated, and slash-dominance no longer holds.

Empirical Validation of Assumption 4.1. We remark that the slash-dominance frequency condition is satisfied by pretrained models. Specifically, we first identify the *active frequencies*, defined as those whose average $\mathrm{InP}(i, j, l)$ exceeds one-tenth of the maximum. These active frequencies contribute primarily to the attention scores. We plot their sinusoidal sum given on the left-hand side of (10). As shown in Figure 23 (in Appendix D), the active frequencies employed in open-source LLMs (e.g., Gemma-7B, Llama3-8B-Instruct, Qwen2.5-7B-Instruct) satisfy this condition.

> **Takeaway 5:** Assumption 4.1 is empirically validated and serves as a sufficient condition on the RoPE frequencies for inducing SDHs.

5. Discussion and Conclusion

In the preceding sections, we investigated the mechanistic interpretability of SDHs, showing that they arise as an intrinsic algorithmic effect induced by RoPE. Beyond interpretability, we also outline potential implications of our findings for improving current model architectures, with detailed discussions deferred to Appendix E.

In this paper, we investigated the mechanism underlying SDHs. Our results demonstrate that, owing to the geometric structure of token embeddings approximately on a cone, slash-dominance is an intrinsic algorithmic effect primarily driven by the medium- to high-frequency components of RoPE. We further establish a slash-dominance frequency condition that is satisfied by real open-source LLMs, and prove these conditions on token embeddings and frequencies are sufficient for the emergence of SDHs by analyzing the training dynamics of a shallow Transformer. Our theoretical and empirical findings provide new insights into how RoPE interacts with attention mechanisms. Our work focused on the analysis of the pretrained LLMs with RoPE, and we leave the investigation of other kinds of PE as future work.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *International Conference on Learning Representations*, 2023.

Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., Dong, Y., Tang, J., and Li, J. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 3119–3137, 2024.

Bai, Y., Tu, S., Zhang, J., Peng, H., Wang, X., Lv, X., Cao, S., Xu, J., Hou, L., Dong, Y., Tang, J., and Li, J. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 3639–3664, 2025.

Barbero, F., Vitvitskyi, A., Perivolaropoulos, C., Pascanu, R., and Veličković, P. Round and round we go! what makes rotary positional encodings useful? In *International Conference on Learning Representations*, 2025.

Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H., and Bottou, L. Birth of a Transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1560–1588, 2023.

Chen, S., Wong, S., Chen, L., and Tian, Y. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Chen, S., Sheen, H., Wang, T., and Yang, Z. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. In *Annual Conference on Learning Theory*, 2024a.

Chen, S., Sheen, H., Wang, T., and Yang, Z. Unveiling induction heads: Provable training dynamics and feature learning in Transformers. In *Advances in Neural Information Processing Systems*, volume 37, pp. 66479–66567, 2024b.

Cheng, Y., Huang, Y., Xiong, Z., Liang, Y., and Tan, V. Y. Transformers provably learn directed acyclic graphs via kernel-guided mutual information. *arXiv preprint arXiv:2510.25542*, 2025.

Ding, Y., Zhang, L. L., Zhang, C., Xu, Y., Shang, N., Xu, J., Yang, F., and Yang, M. LongRoPE: Extending LLM context window beyond 2 million tokens. In *International Conference on Machine Learning*, pp. 11091–11104. PMLR, 2024.

Edelman, E., Tsilivis, N., Edelman, B., Malach, E., and Goel, S. The evolution of statistical induction heads: In-context learning Markov chains. In *Advances in Neural Information Processing Systems*, volume 37, pp. 64273–64311, 2024.

Ekbote, C., Bondaschi, M., Rajaraman, N., Lee, J. D., Gastpar, M., Makkuva, A. V., and Liang, P. P. What one cannot, two can: Two-layer Transformers provably represent induction heads on any-order Markov chains. *arXiv preprint arXiv:2508.07208*, 2025.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for Transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Gemma Team. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Gould, R., Ong, E., Ogden, G., and Conmy, A. Successor heads: Recurring, interpretable attention heads in the wild. In *International Conference on Learning Representations*, 2024.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Huang, R., Liang, Y., and Yang, J. How transformers learn regular language recognition: A theoretical study on training dynamics and implicit bias. In *International Conference on Machine Learning*, 2025a.

Huang, Y., Cheng, Y., and Liang, Y. In-context convergence of Transformers. In *International Conference on Machine Learning*, pp. 19660–19722. PMLR, 2024.

Huang, Y., Wen, Z., Singh, A., Chi, Y., and Chen, Y. Transformers provably learn chain-of-thought reasoning with length generalization. *arXiv preprint arXiv:2511.07378*, 2025b.

Jelassi, S., Sander, M., and Li, Y. Vision Transformers provably learn spatial structure. In *Advances in Neural Information Processing Systems*, volume 35, pp. 37822–37836, 2022.

Jiang, H., Li, Y., Zhang, C., Wu, Q., Luo, X., Ahn, S., Han, Z., Abdi, A. H., Li, D., Lin, C.-Y., et al. Minference 1.0: Accelerating pre-filling for long-context LLMs via dynamic sparse attention. In *Advances in Neural Information Processing Systems*, volume 37, pp. 52481–52515, 2024.

Jin, M., Mei, K., Xu, W., Sun, M., Tang, R., Du, M., Liu, Z., and Zhang, Y. Massive values in self-attention modules are the key to contextual knowledge understanding. In *International Conference on Machine Learning*, 2025.

Kazemnejad, A., Padhi, I., Natesan Ramamurthy, K., Das, P., and Reddy, S. The impact of positional encoding on length generalization in Transformers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24892–24928, 2023.

Kim, J. and Suzuki, T. Transformers provably solve parity efficiently with chain of thought. In *International Conference on Learning Representations*, 2025.

Lai, X., Lu, J., Luo, Y., Ma, Y., and Zhou, X. FlexPrefill: A context-aware sparse attention mechanism for efficient long-sequence inference. In *International Conference on Learning Representations*, 2025.

Li, H., Wang, M., Liu, S., and Chen, P.-Y. A theoretical understanding of shallow vision Transformers: Learning, generalization, and sample complexity. In *International Conference on Learning Representations*, 2023.

Li, W., Zhang, C., Jiang, H., Li, Y., Yang, Y., and Qiu, L. Mtraining: Distributed dynamic sparse attention for efficient ultra-long context training. *arXiv preprint arXiv:2510.18830*, 2025.

Merullo, J., Eickhoff, C., and Pavlick, E. Circuit component reuse across tasks in Transformer language models. In *International Conference on Learning Representations*, 2024.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT press, 2018.

Nichani, E., Damian, A., and Lee, J. D. How Transformers learn causal structure with gradient descent. In *International Conference on Machine Learning*, pp. 38018–38070. PMLR, 2024.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Peng, B., Quesnelle, J., Fan, H., and Shippole, E. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations*, 2024.

Press, O., Smith, N. A., and Lewis, M. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Reddy, G. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *International Conference on Learning Representations*, 2024.

Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Sauvestre, R., Remez, T., et al. Code Llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Singh, A. K., Moskovitz, T., Hill, F., Chan, S. C., and Saxe, A. M. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. In *International Conference on Machine Learning*, pp. 45637–45662. PMLR, 2024.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. RoFormer: Enhanced Transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, M., Yu, R., Wu, L., et al. How Transformers implement induction heads: Approximation and optimization analysis. *arXiv preprint arXiv:2410.11474*, 2024a.

Wang, S., Kobyzev, I., Lu, P., Rezagholizadeh, M., and Liu, B. Resonance RoPE: Improving context length generalization of large language models. In *Findings of the Association for Computational Linguistics*, pp. 586–598, 2024b.

Wen, K., Zhang, H., Lin, H., and Zhang, J. From sparse dependence to sparse attention: Unveiling how chain-of-thought enhances transformer sample efficiency. In *International Conference on Learning Representations*, 2025.

Wu, W., Wang, Y., Xiao, G., Peng, H., and Fu, Y. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, 2025.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022.

Xiong, J., Fan, L., Shen, H., Su, Z., Yang, M., Kong, L., and Wong, N. DoPE: Denoising rotary position embedding. *arXiv preprint arXiv:2511.09146*, 2025.

Xiong, W., Liu, J., Molybog, I., Zhang, H., Bhargava, P., Hou, R., Martin, L., Rungta, R., Sankararaman, K. A., Oguz, B., Khabsa, M., Fang, H., Mehdad, Y., Narang, S., Malik, K., Fan, A., Bhosale, S., Edunov, S., Lewis, M., Wang, S., and Ma, H. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024.

Xu, R., Xiao, G., Huang, H., Guo, J., and Han, S. Xattention: Block sparse attention with antidiagonal scoring. In *International Conference on Machine Learning*, 2025.

Yang, T., Huang, Y., Liang, Y., and Chi, Y. In-context learning with representations: Contextual generalization of trained Transformers. In *Advances in Neural Information Processing Systems*, volume 37, pp. 85867–85898, 2024.

Yang, T., Huang, Y., Liang, Y., and Chi, Y. Multi-head transformers provably learn symbolic multi-step reasoning via gradient descent. In *Advances in Neural Information Processing Systems*, 2025.

Zhang, B. and Sennrich, R. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Zhang, R., Frei, S., and Bartlett, P. L. Trained Transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024a.

Zhang, R., Wu, J., and Bartlett, P. In-context learning of a linear Transformer block: Benefits of the MLP component and one-step GD initialization. *Advances in Neural Information Processing Systems*, 37:18310–18361, 2024b.

Zhang, Y., Zhang, F., Yang, Z., and Wang, Z. What and how does in-context learning learn? Bayesian model averaging, parameterization, and generalization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1684–1692. PMLR, 2025.

Zhao, T., Hong, K., Yang, X., Xiao, X., Li, H., Ling, F., Xie, R., Chen, S., Zhu, H., Zhang, Y., et al. PAROAttention: Pattern-aware reordering for efficient sparse and quantized attention in visual generation models. *arXiv preprint arXiv:2506.16054*, 2025.

Zhong, M., Zhang, C., Lei, Y., Liu, X., Gao, Y., Hu, Y., Chen, K., and Zhang, M. Understanding the RoPE extensions of long-context LLMs: An attention perspective. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8955–8962, 2025.

Supplementary Material

We summarize the content of the appendix as follows. We first discuss related work in Appendix A.

As a supplement to the discussion in Section 3, Appendix B provides a detailed analysis of why the pre-PE queries and keys of SDHs become approximately rank-one. Appendix C extends our empirical study to long-range SDHs with large offsets Δ.

As a supplement to the discussion in Section 4, Appendix D presents the theoretical analysis of learning dynamics of shallow Transformers, including the data model, network architecture, training algorithm, and main theorems.

Appendix E discusses potential applications of our findings, including parameter-efficient query/key matrices and implications for other types of PE.

Appendix F provides experimental details. Appendix G reports the parameter-compression experiment motivated by the observed low-rankness of queries and keys as support of Appendix E

Appendix H reports additional empirical results and complete lists of identified SDHs.

Finally, Appendix I to L provides the detailed proof of theorems in Appendix D.

A. Related Works

Induction Head Circuit. First introduced by Elhage et al. (2021); Olsson et al. (2022), *induction head circuit* is a specialized cascade of attention heads in transformer models that play a central role in *in-context learning*. The induction head circuit typically involves two attention heads, a forwarding head and a feature-matching head, working in tandem through a forward-then-match process to transmit semantics and complete an answer, as in Figure 9. To examine its emergence closely, subsequent empirical work has explored controlled synthetic settings (Reddy, 2024; Bietti et al., 2023; Edelman et al., 2024; Singh et al., 2024). In particular, Edelman et al. (2024) designed a Markov-chain–based in-context learning task and showed that transformers develop "statistical induction heads" through a multi-phase training. Bietti et al. (2023) revealed a two-phase learning process: the rapid acquisition of global bigram statistics, followed by the slower development of induction heads. From a theoretical perspective, Nichani et al. (2024); Chen et al. (2024b); Edelman et al. (2024); Ekbote et al. (2025) investigated induction heads under underlying causal structures such as trees or n-gram Markov chains, (Barbero et al., 2025)and Wang et al. (2024a) demonstrated that two-layer transformers can efficiently represent both standard and generalized induction head mechanisms. However, all these works neglected the contribution of RoPE to the induction head. Concretely, they employed either vanilla one-hot PE (Nichani et al., 2024) or ALiBi (Wang et al., 2024a) and its variants (Chen et al., 2024b; Ekbote et al., 2025). As a result, their assumptions and conclusions did not fully reflect the empirical behavior of many real-world LLMs. Xie et al. (2022); Zhang et al. (2025) considered in-context learning from a different Bayesian perspective.

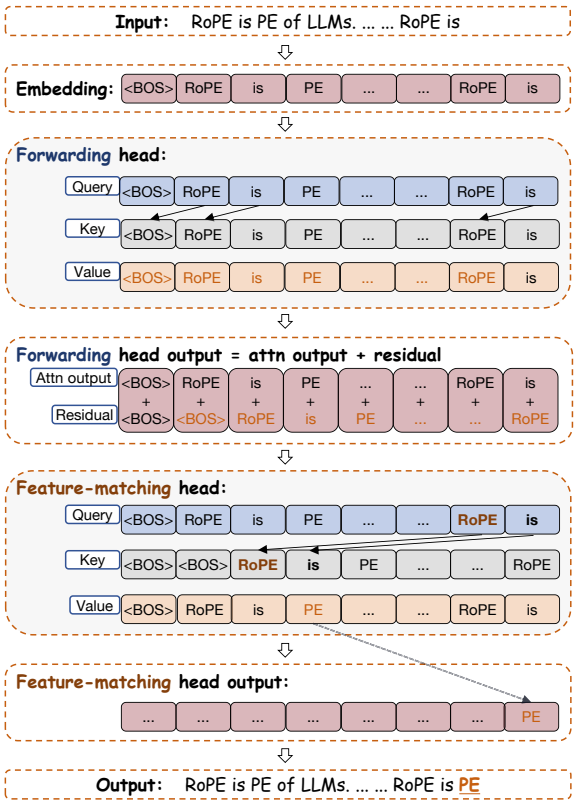


Figure 9. The forwarding head forwards semantic information from the prefix to the current token. In the feature-matching head, the target question token ("is") matches tokens whose prefixes exhibit similar semantics to the target, enabling the model to generate the correct continuation "PE".

Other Special Attention Patterns and Heads. In addition to the slash patterns and SDHs, a line of work proposed different kinds of heads with special patterns to explain the information aggregation in transformers. For example, successor heads (Gould et al., 2024), content-gathering head (Merullo et al., 2024), and retrieval head (Wu et al., 2025). Different from them, in our work, we also show that SDHs can be generalized to out-of-distribution cases. Moreover, by understanding certain attention score patterns of special heads (Jiang et al., 2024; Xu et al., 2025; Zhao et al., 2025; Lai et al., 2025; Li et al., 2025), one can improve the efficiency of LLMs. Specifically, Xu et al. (2025) observed that the sum of the antidiagonal values in the attention score matrix serves as a strong proxy for block importance, based on which they proposed a plug-and-play framework that accelerates long-context inference in transformer models using sparse attention. Li et al. (2025) proposes a distributed dynamic sparse attention framework for ultra-long context training. Their system leverages the empirical observation that many attention heads attend primarily along fixed relative offsets induced by RoPE, enabling efficient training at a million-token scale.

Rotary Position Embeddings (RoPE). Proposed by Su et al. (2024), RoPE introduced multiplicative "rotations" on queries and keys so attention implicitly depends on relative positions, and became the default PE in many LLMs (Qwen Team, 2025; Yang et al., 2024; Gemma Team, 2024; Grattafiori et al., 2024). Following RoPE, a series of works extended the context length window of the pretrained models by modifying the base frequency of RoPE (Xiong et al., 2024; Roziere et al., 2023; Xiong et al., 2025), interpolating the position indexes and frequencies (Chen et al., 2023; Peng et al., 2024; Ding et al., 2024; Zhong et al., 2025), and controlling the RoPE feature gaps (Wang et al., 2024b). Another line of work investigated the mechanism behind RoPE. Barbero et al. (2025) showed that rather than simply inducing attention decay with distance, models like Gemma-7B used RoPE's high frequencies to build robust positional attention heads and low frequencies to carry semantic information, and further proposed a modified variant (p-RoPE) that improves performance. Xiong et al. (2025) proposes DoPE, a training-free denoising method for RoPE that suppresses noisy positional frequency components in attention maps via truncated matrix entropy, improving long-context length extrapolation. In addition, RoPE has been shown to induce various distinctive structural patterns in LLMs. Jin et al. (2025) empirically shows that RoPE-induced "massive values" in the queries and keys of self-attention emerge consistently across layers and play a critical role in contextual knowledge understanding in LLMs, with perturbing these values significantly degrading performance on context-dependent tasks.

Training Dynamics of Transformer. Huang et al. (2024) and Zhang et al. (2024a) are the first two works to establish the in-context convergence results for the softmax attention and linear attention trained with gradient descent, respectively. Then, a series of works relaxed their assumptions and studied more general network structures. Chen et al. (2024a) first explored the training dynamics of the multi-task in-context learning using multi-head transformers. Yang et al. (2024) explored in-context learning in multi-head transformers applied to nonlinear task functions. More recently, Nichani et al. (2024); Chen et al. (2024b); Edelman et al. (2024); Cheng et al. (2025) have investigated transformer learning dynamics in settings underpinned by causal structure, deepening our understanding of how such models acquire and represent inductive capabilities. Wen et al. (2025); Kim & Suzuki (2025); Yang et al. (2025); Huang et al. (2025a;b) investigated transformer dynamics in the chain-of-thought reasoning setting.

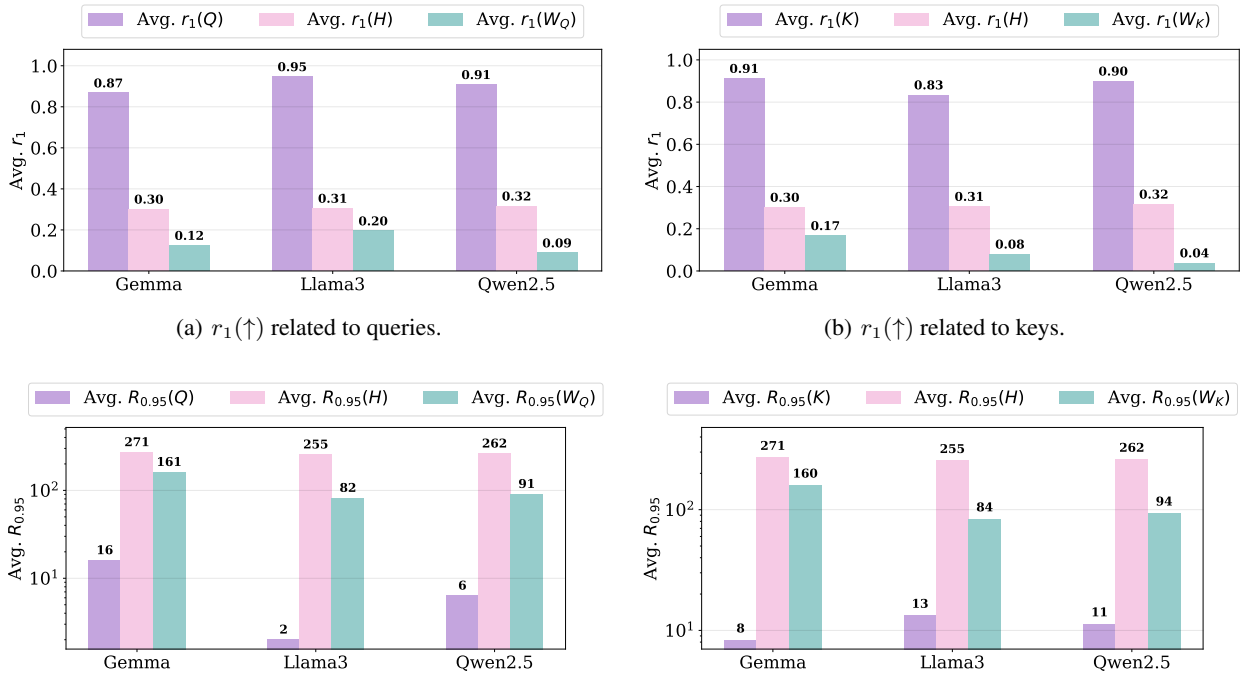
Another line of work studies the learning dynamics of transformers on vision tasks. For example, Jelassi et al. (2022) analyzed the inductive biases of Vision Transformers (ViTs) with a focus on specialized positional attention. Building on this foundation, Li et al. (2023) analyzed the training process of shallow ViTs in supervised classification settings and extended their study to in-context learning, under strict assumptions on network initialization.

B. How to Get Approximately Rank-One Queries and Keys?

In this section, we answer the follow-up question:

How SDHs achieve the approximately rank-one queries and keys, and what the shared low-dimensional space induced by Q and K represents?

Since the queries and keys are obtained by $Q = HW_Q$ and $K = HW_K$, where H consists of the hidden states, and W_Q and W_K are the weight matrices, a natural conjecture is that either H or one of W_Q and W_K is low-rank. To test this hypothesis, we compute the average values of r_1 and $R_{0.95}$ for H, W_Q (resp. W_K), and Q (resp. K) over the identified SDHs. These values are plotted in Figure 10. This figure clearly shows that the effective ranks of H, W_Q, and W_K are much higher than those of Q and K, and thus clearly refutes this hypothesis. Therefore, low-rankness of Q and K must arise from the interaction between the hidden states H and weight matrices W_Q and W_K.



(a) $r_1(\uparrow)$ related to queries.

(b) $r_1(\uparrow)$ related to keys.

(c) $R_{0.95}(\downarrow)$ related to queries.

(d) $R_{0.95}(\downarrow)$ related to keys.

Figure 10. Comparison of average r_1 and $R_{0.95}$ for query/key matrices relative to the hidden states and their corresponding weight matrices, for SDHs with *small* Δ. Both hidden states H and the weight matrices W_Q, W_K exhibit much higher rank than the resulting pre-PE queries and keys.

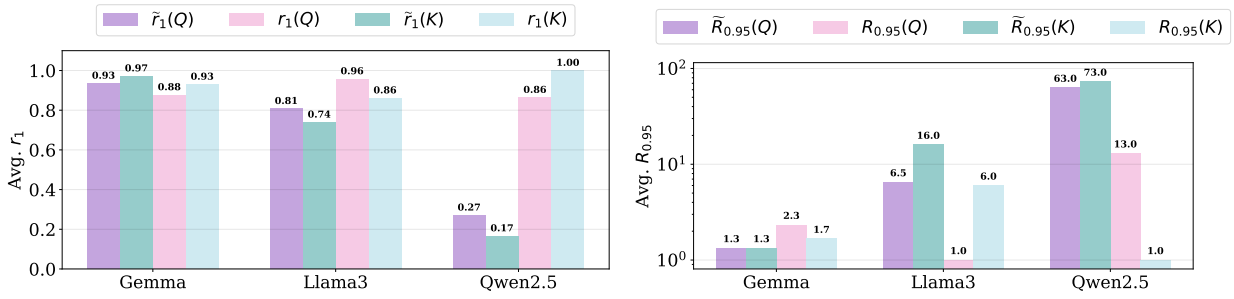
Analysis of Token Embeddings. In general, analyzing the hidden states H is difficult, since at each layer they are formed through complex transformations of all preceding layers and vary with the prefix sequence due to prior attention. To simplify, we focus on the 0-th layer, where H is the input to the first transformer layer, i.e., the token embeddings. In this case, we set the prompt P as the concatenation of all tokens, whose length is the size of the alphabet. Let $H = H(P)$ denote the token embeddings of this prompt, whose i-th vector is denoted by $\mathbf{h}_i \in \mathbb{R}^d$ for each i. With these hidden states, in the first transformer layer, the i-th query of CSA is given by $\mathbf{q}_i = W_Q^\top \mathbf{h}_i$ for Gemma-7B and Llama3-8B-Instruct, and $\mathbf{q}_i = W_Q^\top \mathbf{h}_i + \mathbf{b}_Q$ for Qwen2.5-7B-Instruct. Recall that Qwen2.5-7B-Instruct has additional bias terms. We have similar expressions for the i-th key $\mathbf{k}_i$ using $\mathbf{h}_i$, W_K (and $\mathbf{b}_K$).

Metrics for Subspace Alignment. Similar to r_ℓ and R_τ defined in (7), we introduce metrics that characterize how well a vector $\mathbf{x} \in \mathbb{R}^d$ aligns with the leading subspaces of W_Q and W_K. Recall that $W_Q \in \mathbb{R}^{d \times d_1}$ where d is the latent dimension of the model and $d_1 = d/\texttt{num_heads}$ with $\texttt{num_heads}$ being the number of attention heads. Let $W_Q = \sum_{\ell=1}^{d_1} \sigma_\ell \mathbf{v}_\ell \mathbf{u}_\ell^\top$ denote the SVD of W_Q, where $\{\sigma_\ell\}_{\ell=1}^{d_1}$ are the singular values in descending order. Note that $\{\mathbf{v}_\ell\}_{\ell=1}^{d_1} \subseteq \mathbb{R}^d$. For any $\ell \in [d_1]$, we let $\ell(\mathbf{x})$ denote the index of ℓ-th largest element among $\{(\sigma_i \cdot \mathbf{v}_i^\top \mathbf{x})^2\}_{i=1}^{d_1}$. Then we define

$$\tilde{r}_\ell(\mathbf{x}, W_Q) := \frac{\left(\sigma_{\ell(\mathbf{x})} \mathbf{v}_{\ell(\mathbf{x})}^\top \mathbf{x}\right)^2}{\sum_{i=1}^{d} \left(\sigma_i \mathbf{v}_i^\top \mathbf{x}\right)^2}, \quad \text{and} \quad \tilde{R}_\tau(\mathbf{x}, W_Q) := \min\left\{\ell \in [d_1] \,\middle|\, \sum\nolimits_{i=1}^{\ell} \tilde{r}_i(\mathbf{x}, W_Q) \geq \tau\right\}, \tag{11}$$

where $\tau \in [0, 1]$. We can define $\tilde{r}_\ell(\mathbf{x}, W_K)$ and $\tilde{R}_\tau(\mathbf{x}, W_K)$ in a similar way. Intuitively, $\tilde{r}_\ell(\mathbf{x}, W_Q)$ characterizes how well $\mathbf{x}$ aligns with the $\ell(\mathbf{x})$-th singular vector of W_Q, and $\tilde{R}_\tau(\mathbf{x}, W_Q)$ is small when $\mathbf{x}$ is only spanned by a limited number of vectors among $\{\mathbf{v}_i\}_{i=1}^{d_1}$. This quantity enables us to quantify how transformed hidden states allocate energy across parameter-induced subspaces and, in turn, assess the coupling between hidden states and the attention weights. In the following, for brevity, we write $\tilde{r}_\ell(Q, W_Q)$ (resp. $\tilde{R}_\tau(Q, W_Q)$) as $\tilde{r}_\ell(Q)$ (resp. $\tilde{R}_\tau(Q)$), and analogously for K.

Demystifying Why Q and K are Approximately Rank-One in Zeroth Layer. We plot the average values of $\tilde{r}_1$ and $\tilde{R}_{0.95}$ in Figures 11, and compare them with r_1 and $R_{0.95}$ respectively. These average values are computed over all the tokens and



(a) $\tilde{r}_1(\uparrow)$ and $r_1(\uparrow)$ related to queries.

(b) $R_{0.95}(\downarrow)$ and $\widetilde{R}_{0.95}(\downarrow)$ related to queries.

Figure 11. Comparison of average r_1, $R_{0.95}$, $\tilde{r}_1$, and $\tilde{R}_{0.95}$ for query/key matrices relative to the 0-th layer hidden states (token embeddings) for SDHs with *small* Δ in Gemma-7B, Llama3-8B-Instruct, and Qwen2.5-7B-Instruct.

the layer zero SDHs. This figure shows that the behaviors of $\tilde{r}_1$ and $\tilde{R}_{0.95}$ closely match those of r_1 and $R_{0.95}$ in Gemma-7B and Llama3-8B-Instruct, but differ in Qwen2.5-7B-Instruct. This confirms that in Gemma-7B and Llama3-8B-Instruct, each token embedding concentrates most of its energy approximately in a one-dimensional principal subspace of W_Q and W_K. That is, each token embedding aligns well with low-rank (and almost rank-one) subspaces of W_K and W_Q.

It is natural to ask whether there exists a special one-dimensional principal subspace of each W_Q or W_K that aligns with all the token embeddings. To examine this, we define the dominating singular vector of each W_Q as

$$\{\mathbf{v}_Q, \mathbf{u}_Q\} = \{\mathbf{v}_{\ell^*}, \mathbf{u}_{\ell^*}\}, \qquad \text{where } \ell^* = \operatorname{argmax}_{\ell \in [d]} \sum\nolimits_{\mathbf{x} \in \mathcal{X}} \left(\sigma_\ell \mathbf{v}_\ell^\top \mathrm{RMSN}(\mathbf{x})\right)^2.$$

Here, $\mathcal{X}$ is the set of token embeddings, and $\mathrm{RMSN}(\cdot)$ denotes root-mean-square (RMS) normalization (Zhang & Sennrich, 2019), which is implemented before the attention modules in the pretrained models. We define $\{\mathbf{v}_K, \mathbf{u}_K\}$ similarly for each W_K. Note that each head has a different $\{W_K, W_Q\}$ and thus it is possible to have different dominating singular vectors. We examine how well the token embeddings align with these vectors by computing the relative variation (RV), which is defined as

$$\mathrm{RV}(\mathbf{v}) = \frac{\mathrm{std}\left(\{\mathbf{v}^\top \mathrm{RMS}(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}\right)}{\left|\mathrm{mean}\left(\{\mathbf{v}^\top \mathrm{RMS}(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}\right)\right|}.$$

Here, $\mathrm{std}(\cdot)$ and $\mathrm{mean}(\cdot)$ denote the standard deviation and mean of a set, respectively. We report the average values of $\mathrm{RV}(\mathbf{v}_Q)$ and $\mathrm{RV}(\mathbf{v}_K)$ over the attention heads in layer zero in Figure 12. The detailed values of each head are deferred to Table 11 in Appendix H.5. As a baseline, we also report $\mathrm{RV}(\mathbf{v}_{\text{rand}})$ where $\mathbf{v}_{\text{rand}}$ is drawn from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, I_d)$ on $\mathbb{R}^d$. This figure shows that $\mathbf{v}_Q^\top \mathrm{RMS}(\mathbf{x})$ and $\mathbf{v}_K^\top \mathrm{RMS}(\mathbf{x})$ are highly concentrated for Gemma-7B and moderately concentrated for Llama3-8B-Instruct. In other words, for each $\mathbf{x}$, the component perpendicular to $\mathbf{v}_Q$ or $\mathbf{v}_K$ has a small magnitude, compared with the parallel component. This implies that all token embeddings align well with $\mathbf{v}_Q$ and $\mathbf{v}_K$ on each head. Thus, for Gemma-7B and Llama3-8B-Instruct, in the zeroth layer, Q (resp. K) is approximately rank-one because all the token embeddings lie approximately on a cone centered around $\mathbf{v}_Q$ (resp. $\mathbf{v}_K$).

For Qwen2.5-7B-Instruct, we additionally calculate the norms of bias parameters $\|\mathbf{b}_Q\|$ (resp. $\|\mathbf{b}_K\|$) and the average values of $\|W_Q^\top \mathbf{h}_i\|$ (resp. $\|W_K^\top \mathbf{h}_i\|$) over the alphabet, which are reported in Figure 13. Detailed values of each SDHs are deferred to Table 12 in Appendix H.5. This figure shows that the average norm of hidden states is much larger than that of bias parameters, suggesting that the approximate rank-one structure of Q and K is primarily driven by overly large bias parameters.

Combining all these facts, we have a complete picture of why Q and K are approximately rank-one in the zeroth layer for the three models. The conclusion is summarized as follows.

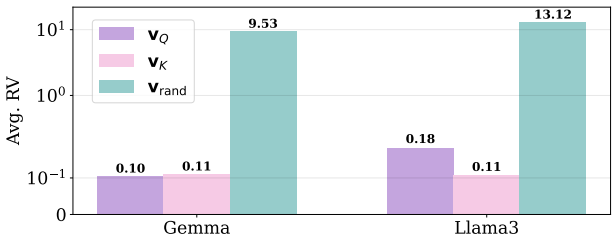


Figure 12. Average of the relative variation of token embeddings projected onto the dominant subspace of the SDHs in the 0-th layer of Gemma-7B and Llama3-8B-Instruct.

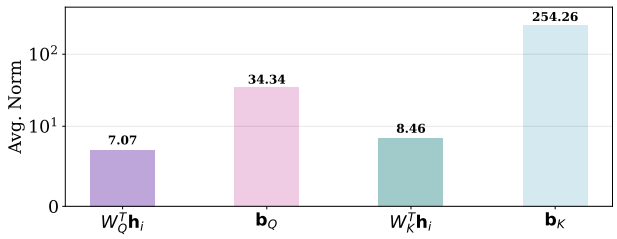
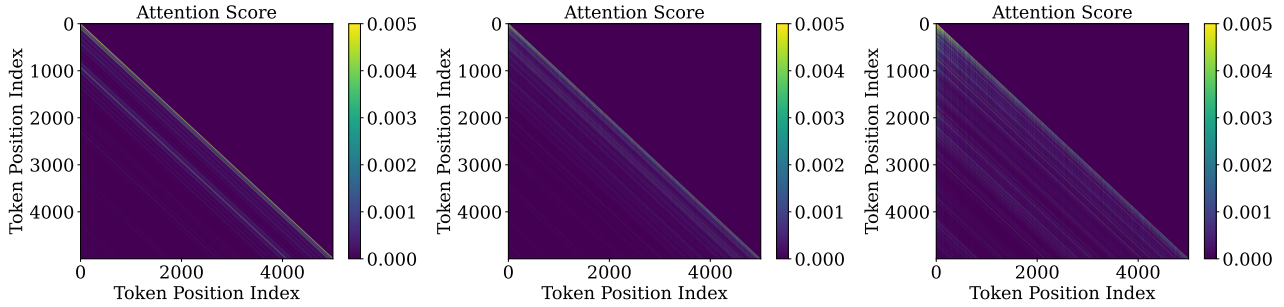


Figure 13. The average norms of $\|W_Q^\top \mathbf{h}_i\|$ and $\|W_K^\top \mathbf{h}_i\|$ over the alphabet and SDH, together with the average norms of $\|\mathbf{b}_Q\|$ and $\|\mathbf{b}_K\|$ over SDH in the 0-th layer of Qwen2.5-7B-Instruct.

Takeaway 3: In the zeroth layer, Q or K of SDHs are approximately rank-one because of one of the following two reasons: (i) the overly large bias parameter $\boldsymbol{b}_Q$ or $\boldsymbol{b}_K$ (Qwen2.5-7B-Instruct) or (ii) the fact that token embeddings are approximately on a cone $\mathcal{C}(\mathbf{v}, c) = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{v}^\top \mathbf{x}/\|\mathbf{x}\| = c\}$ for some axis vector $\mathbf{v}$ and scalar c (Gemma-7B and Llama3-8B-Instruct).

C. Extensions to SDHs with Large Δ

In this section, we show that the definition of SDHs for small Δ introduced in Definition 3.1 can be naturally extended to the large Δ (e.g., $\Delta > 500$) regime, by appropriately adjusting the threshold κ. This complements our findings for small $\Delta \in \{0, 1, \ldots, 4\}$ in Section 3. Intriguingly, most of the empirical findings hold similarly to small Δ, which demonstrates the robustness of our insights.



(a) Average attention score matrix of L0H7.
(b) Average attention score matrix of L1H11.
(c) Average attention score matrix of L2H6.

Figure 14. Average attention score matrices of SDHs with *large* Δ in Qwen2.5-7B-Instruct with prompts whose tokens are i.i.d. samples from the uniform distribution over the alphabet.

Definition of SDHs with Large Δ. Definition 3.1 identifies slash patterns in attention score matrices by examining the average slash scores. When κ is chosen sufficiently large relative to the context length (e.g., $\kappa = 0.1$ for a context length of 6000 in Section 3), this definition reliably detects slash patterns with small Δ. As illustrated in Figure. 1(d)–(f), however, there also exist slash patterns with large Δ. These long-range slash patterns have much smaller average slash scores, on the order of 10^{-3}. However, directly setting $\kappa = 10^{-3}$ leads to many spurious detections. For example, it would label offsets $\Delta < 10$ as slash-dominant for almost all heads. This high false-positive rate arises from the locality of natural language, which induces irregularly high attention values to nearby tokens rather than consistent slash patterns. To mitigate this locality effect, we restrict the offset range to $500 \leq \Delta \leq 5000$ and set $\kappa = 10^{-3}$. The goal of this section is to verify that the properties of SDHs with small Δ generalize to a *subset* of SDHs with large Δ. Empirically, with the hyperparameters $500 \leq \Delta \leq 5000$ and $\kappa = 10^{-3}$, we identify SDHs (with large Δ) only in Llama3-8B-Instruct and Qwen2.5-7B-Instruct since the context length of Gemma-7B is insufficient for this setting. Accordingly, we report experimental results only for these two models in this section. The detailed lists of selected heads are provided in Appendix H. A refined definition and a

more comprehensive investigation of SDHs with large Δ are left to future work.

OOD Generalization of SDHs. Similar to Section 3.1, we will show that SDHs with large Δ is OOD generalizable and intrinsic to the models. We test OOD generalization by evaluating the average slash scores under a special OOD prompt distribution $\mathcal{D}'$, in which every token is sampled independently from a uniform alphabet. We plot the average attention score matrices of identified SDHs under OOD prompts in Figure 14. These SDHs are from Qwen2.5-7B-Instruct and include L0H7, L1H11 and L2H6 for $\Delta = 937, 804, 1934$, which are the same SDHs reported in Figure 1. Figure 14 shows that the same slash pattern persists under OOD prompts. Same results are also observed for Llama3-8B-Instruct (see Appendix H).

We further quantify OOD generalization by computing the average slash scores for the identified SDHs with in-distribution and OOD prompts, and comparing the ratios. The resulting box plot of these ratios is shown in Figure 15. The average slash scores under OOD prompts are generally higher, or at least comparable to those obtained from in-distribution prompts. Detailed head-level results for the OOD case are provided in Table 6 in Appendix H.

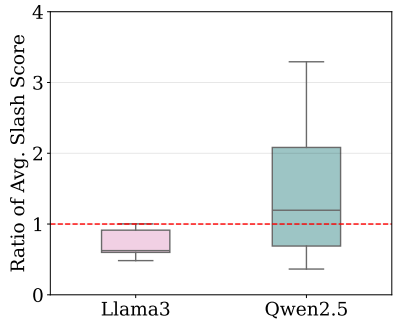
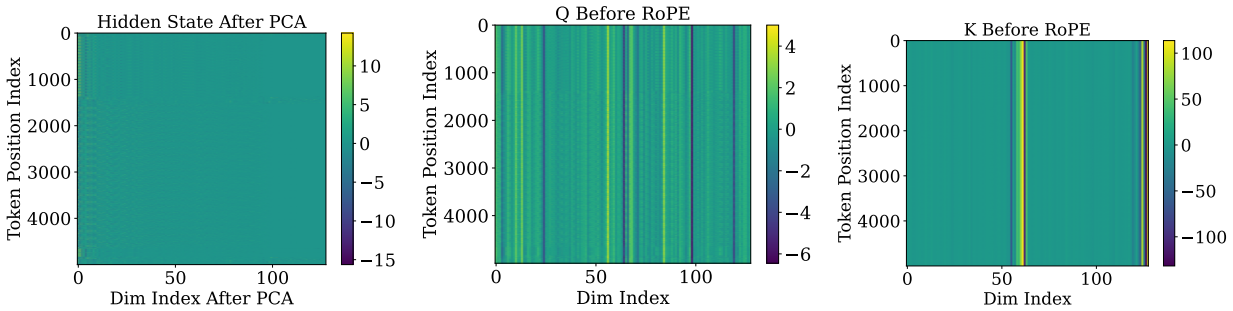


Figure 15. Ratio of average slash scores with OOD and in-distribution prompts (large Δ).

Hence, our Takeaway 1 generalizes to large Δ. The emergence of the slash-dominance pattern is not relevant to the semantic meaning of the prompts, but is intrinsic to the model architecture.



(a) Hidden state H of L0H7 after PCA. (b) Queries Q of L0H7. (c) Keys K of L0H7.

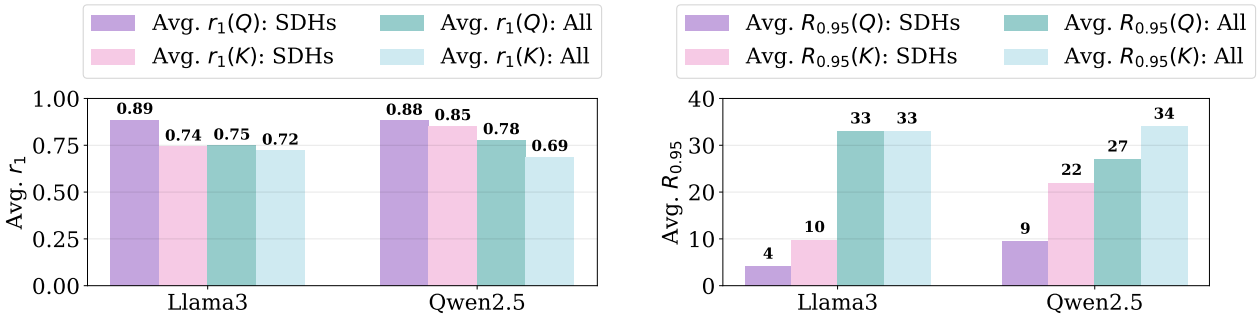
Figure 16. These figures show the queries and keys before the RoPE implementation, as well as the hidden states for SDHs with large Δ in Qwen2.5-7B-Instruct. The queries and keys each have dimension 128, and the hidden states are reduced to 128 dimensions for demonstration.

Approximate Low-Rankness of pre-PE Queries and Keys. Similar to Section 3.2, we show the approximate low-rankness of pre-PE queries and keys. We first visualize a SDH of Qwen2.5-7B-Instruct L0H7 in Figure 16. This SDH demonstrates slash dominance across multiple large offsets, all with $\Delta \geq 900$. As shown in Figure 16 (b) and (c), the pre-PE queries and keys are highly similar across tokens, implying that the query and key matrices are almost low-rank, particularly rank-one. Additional visualizations for more heads and models are provided in Figure 30 in Appendix H, which consistently exhibit the same structural pattern.

We then quantify the low-rankness of the pre-PE queries and keys in detail by r_1 and $Q_{0.95}$ according to (7) in Figure 17. In particular, Figure 17(a) shows that, for each model, at least one of $r_1(Q)$ and $r_1(K)$ strictly exceeds 0.88 on average over SDHs only, and these values are strictly lower on the other heads. Figure 17(b) shows that the effective ranks of Q and K are low only on the SDHs, while these matrices on the other heads have much higher ranks.

Hence, our Takeaway 2 generalizes to the large Δ regime. The pre-PE queries Q and keys K all have low effective ranks. Moreover, at least one of Q and K is close to being rank-one.

How to Get Approximately Rank-One Queries and Keys? Similar to Section B, we next answer how SDHs achieve the approximately rank-one queries and keys. As in Figures 18, we compute the average values of r_1 and $R_{0.95}$ for H, W_Q



(a) Average $r_1(\uparrow)$ of all heads and SDHs with large Δ.

(b) Average $R_{0.95}(\downarrow)$ of all heads and SDHs with large Δ.

Figure 17. Comparison of average r_1, $R_{0.95}$ of SDHs with those of all heads. For large Δ, at least one of the Q- or K exhibits *substantially lower average ranks* for SDHs compared to all heads.

(resp. W_K), and Q (resp. K) and average these values over the identified SDHs. Figures 18 clearly shows that the effective ranks of H, W_Q, and W_K are much higher than those of Q and K. Thus, low-rankness of Q and K must arise from the interaction between the hidden states H and weight matrices W_Q and W_K.

We further plot the average values of $\tilde{r}_1$ and $\tilde{R}_{0.95}$, and compare them with r_1 and $R_{0.95}$ respectively in Figures 19. These average values are computed over all the tokens and the SDHs in Layer 0. Figures 19 shows that the behaviors of $\tilde{r}_1(Q)$ and $\tilde{R}_{0.95}(Q)$ closely match those of $r_1(Q)$ and $R_{0.95}(Q)$ in Llama3-8B-Instruct, but differ in Qwen2.5-7B-Instruct. This confirms that in Gemma-7B and Llama3-8B-Instruct, each token embedding concentrates most of its energy approximately in a one-dimensional principal subspace of W_Q and W_K. That is, each token embedding aligns well with low-rank (and almost rank-one) subspaces of W_K and W_Q. Hence, our Takeaway 3 generalizes to large Δ.

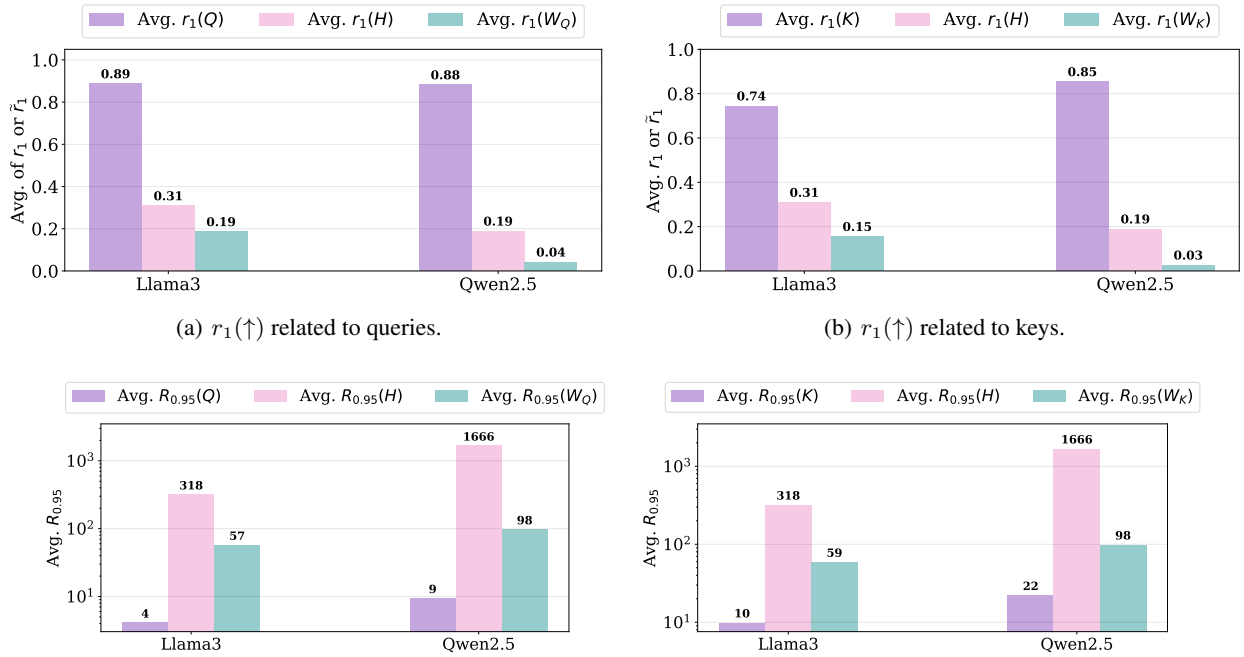
Collaboration of Frequencies in RoPE Determines Slash Pattern. Similar to Section 3.3, given that pre-PE queries and keys are approximately rank-one, we proceed to examine how the different RoPE frequencies collectively contributed to the slash pattern according to (8). We compute the relative variation of the norms of the query and key vectors and plot their averages over tokens and SDHs in Figure 21. As shown in the figure, the relative variation is at most 0.093, indicating that the norms of $\mathbf{q}_i$ (and similarly $\mathbf{k}_j$) are nearly constant. As a result, $\mathbf{q}_i$ is approximately the same for all i, and similarly for $\mathbf{k}_j$. This implies that the amplitudes A_ℓ and initial phases φ_ℓ in (9) are approximately identical across token pairs, which mirrors the behavior observed in the small-Δ regime and explains how RoPE gives rise to SDHs in the large Δ setting.

To further characterize the collaboration of frequencies, we then visualize $\mathrm{InP}(i, j, \ell)$ with i fixed to $i = 5000$ and j varying within $[1, 5000]$ in Figure 20, based on Qwen2.5-7B-Instruct. We observe that high- and medium-frequency components ($\ell \in [1, 42]$) exhibit greater variation than low-frequency components ($\ell \in [43, 64]$). For a more quantitative characterization, we evaluate slash-dominance after selectively removing certain frequencies when computing the attention score using the same criteria as the small Δ regime. We compute the ratios of the new score to the original score for each identified SDHs in Llama3-8B-Instruct and Qwen2.5-7B-Instruct, and show the box plot in Figure 22. We observe that, when removing high frequencies, the average slash score decreases in both models. Removing the medium frequencies leads to a mild change, or even no decrease, while the low frequencies yield the least drop.

Hence, our Takeaway 4 generalizes to large Δ. The high- and medium-frequency components in RoPE are more important than the low-frequency components for SDHs.

D. Theoretical Study of Shallow Transformers

In this section, we provide theoretical support for the empirical findings in the small Δ regime in Section 3. We show that under two conditions similar to those identified by experiments (Takeaways 2 to 4): (i) token embeddings lie approximately on a cone and (ii) RoPE is dominated by medium- and high-frequency components, SDHs provably emerge. Moreover, the learned SDHs are provably OOD generalizable (Takeaway 1). In particular, we focus on a simple but fundamental



(a) $r_1(\uparrow)$ related to queries.

(b) $r_1(\uparrow)$ related to keys.

(c) $R_{0.95}(\downarrow)$ related to queries.

(d) $R_{0.95}(\downarrow)$ related to keys.

Figure 18. Comparison of average r_1 and $R_{0.95}$ for query/key matrices relative to the hidden states and their corresponding weight matrices, for SDHs with *large* Δ. Both hidden states H and the weight matrices W_Q, W_K exhibit much higher rank than the resulting pre-PE queries and keys, indicating that the low-rankness arises from the interaction between H and the weights W_Q, W_K.

ICL setting, and train a shallow transformer with RoPE via Gradient Descent (GD). We show that the transformer learns a two-layer induction-head circuit, and the first-layer attention head is an SDH.

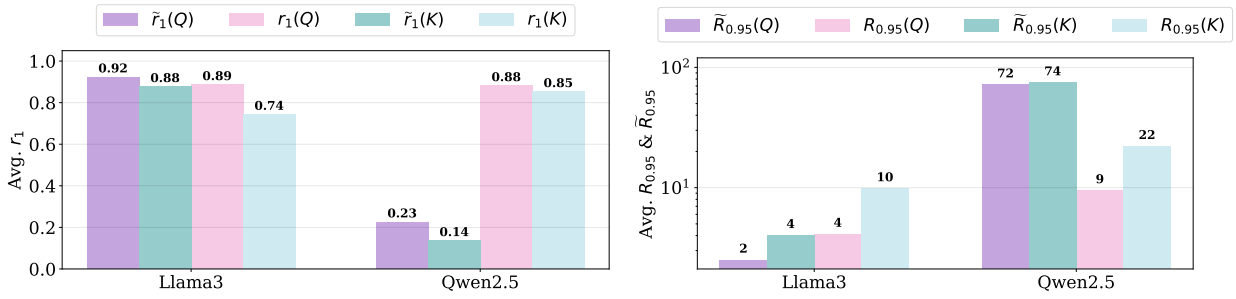
The theoretical results are organized as follows. We present the ICL data model in Section D.1 and then propose a slash-dominance frequency condition that quantitatively characterizes the RoPE frequency interactions in Section D.2. It can be viewed as a mathematically rigorous formulation of Takeaway 4. We present the transformer architecture and training algorithm in Section D.4 and D.5. Finally, we present the main theoretical results in Section D.6.

D.1. Data Model

We consider a regression task under the ICL framework, where the SDH plays an important role. Such a setting is commonly adopted in theoretical studies of ICL and induction heads (Zhang et al., 2024a; Huang et al., 2024; Chen et al., 2024a; Yang et al., 2024; Zhang et al., 2024b). The goal of in-context regression is to correctly learn a predictor $\hat{y}_{\mathrm{q}} := \widehat{y}(\mathbf{x}_{\mathrm{q}})$ that approximates the true response $y = f(\mathbf{x}_{\mathrm{q}})$ for a question $\mathbf{x}_{\mathrm{q}} \sim \mathcal{D}_{\mathcal{X}}$ and a target function $f \sim \mathcal{D}_{\mathcal{F}}$, based on a prompt P containing demonstration examples with the same function f. Specifically, the data are sampled as follows. We first draw a function $f \sim \mathcal{D}_{\mathcal{F}}$. Next, we independently sample a collection of N_{in} inputs $\mathbf{x}_1, \ldots, \mathbf{x}_{N_{\mathrm{in}}}$ together with a question $\mathbf{x}_{\mathrm{q}}$, all from $\mathcal{D}_{\mathcal{X}}$. For each $\mathbf{x}_i$, we define the label as $y_i = f(\mathbf{x}_i)$. Then we construct the prompt $P = (\mathbf{x}_1, y_1, \cdots, \mathbf{x}_{N_{\mathrm{in}}}, y_{N_{\mathrm{in}}}, \mathbf{x}_{\mathrm{q}})$, which has a fixed length of $N = 2N_{\mathrm{in}} + 1$.

Task Distribution. We consider a specialized *linear regression* family: $\mathcal{F} = \left\{ f : \mathcal{X} \to \mathbb{R} \mid f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle \text{ with } \mathbf{w} \in \mathbb{R}^{d_{\mathcal{X}}}, \|\mathbf{w}\|_2 \leq \sqrt{d_{\mathcal{X}}}, \mathcal{X} \subseteq \mathbb{R}^{d_{\mathcal{X}}} \right\}$, which is widely adopted in recent studies for in-context learning (Zhang et al., 2024a; Huang et al., 2024). As a result, the distribution $\mathcal{D}_{\mathcal{F}}$ is induced by the distribution of the random weight vector $\mathbf{w}$, denoted by $\mathcal{D}_{\Omega}$. For simplicity, we consider a normalized case where $\mathbb{E}[\mathbf{w}] = 0$ and $\mathrm{Var}[\mathbf{w}] = I_{d_{\mathcal{X}}}$.

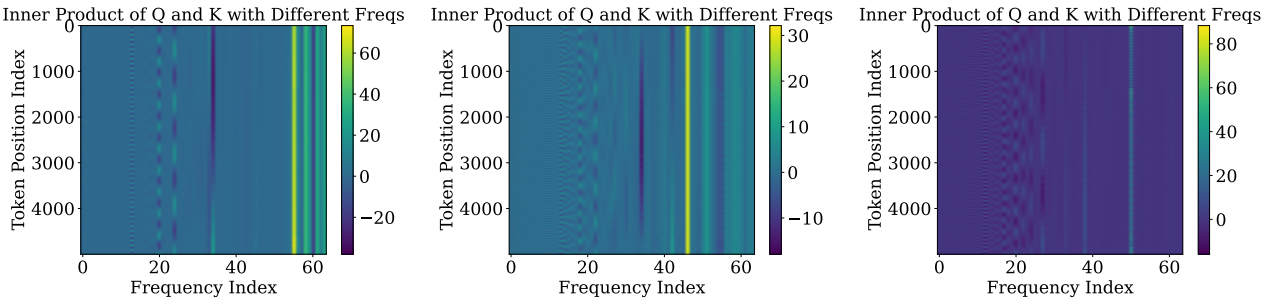
Data Distribution. We consider the *feature learning* setting, where $\mathcal{X} = \{\mathbf{v}_1, \ldots, \mathbf{v}_K\} \subseteq \mathbb{R}^{d_{\mathcal{X}}}$ are a finite set of K orthogonal and normalized features. Here, with slight abuse of notation, we use K to denote the number of features. Each data point $\mathbf{x}$ is sampled from $\mathcal{X}$ with the probability p_k for sampling $\mathbf{v}_k$, where $p_k \in (0, 1)$ for $k \in [K]$ and $\sum_{k \in [K]} p_k = 1$. Such a data model has been widely employed in the theoretical studies of deep learning, including ensemble methods (Allen-



(a) $\tilde{r}_1(\uparrow)$ and $r_1(\uparrow)$ related to queries.

(b) $R_{0.95}(\downarrow)$ and $\widetilde{R}_{0.95}(\downarrow)$ related to queries.

Figure 19. Comparison of average r_1, $R_{0.95}$, $\tilde{r}_1$, and $\tilde{R}_{0.95}$ for query/key matrices relative to the 0-th layer hidden states (token embeddings) for SDHs with *large* Δ in Gemma-7B, Llama3-8B-Instruct, and Qwen2.5-7B-Instruct.



(a) $\mathrm{InP}(5000, j, l)$ of L0H7.

(b) $\mathrm{InP}(5000, j, l)$ of L1H11.

(c) $\mathrm{InP}(5000, j, l)$ of L2H6.

Figure 20. Heatmaps of $\mathrm{InP}(i, j, \ell)$ on various SDHs of Qwen2.5-7B-Instruct. where we set $i = 5000$ and let j vary. Qwen2.5-7B-Instruct has 64 frequencies ($1 \leq \ell \leq 64$). The variation within each column shows the influence of each frequency. Qualitatively, high- and medium-frequency components $\ell \in [1, 42]$ exhibit greater variation.

Zhu & Li, 2023), and in-context learning (Huang et al., 2024). For simplicity, we consider the balanced case where $p_k = 1/K$ for all $k \in [K]$. The extension to the imbalanced case, where p_k takes non-uniform values, can be derived similarly to Huang et al. (2024).

D.2. Slash-Dominance Frequency Condition

From observation in Section 3.3, we find that RoPE and, especially, its frequencies play a crucial role in the slash-dominant pattern. We introduce a quantitative slash-dominance frequency condition that precisely characterizes frequency interaction in this section. Because RoPE operations depend on the token embeddings, we first introduce the token embedding of prompts as follows.

Token Embedding. For each $i \in [N_{\mathrm{in}}]$, we consider embedding information of input $\mathbf{x}_i$ at position $2i - 1$ and its label y_i at position $2i$ into orthogonal subspaces. Concretely, we consider $E_{2i-1} = [\mathbf{c}^\top \ \mathbf{x}_i^\top \ 0 \ 0]^\top$, and $E_{2i} = [\mathbf{c}^\top \ 0 \ 1 \ y_i]^\top \in \mathbb{R}^d$, where $\mathbf{c} \in \mathbb{R}^{d_\mathbf{c}}$, $\|\mathbf{c}\|_2 = 1$ represents the semantically independent information and is identical across all token embeddings, and $d = d_\mathbf{c} + d_\mathcal{X} + 2$. In addition, $E_{\mathrm{q}} = [\mathbf{c}^\top \ \mathbf{x}_{\mathrm{q}}^\top \ 0 \ 0]^\top$. As a result, the token embedding E is defined as follows.

$$E = E(P) = \begin{bmatrix} E_1 & \cdots & E_{2N_{\mathrm{in}}} & E_{\mathrm{q}} \end{bmatrix}^\top \in \mathbb{R}^{N \times d}.$$

In the definition above, all token embeddings lie on a cone whose axis is given by $[\mathbf{c}^\top \ 0 \ 0 \ 0]^\top$. This is motivated by Takeaway 3, and to further distinguish the semantically independent information (the cone axis) from the semantically dependent information, we embed the semantically dependent and independent components in orthogonal coordinate subspaces. For brevity, we further denote the semantically dependent subspace $E^{\mathbf{x},y} = E_{:,d_\mathbf{c}+1:d} \in \mathbb{R}^{N \times (d_\mathcal{X}+2)}$ and $E^y = E_{:,d} \in \mathbb{R}^N$, which are the last $d_\mathcal{X} + 2$ columns and the last column of E, respectively.

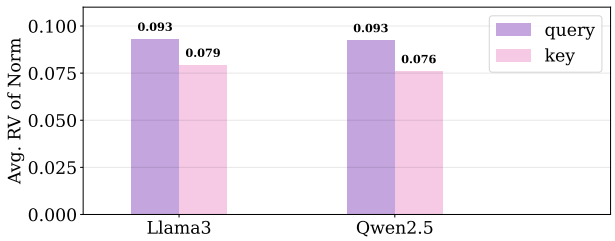


Figure 21. Average of the relative variation of the norm of the query vector $\|\mathbf{q}_i\|$ and key $\|\mathbf{k}_i\|$ for large Δ in Llama3-8B-Instruct and Qwen2.5-7B-Instruct.

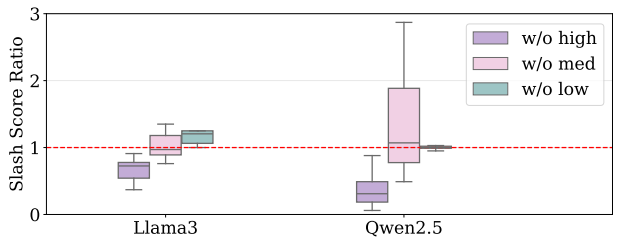


Figure 22. The figure quantifies the effect of low-, medium-, and high-frequency components on SDHs by reporting, for each band, the ratio of the average slash score after removing that band to the original average score.

Note that RoPE has $d/2$ frequencies. In the following, for theoretical analysis, we restate the Theorem 4.1 for the cone component (the first $d_{\mathbf{c}}$ dimensions) and the semantic component (the last $d - d_{\mathbf{c}}$ dimensions), respectively.

D.3. A sufficient Slash-Dominance Frequency of RoPE

Assumption D.1 (Restatement of Slash-Dominance Frequency sequence)**.** *Let N be the fixed length of the prompt P and d be the hidden dimension of embeddings. Denote the Kronecker delta function by $\delta_0(x) := \mathbb{1}\{x = 0\}$. We assume that frequency sequences $\boldsymbol{\vartheta} = (\theta_1, \cdots, \theta_{d/2})$ satisfy the following:*

1. ***The last** $(d - d_{\mathbf{c}})/2$ **low frequencies are small enough.** For any $d_{\mathbf{c}}/2 + 1 \leq s \leq d/2$, $\theta_s \leq O(N^{-\alpha})$ is a low frequency with $\alpha \geq 2$.*

2. ***Sinusoal components of the first** $d_{\mathbf{c}}/2$ **frequencies approximate the pulse.** There exists constants $L, C_1 \geq 0, C_2 \in \mathbb{R}$, and a small enough noise tolerance $\epsilon_{\text{FN}} \leq LC_1/N$ such that for any $x \in \mathbb{Z}$ and $|x| \leq N$, we have*

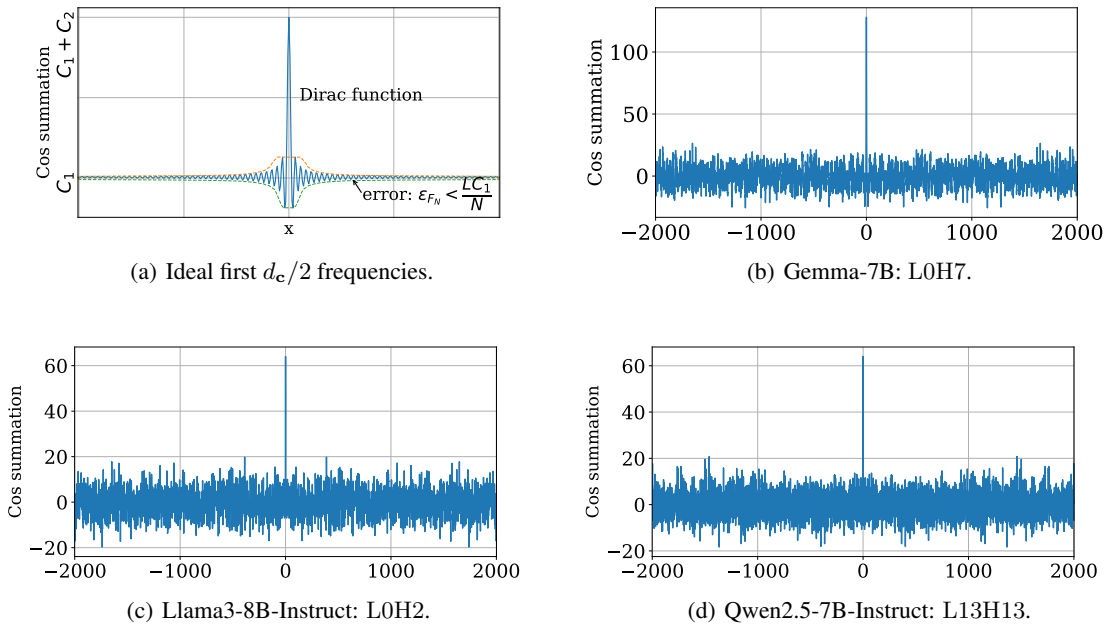
$$\Big|\sum_{s=1}^{d_{\mathbf{c}}/2} \cos(\theta_s x) - C_1 \cdot \delta_0(x) - C_2\Big| \leq \epsilon_{\text{FN}}. \tag{12}$$

We assume that the *last* $(d - d_{\mathbf{c}})/2$ *frequencies*, corresponding to semantically dependent content $\mathbf{x}, y$, are all low. This is aligned with the observation in (Barbero et al., 2025) that low frequencies act primarily on semantic content. Conversely, the *first* $d_{\mathbf{c}}/2$ *frequencies* correspond to the semantically independent content $\mathbf{c}$. This assumption states that if the selected frequency components have identical amplitudes 1, then the resulting sum of sinusoidal components approximates a pulse, which is represented by the Kronecker delta function $\delta_0(x)$ as in (12). Here, the uniform amplitudes are intrinsic to the Fourier expansion of the approximation target $\delta_0(x)$, which distributes equal amplitude across all basis frequencies.

After training, however, the relative contributions of frequencies need not remain uniform, as their effective amplitudes are determined by the learned weight matrices and token embeddings. Assumption D.1 further verifies that not all frequencies in RoPE are required for the SDHs. In particular, the low frequencies do not affect the validity of Assumption D.1, consistent with Takeaway 4. On the contrary, if medium-frequencies or high-frequencies are set to zero, Assumption D.1 may be violated, and slash-dominance no longer holds. As shown in Figure 8, setting part of the medium- or high-frequencies to zero substantially reduces the average slash scores.

Empirical Validation of Assumption D.1. We remark that the slash-dominance frequency condition is satisfied by pretrained models. Specifically, we first identify the *active frequencies*, defined as those whose average $\text{InP}(i, j, l)$ exceeds one-tenth of the maximum. These active frequencies contribute primarily to the attention scores. We plot their sinusoidal sum given on the left-hand side of (12). As shown in Figure 23, the active frequencies employed in open-source LLMs (e.g., Gemma-7B, Llama3-8B-Instruct, Qwen2.5-7B-Instruct) satisfy this condition.

Takeaway 5: Assumption D.1 is empirically validated and serves as a sufficient condition on the RoPE frequencies for inducing SDHs.



(a) Ideal first $d_{\mathbf{c}}/2$ frequencies.

(b) Gemma-7B: L0H7.

(c) Llama3-8B-Instruct: L0H2.

(d) Qwen2.5-7B-Instruct: L13H13.

Figure 23. Summed cosine functions of frequencies. Panel (a) is the ideal case using the first $d_{\mathbf{c}}/2$ frequencies in Assumption D.1; Panel (b)–(d) use the *active* frequencies in practice (defined as those with average $\mathrm{InP}(i,j,l)$ greater than one-tenth of the largest $\mathrm{InP}(i,j,l)$). Results (b)–(d) for Gemma-7B, Llama3-8B-Instruct, and Qwen2.5-7B-Instruct show a pattern consistent with (a).

D.4. Network Architecture

We next introduce the network structure, which takes the token embedding E as input and outputs the desirable prediction $\hat{y}_{\mathrm{q}}$.

Reduced Two-layer single head Disentangled Transformer. We consider a *two-layer single head transformer*, and for simplicity in analyzing the residual stream and output of each CSA layer separately, we adopt the *disentangled transformer* (see Definition D.2 and (13)), which feeds the concatenation, rather than the sum, of the residual and CSA layer output as the input into the subsequent layer. Notably, the *disentangled transformer* is equivalent to a standard decoder-based, attention-only transformer (Nichani et al., 2024), and thus its use entails no loss of generality.

Definition D.2 (Two-layer Disentangled Transformer)**.** *Let $E \in \mathbb{R}^{N\times d_0}$ with $d_0 = d$ be the input token embedding and $\boldsymbol{\vartheta}^{(0)} = \boldsymbol{\vartheta}$ denote the initial RoPE frequency sequence. Define the hidden dimensions of two subsequent layers as $d_1 = 2d_0$ and $d_2 = 2d_1$. For each layer $\ell = 1, 2$, let $W^{(\ell)}_{\{Q,K,V\}} := \{W^{(\ell)}_Q, W^{(\ell)}_K, W^{(\ell)}_V\} \subset \mathbb{R}^{d_{\ell-1}\times d_{\ell-1}}$ be weight matrices, and let $W_O \in \mathbb{R}^{d_2\times d_{\mathrm{out}}}$ be the output matrix. Set all parameters $\theta = \big\{W^{(\ell)}_{\{Q,K,V\}}\big\}_{\ell=1}^{2} \bigcup \{W_O\}$. The frequency sequence is updated at each layer by $\boldsymbol{\vartheta}^{(\ell)} = (\boldsymbol{\vartheta}^{(\ell-1)}, \boldsymbol{\vartheta}^{(\ell-1)})$. Then the hidden states $H^{(\ell)} \in \mathbb{R}^{N\times d_\ell}$ at each layer $\ell = 0, 1, 2$ and the final output of TF_θ are defined as follows:*

$$
\begin{aligned}
&H^{(0)} = E,\\
&H^{(\ell)} = \big[H^{(\ell-1)},\, \mathrm{CSA}\big(H^{(\ell-1)}; W^{(\ell)}_{\{Q,K,V\}}, \boldsymbol{\vartheta}^{(\ell)}\big)\big], \quad \ell = 1, 2\\
&\mathrm{TF}_\theta(E) = H^{(2)} W_O.
\end{aligned}
$$

We abbreviate $\mathrm{CSA}\big(H^{(\ell-1)}; W^{(\ell)}_{\{Q,K,V\}}, \boldsymbol{\vartheta}^{(\ell)}\big)$ as $\mathrm{CSA}\big(H^{(\ell-1)}\big)$ whenever the context is clear. Moreover, as in (4), we denote Layers 1 and 2 attention scores by $S^{(1)}(E;\theta)$ and $S^{(2)}(E;\theta)$, respectively.

Before formally giving the expression of the transformer output, we first simplify the weight matrices $\theta = \big\{W^{(\ell)}_{\{Q,K,V\}}\big\}_{\ell=1,2} \bigcup \{W_O\}$ for ease of analysis. A visualization of the simplification is provided in Figures 34 and 35 in Appendix I.

First, for $\ell = 1, 2$, we reduce $W_{K,Q}^{(\ell)}$ to sparse matrices as follows.

$$W_Q^{(1)} = \begin{bmatrix} \widetilde{W}_Q^{(1)} & 0_{d_{\mathbf{c}} \times (d_{\mathcal{X}}+2)} \\ 0_{(d_{\mathcal{X}}+2) \times d_{\mathbf{c}}} & 0_{(d_{\mathcal{X}}+2) \times (d_{\mathcal{X}}+2)} \end{bmatrix},$$
$$W_Q^{(2)} = \left[\begin{array}{cc|c} 0_{d_{\mathbf{c}} \times d_{\mathbf{c}}} & 0_{d_{\mathbf{c}} \times (d_{\mathcal{X}}+2)} & 0_{d \times d} \\ 0_{(d_{\mathcal{X}}+2) \times d_{\mathbf{c}}} & \widetilde{W}_Q^{(2)} & \\ \hline \multicolumn{2}{c|}{0_{d \times d}} & 0_{d \times d} \end{array}\right], \tag{13}$$

where the block $\widetilde{W}_Q^{(1)} \in \mathbb{R}^{d_{\mathbf{c}} \times d_{\mathbf{c}}}$ is the only non-zero block in $W_Q^{(1)}$. This reduction follows from Takeaway 2 and 3. Concretely, in this simplified model, query Q exhibits almost rank-one and lies within the subspace generated by the cone axis $[\mathbf{c}^\top \ 0 \ 0 \ 0]^\top$. Consequently, the first-layer weights need only take nonzero values on the corresponding subspace to map the token embeddings distributed on a cone onto the cone axis. In addition, the block $\widetilde{W}_Q^{(2)} \in \mathbb{R}^{(d_{\mathcal{X}}+2) \times (d_{\mathcal{X}}+2)}$ is the only non-zero block in $W_Q^{(2)}$. The sparsity of $W_Q^{(2)}$ follows from the structure of the disentangled transformer, and the second layer focuses on the semantically dependent part of the input $H^{(1)}$.

In addition, since only the relative correlation $W_Q^{(\ell)}$ and $W_K^{(\ell)}$ matters when calculating attention scores, we only make $W_Q^{(\ell)}$ *trainable* while keeping $W_K^{(\ell)}$ *fixed* during training as follows.

$$W_K^{(1)} = \begin{bmatrix} \widetilde{W}_K^{(1)} & 0_{d_{\mathbf{c}} \times (d_{\mathcal{X}}+2)} \\ 0_{(d_{\mathcal{X}}+2) \times d_{\mathbf{c}}} & 0_{(d_{\mathcal{X}}+2) \times (d_{\mathcal{X}}+2)} \end{bmatrix},$$
$$W_K^{(2)} = \begin{bmatrix} 0_{d \times d} & 0_{d \times d} \\ I_d & 0_{d \times d} \end{bmatrix}, \tag{14}$$

where we assume that $\widetilde{W}_K^{(1)}$ is fixed and satisfies $(\widetilde{W}_K^{(1)})^\top \mathbf{c} = (1, 0, \cdots, 1, 0)^\top := \tilde{\mathbf{c}} \in \mathbb{R}^{d_{\mathbf{c}}}$.
Finally, we specify the value and output matrices as

$$W_V^{(1)} = I_d, W_V^{(2)} = I_{2d}, W_O = [0_{d \times d} \ \ 0_{d \times d} \ \ I_d \ \ 0_{d \times d}]^\top. \tag{15}$$

As a result, the trainable weights reduce to $\tilde{\theta} = \{\widetilde{W}_Q^{(1)}, \widetilde{W}_Q^{(2)}\}$.

Output of the Reduced Model. With reduced trainable $\tilde{\theta}$ defined above, and since we are only interested in the predicting $\mathbf{x}_{\mathrm{q}}$ at position N, the predictor is given by $\hat{y}_{\mathrm{q}}(E; \tilde{\theta}) = \mathrm{TF}_{\tilde{\theta}}(E)_{(N,d)}$. The explicit form of this predictor is obtained by plugging in the fixed parameters in (14) and (15), and is provided in (22) in Appendix I.

D.5. Training Settings

Loss Function. We choose the standard squared loss used in a linear regression task as follows.

$$L(\tilde{\theta}) = \frac{1}{2} \mathbb{E}_{\mathbf{w}, P} \left[(\hat{y}_{\mathrm{q}}(E(P); \tilde{\theta}) - \langle \mathbf{w}, \mathbf{x}_{\mathrm{q}} \rangle)^2 \right], \tag{16}$$

where the expectation is taken with respect to the joint distribution of the task $\mathbf{w}$ and the prompt P.

Training Algorithm. As shown in Algorithm 1, we train the reduced two-layer disentangled transformer with two-stage gradient descent on the squared loss. This two-stage procedure is motivated by the empirical observation that different layers tend to converge in distinct phases (Bietti et al., 2023). The reduced weights are initialized as $\widetilde{W}_Q^{(1)} = 0_{d \times d}$, $\widetilde{W}_Q^{(2)} = I_d$. In the first stage, $\widetilde{W}_Q^{(2)}(0) = I_d$ is fixed, and gradient descent operates on $\widetilde{W}_Q^{(1)}(0) = I_d$ with learning rate η_1 until timestep $\tau_1 + 1$, and outputs $\widetilde{W}_Q^{(1)}(\tau_1 + 1)$. Then, in the second stage, the weight $\widetilde{W}_Q^{(1)}(\tau_1 + 1)$ is fixed, and gradient descent operates on $\widetilde{W}_Q^{(2)}(0) = I_d$ with learning rate η_2 until timestep $\tau_1 + \tau_2 + 1$.

Algorithm 1 Two-stage Training Algorithm

Input: learning rate $\eta_1, \eta_2 > 0$; iteration τ_1, τ_2.
Initialize $\widetilde{W}_Q^{(1)}(0) = 0_{d\times d}$, $\widetilde{W}_Q^{(2)}(0) = I_d$.
// Stage I
for $t = 1$ **to** $\tau_1 + 1$ **do**
  $\widetilde{W}_Q^{(1)}(t) = \widetilde{W}_Q^{(1)}(t-1) - \eta_1 \nabla_{\widetilde{W}_Q^{(1)}} L(\tilde{\theta}^{(t-1)})$.
  $\tilde{\theta}(t) = (\widetilde{W}_Q^{(1)}(t), \widetilde{W}_Q^{(2)}(0))$.
end for
// Stage II
for $t = \tau_1 + 1$ **to** $\tau_1 + \tau_2 + 1$ **do**
  $\widetilde{W}_Q^{(2)}(t) = \widetilde{W}_Q^{(2)}(t-1) - \eta_2 \nabla_{\widetilde{W}_Q^{(2)}} L(\tilde{\theta}^{(t-1)})$.
  $\tilde{\theta}(t) = (\widetilde{W}_Q^{(1)}(\tau_1 + 1), \widetilde{W}_Q^{(2)}(t))$.
end for
Output: $\hat{\theta} = \tilde{\theta}(\tau_1 + \tau_2 + 1)$.

D.6. Main Theoretical Result

In this section, we characterize the convergence of Algorithm 1, and analyze its performance in Stages I and II. Our theoretical regression setting follows a general induction head circuit mechanism as introduced in Figure 9 in Section A. To output an accurate prediction, (i) after Stage I, the first layer focuses on the prefix positions and thereby forms an SDH, enabling each y_i to integrate information from its associated prefix $\mathbf{x}_i$; (ii) After Stage II, the learned second layer identifies the y_i whose prefix matches the question token, i.e., $\mathbf{x}_i = \mathbf{x}_{\mathrm{q}}$, and outputs the corresponding value.

To formalize these behaviors, we introduce notations for tracking the attention scores during training. Under the reduced model with parameters $\tilde{\theta}$, we denote the first layer attention score from the position i to the position j by $S_{i,j}^{(1)}(E;\tilde{\theta})$, and the second layer attention score from the question token to the position i by $S_i^{(2)}(E;\widetilde{\theta})$. To further characterize Layer 2 behavior, for any $k \in [K]$, we define the aggregated attention score from the question token to the feature k by

$$\mathcal{S}_k^{(2)}(E;\tilde{\theta}) := \sum\nolimits_{i\in\mathcal{V}_k(P)} S_{2i}^{(2)}(E;\tilde{\theta}), \tag{17}$$

where $\mathcal{V}_k := \mathcal{V}_k(P) \subset [N_{\mathrm{in}}]$ is the feature k index set, such that $\mathbf{x}_i = \mathbf{v}_k$ for $i \in \mathcal{V}_k(P)$. For simplicity, given the parameters $\widetilde{\theta}(t)$ at step t, we abbreviate $S_{i,j}^{(1)}(E;\widetilde{\theta}(t))$, $\mathcal{S}_k^{(2)}(E;\widetilde{\theta}(t))$ and $\hat{y}_{\mathrm{q}}(E;\widetilde{\theta}(t))$ as $S_{i,j}^{(1)}(t)$, $\mathcal{S}_k^{(2)}(t)$, and $\hat{y}_{\mathrm{q}}(t)$, respectively.

These quantities allow us to monitor how the network forms an SDH in Layer 1 and outputs an accurate prediction in Layer 2. Specifically, two attention patterns are desired: (i) In the first layer, an SDH requires $S_{i,i-1}^{(1)}(\tau_1+1) \approx 1$. (ii) In the second layer, when $\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k$, we similarly require $\mathcal{S}_k^{(2)}(\tau_1+\tau_2+1) \approx 1$. Under these conditions, the model outputs an accurate prediction:

$$\hat{y}_{\mathrm{q}}(\tau_1+\tau_2+1) = \sum\nolimits_{i\in[N_{\mathrm{in}}]} S_{2i}^{(2)}(\tau_1+\tau_2+1)\cdot y_i \approx \langle \mathbf{w}, \mathbf{v}_k\rangle. \tag{18}$$

We formally state our main theorem below, which characterizes the convergence of loss and the dynamics of these attention scores.

Theorem D.3 (Training Dynamics)**.** *Suppose Assumption D.1 holds. Consider $N \gg \mathrm{poly}(K) \gg \mathrm{polylog}(N)$, $N \gg d$, and $p_k = 1/K$ for any $k \in [K]$. We set $\epsilon_1 = O(N^{-\frac{1}{2}})$ and $\epsilon_2 = O(N^{-\frac{1}{4}}) \bigcap \Omega(\epsilon_1)$ as the attention concentration errors of the first and second layer, respectively. Then the following holds for the training of Algorithm 1,*

Stage I: SDH Emergence in First Layer. *In Stage I with τ_1 at most*

$$O\left(\frac{KN\log(N)}{\eta_1 C_1} + \frac{K\log(K^{-1}\epsilon_1^{-1})}{\epsilon_1\eta_1 C_1}\right),$$

the attention head of the first layer is trained to be $1-\epsilon_1$ slash-dominant at 1. Formally, for any token at position $i \in [N]$, it almost attends to the immediately previous token, i.e., $1-\epsilon_1 \le S_{i,i-1}^{(1)}(\tau_1+1)$.

Stage II: Feature Matching in Second Layer. *In Stage II with τ_2 at most*

$$O\left(\frac{K^2\log(K)}{\eta_2}+\frac{K\log(K\epsilon_2^{-1/2})}{\eta_2\epsilon_2}\right),$$

*the attention head of second layer has a **concentrated** attention score on the same feature with the question token. Formally, if $\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k$, then with high probability, the second layer almost attends to input tokens featuring $\mathbf{v}_k$: $(1-\mathcal{S}_k^{(2)}(\tau_1+\tau_2+1))^2\leq O(\epsilon_2)$.*

End: The Loss Convergence. *At the end of Stage II, the loss converges. Formally, $L(\hat{\theta})\leq\epsilon_2$.*

The above theorem shows that if the frequencies satisfy Assumptions D.1, the training of Algorithm 1 converges to the minimum of loss via GD, with polynomial time efficiency, and proceeds in two distinct stages: (i) The first layer captures positional dependencies and rapidly learns a slash-dominant attention pattern with $\Delta=1$, given a sequence of frequencies satisfying Assumption D.1, and (ii) The second layer captures semantic and feature dependencies and achieves feature matching. Notably, the error in the second layer, ϵ_2, is larger than that in the first layer, ϵ_1, because feature matching relies on accurate prefix recognition by the SDH formed in the first layer. Detailed proofs of Theorem D.3 are provided in Appendices J to L.

Extension to SDHs with General offsets Δ. Theorem D.3 focused on the specific case of small offset $\Delta=1$. This is because, in our prompt, each y_i is paired with its desired prefix $\mathbf{x}_i$ at a distance of 1. This setup requires the SDH to attend to the immediately preceding token. However, the proof is not restricted to this particular offset. In the general case, if the data were constructed such that the distance between y_i and $\mathbf{x}_i$ were an arbitrary Δ, the same reasoning would demonstrate the emergence of an SDH at offset Δ.

OOD Generalization. Given the learned model $\hat{\theta}$ and a test prompt for a linear task $\mathbf{w}$ (possibly outside the support of $\mathcal{D}_\Omega$), let the question token be $\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k$. By Stage II attention concentration, Eq. (18) shows that, with high probability, the question prediction $\hat{y}_{\mathrm{q}}$ is $\mathcal{S}_k^{(2)}(\tau_1+\tau_2+1)\langle\mathbf{w},\mathbf{v}_k\rangle+\sum_{m\neq k}\mathcal{S}_m^{(2)}(\tau_1+\tau_2+1)\langle\mathbf{w},\mathbf{v}_m\rangle\approx\langle\mathbf{w},\mathbf{v}_k\rangle$. This showcases the remarkable OOD generalization capability of SDHs.

> **Takeaway 6:** Theorem D.3 shows that, under the structural conditions in Takeaways 2-4 and the slash-dominance frequency condition Assumption D.1, a shallow transformer learns an SDH that are OOD generalizable to tasks.

We further remark that the convergence rate in Theorem D.3 does not depend on the two parameters α and $d_{\mathbf{c}}$ in Assumption D.1, as $\alpha\geq 2$ and $N\gg d\gg d_{\mathbf{c}}$, making the terms involving α and $d_{\mathbf{c}}$ a higher order term, and asymptotically negligible in the final convergence bound.

We finally comment on why GD performs well despite the nonconvex loss in (16). The key is that the optimization landscape is highly structured, containing *many* sub–global-optimal parameter configurations. In particular, for both Layer 1 and Layer 2, any configuration that produces sufficiently concentrated attention scores, as characterized in the two stages of Theorem D.3, already lies in such a sub–global-optimal region. Owing to the monotonicity and saturation properties of the softmax function, once a logit at a given position becomes sufficiently larger than the others, the attention scores concentrate almost entirely on that position. As a result, GD does not require a carefully tuned descent direction or precise learning-rate scheduling. The gradients naturally guide the parameters toward these large regions of concentrated-attention solutions. Indeed, our convergence rate bounds show that in both Stage I and Stage II, sufficiently large learning rates η_1 and η_2 move the parameters into their sub–global-optimal regions in essentially *one step*. In much deeper real-world LLMs, interactions across many layers introduce considerably more complex dynamics, and the behavior of GD may not be nearly as simple or predictable. Nevertheless, the emergence of SDHs is still expected.

E. Discussion of Potential Applications

In the preceding sections, we investigated the mechanistic interpretability of SDHs, showing that this phenomenon arises as an intrinsic algorithmic effect contributed by RoPE. In this discussion part, we first outline two potential directions where our findings could be applied to improve the efficiency and efficacy of the current model architectures. Then we discuss the implications of our results for the models with other kinds of PE.

Parameter Efficiency of Queries and Keys. Motivated by the observed *approximate low-rankness*, we suggest that the parameter matrices W_Q and W_K of some heads can be constrained to low rank. This could reduce the number of parameters and save computational resources during both training and inference, without significant loss in performance. We take an initial step to verify this. In Table 1 in Appendix G, we compress the query and key matrices of pretrained Qwen2.5-7B-Instruct and Llama3-8B-Instruct, which yields notable parameter reduction while preserving accuracy. A promising direction for future work is to explicitly enforce low-rank constraints during training to see whether we can achieve further efficiency gains.

Effective Length Generalization. Our result reveals that high- and medium-frequency components of RoPE predominantly contribute to slash-dominant behavior, whereas low frequencies play a negligible role. Based on this observation, one potential application is to edit or reweight the low-frequency components of RoPE to enhance length generalization in LLMs. We leave the systematic exploration of this idea to future experimental work.

Implications for Other Kinds of PEs. Our analysis focuses on models that use RoPE (e.g., the Gemma, Llama, and Qwen families). Other positional-encoding schemes, ALIBI (Press et al., 2021), NOPE (Kazemnejad et al., 2023), and sinusoidal positional embeddings (Vaswani et al., 2017), can exhibit behavior similar to or distinct from RoPE. For example, ALIBI adds a relative bias directly to query–key inner products; it is length-extrapolative by design and may therefore support OOD generalization. In contrast, NOPE (no explicit positional signal) and sinusoidal embeddings (often injected only at the input layer) may render SDHs more context dependent. A detailed study of SDHs under these alternatives is left to future work.

F. Experimental Details

In our experiments, we identify SDHs in Qwen2.5-7B-Instruct (Qwen Team, 2025), Llama3-8B-Instruct (Grattafiori et al., 2024), and Gemma-7B (Gemma Team, 2024). We evaluate two offset regimes that allow clean isolation of the slash pattern: (i) local offsets ($\Delta < 5$), which capture short-range structural effects, and (ii) extreme long-range offsets ($500 \leq \Delta \leq 5000$), where slash lines—if present—manifest as stable diagonal structures across large positional gaps. We intentionally exclude the intermediate offsets ($5 \leq \Delta < 500$), because attention in this range is dominated by semantic relatedness between nearby tokens. In this confounded regime, the slash-dominance score $S_{i,i-\Delta}$ reflects a multiplicative interaction between semantic dependencies and positional alignment, making it impossible to isolate the geometric slash pattern. By focusing on the short- and long-range extremes—where semantic contributions are either concentrated near the diagonal or negligible at large distances—we obtain measurements of slash-dominance that cleanly reflect positional structure.

To approximate the expectation in (6), we prefill each model with 500 diverse prompts from LongBenchV2 (Bai et al., 2025), which provides heterogeneous long-context distributions and yields stable cross-model statistics. Each prompt is truncated to 6000 tokens to ensure consistent context length across models and to avoid context-extension artifacts. We collect the prefill attention matrices, which deterministically reflect each model's attention geometry without introducing sampling noise. Following prior observations that early tokens (including BOS and warm-up positions) exhibit disproportionately large attention scores (Xiao et al., 2023), we exclude attention entries involving positions 1–4 from the computation of $S_{i,i-\Delta}$. Removing these anomalous early positions prevents them from dominating the slash-dominance statistics. After excluding these positions, each prompt still contributes at least 1000 valid pairs $(i, i-\Delta)$ for the Δ ranges we study. In total, we average over more than 5×10^5 valid samples.

For slash-dominance detection, we match the threshold κ to the natural scale of attention in each regime. For local offsets ($\Delta < 5$), we set $\kappa = 0.1$, reflecting the high concentration of attention near the diagonal. However, the number of SDHs for $\Delta = 0, 1, 2$ is quite large, so we only report the first three heads for each Δ in the tables in Appendix H. For extreme long-range offsets ($\Delta \geq 500$), we use $\kappa = 10^{-3}$. Although small in magnitude, this threshold corresponds to a 5× enrichment over the uniform baseline $1/5000 = 2 \times 10^{-4}$ at context length 5000, making it a meaningful marker of strong long-range alignment. Empirically, we do not find any SDHs with large Δ in Gemma-7B. Thus, we only report results for SDHs with large Δ in Llama3 and Qwen2.5.

For the experiments on OOD prompts, we construct prompts from i.i.d. tokens. Specifically, we independently sample each token from the uniform distribution over the model's vocabulary. For example, for Qwen2.5-7B-Instruct, the vocabulary size is 151,642, so each token is sampled with probability $1/151{,}642$. Using this procedure, we generate 500 prompts of length 6000 by sampling 500×6000 tokens in total.

Our experiments are conducted on NVIDIA A100 GPUs with 80G memory.

G. Parameter Compression Experiment

Motivated by the low-rankness of the queries and keys, we can compress the parameters of the W_Q and W_K matrices in the attention modules.

Experimental Setup. As we will compute the effective singular values/subspaces of the W_Q and W_K over LongBench v2 dataset (Bai et al., 2025) with heterogeneous prompts, we extend our power computation method in (11) to the hidden states H of a sequence of tokens, which is originally defined for a single token embedding $\mathbf{x}$.

To take the bias $\boldsymbol{b}$ of Qwen2.5-7B-Instruct model into account, for simplicity, we define $\sigma_0 = 1$ and $\mathbf{v}_0 = \boldsymbol{b}$. For other models without the bias term, we define $\sigma_0 = 1$ and $\mathbf{v}_0 = 0$. For each head, we first compute

$$r_l(H) = \frac{\|\sigma_l \mathbf{v}_l^\top H\|_2^2}{\sum_{i=0}^{d} \|\sigma_i \mathbf{v}_i^\top H\|_2^2} \tag{19}$$

for all $l \in \{0\} \cup [d]$, hidden states H of prompts in LongBench v2 dataset (Bai et al., 2025), where $\mathbf{v}_l$ is the singular vector of the weights, which is defined in (11). This quantity characterizes the power of this singular value σ_l for this prompt x. We then average the power over all the prompts to get the average power

$$r_l = \frac{\sum_H r_l(H)}{|\text{LongBench v2}|}. \tag{20}$$

Subsequently, we sort the singular values in the decreasing order of $\{r_l\}_{l\in[d]}$ to get $\{r_{\text{sorted}(l)}\}_{l\in[d]}$, where we force $r_{\text{sorted}(0)} = r_0$ to keep the bias term. We then calculate the effective rank R_{thre} such that the cumulative average power hits the power threshold:

$$R_{\text{thre}} = \min\Big\{l \,\Big|\, \sum_{i=0}^{l}\sum_{j=1}^{d} r_j \mathbb{1}\{\text{sorted}(j) = i\} > \text{thre}\Big\}. \tag{21}$$

For those W_Q and W_K weight matrices whose effective ranks are smaller than compression threshold $\text{rank}_{\text{thre}}$, we adopt its low rank approximations corresponding to the singular values $\{r_l | \text{sorted}(l) \le R_{\text{thre}}\}$; otherwise, we leave the weight matrices untouched. In the end, we obtain the low-rank model.

We evaluate low-rank models with different energy thresholds thre and compression limits $\text{rank}_{\text{thre}}$ on LongBench (Bai et al., 2024), which spans 21 benchmarks across 6 task types and consists of in total 4750 prompts. For each model, we report the mean score within each task type, the overall mean across all 21 benchmarks, and the reduction in effective parameter count computed via $(N_{\text{total}}^{QK} - N_{\text{eff}}^{QK})/N_{\text{total}}^{QK}$, where N_{total}^{QK} and N_{eff}^{QK} denote, respectively, the total and effective parameter counts of all W_Q and W_K matrices. For a low-rank matrix (W_Q or W_K for an attention head) with retained rank R_{thre} (after truncation), the effective parameter count for a low-rank weight matrix is $R_{\text{thre}} \times (d_{\text{head}} + d_{\text{model}})$, instead of the original size of $d_{\text{head}} \times d_{\text{model}}$.

Experimental Result. We present the result in Table 1. It can be observed that we can save decent parameter count while preserving the performance of the original model. Since the Qwen2.5-7B-Instruct model includes an additional bias term whose norm is relatively large compared to that of the weight matrix (see Table 12), we are able to discard more singular values while still maintaining performance. Specifically, we can set the power threshold $\text{thre} = 0.92$ without compromising its overall performance (37.84 v.s. 37.80 from the baseline). It saves 6.51% of the original W_Q and W_K weight parameters.

H. More results of Slash-Dominant Heads

In this section, we present additional results on SDHs with small Δ. For notational convenience, we slightly abuse $\mathbb{E}[S_{i,i-\Delta}]$ to represent the average slash score reported in the tables. In the following, we will abbreviate Gemma-7B, Llama3-8B-Instruct, and Qwen2.5-7B-Instruct as Gemma, Llama3, and Qwen2.5, respectively.

Model	SdQA	MdQA	Sum.	Few-shot	Syn.	Code	Avg.	Reduced Params
Llama-3-8B-Instruct	27.86	23.07	25.86	60.72	54.04	49.27	38.60	0
$\hookrightarrow$ $\texttt{thre} = 0.98, \texttt{rank}_{\texttt{thre}} = 64$	26.96	22.21	24.46	**61.73**	47.03	47.34	37.01	8.62%
$\hookrightarrow$ $\texttt{thre} = 0.96, \texttt{rank}_{\texttt{thre}} = 32$	26.46	**24.10**	25.40	**62.59**	**57.29**	35.99	38.00	5.90%
Qwen2.5-7B-Instruct	18.82	14.65	22.98	62.50	63.55	63.65	37.80	0
$\hookrightarrow$ $\texttt{thre} = 0.96, \texttt{rank}_{\texttt{thre}} = 64$	18.53	**15.59**	22.76	60.85	61.40	60.88	36.99	9.54%
$\hookrightarrow$ $\texttt{thre} = 0.92, \texttt{rank}_{\texttt{thre}} = 32$	**19.62**	**17.34**	**25.09**	62.04	61.36	57.13	**37.84**	6.51%

Table 1. **Parameter Compression Results.** The **bold** number indicates better performance than that of the baseline. The models are evaluated across 6 task types: Single-doc QA (SdQA), Multi-doc QA (MdQA), Summarization (Sum.), Few-shot, Synthetic (Syn.) and Code. The performance of the model is maintained with decent parameter reductions.

H.1. More Results of Slash-dominant Heads with Small Δ

For each Δ, we list at most three heads with the average slash score larger than 0.1. Due to space limitations, only a subset is shown; more slash-dominant heads with small Δ exist but are not included here.

H.1.1. Full List of SDHs with Small Δ

We report the full list of SDHs with $\Delta \in \{0, 1, 2, 3, 4\}$ for each model as follows. These SDHs are identified by using (6) with $\kappa = 0.1$, where the average slash scores are computed over 500 random prompts from `LongBenchV2` (Bai et al., 2025) dataset.

Gemma-7B Model.

- For $\Delta = 0$, the found SDHs are:

(Layers 0-2) L0H2, L0H4, L0H6, L0H10, L0H15, L1H2, L1H7, L2H0, L2H3, L2H8;
(Layers 3-5) L3H1, L3H7, L3H8, L4H0, L4H9, L4H12, L5H6, L5H11, L5H13;
(Layers 6-8) L6H1, L6H3, L6H11, L7H3, L7H6, L7H11, L7H14, L8H1, L8H4, L8H5, L8H9;
(Layers 9-11) L9H5, L9H8, L9H14, L10H2, L10H5, L10H7, L10H15, L11H4, L11H7, L11H8;
(Layers 12-20) L12H12, L13H1, L13H4, L14H14, L16H0, L17H8, L19H1, L19H15, L20H3, L20H7;
(Layers 21-25) L20H3, L20H7, L21H10, L22H5, L23H2, L23H14, L24H0, L24H12, L25H2, L25H4, L25H8;
(Layer 26) L26H0, L26H1, L26H3, L26H7, L26H9, L26H15;
(Layer 27) L27H0, L27H1, L27H2, L27H3, L27H4, L27H5, L27H6, L27H7, L27H8, L27H10, L27H13, L27H14.

- For $\Delta = 1$, the found SDHs are:

(Layers 0-2) L0H0, L0H1, L0H7, L0H8, L0H9, L0H14, L1H0, L1H7, L1H9, L1H14, L1H15, L2H2, L2H5;
(Layers 3-5) L3H2, L3H5, L3H9, L3H11, L3H14, L3H15, L4H6, L4H14;
(Layers 6-18) L6H7, L9H10, L10H10, L11H15, L13H3, L16H9, L17H0, L17H15, L18H11;
(Layers 19-27) L19H13, L20H5, L22H8, L23H3, L23H15, L25H2, L26H4, L26H11, L27H5.

- For $\Delta = 2$, the found SDHs are: L0H1, L1H7, L2H2, L2H9, L3H9, L13H3.
- For $\Delta = 3$, the found SDHs are: L0H1, L2H9.
- For $\Delta = 4$, the found SDH is: L0H1.

Llama3-8B-Instruct Model.

- For Llama3 and $\Delta = 0$, the found SDHs are:

(Layers 0-8) L0H2, L0H24, L4H5, L4H26, L6H5, L6H18, L7H0, L8H5, L8H13, L8H16, L8H23;
(Layers 9-14) L9H10, L10H0, L10H10, L12H12, L12H20, L12H30, L13H28, L14H8, L14H16, L14H19;
(Layers 15-19) L15H6, L15H10, L15H24, L16H22, L16H29, L16H30, L17H6, L17H12, L19H8, L19H11;
(Layers 20-26) L20H0, L20H3, L21H3, L21H8, L21H10, L22H9, L22H19, L25H21, L26H0, L26H1, L26H3, L26H22;

(Layers 27-29) L27H11, L28H10, L28H11, L28H17, L28H21, L28H29, L28H30, L29H8, L29H10, L29H25;
(Layer 30) L30H0, L30H24, L30H25, L30H26, L30H27;
(Layer 31) L31H0, L31H1, L31H2, L31H3, L31H5, L31H6, L31H7, L31H9, L31H11, L31H12, L31H14, L31H15, L31H18, L31H23, L31H24, L31H25, L31H26, L31H27.

- For Llama3 and $\Delta = 1$, the found SDHs are:

(Layers 0-1) L0H0, L0H1, L0H2, L0H3, L0H6, L0H10, L0H24, L0H26, L1H16, L1H18, L1H20;
(Layers 2-11) L4H12, L6H8, L6H16, L7H1, L7H2, L7H5, L8H22, L9H10, L9H11, L11H16, L11H19;
(Layers 12-17) L12H19, L14H8, L14H9, L14H11, L14H26, L15H0, L15H7, L16H22, L16H29, L17H4, L17H5;
(Layers 18-28) L18H26, L19H8, L21H4, L21H7, L21H10, L25H20, L25H22, L25H23, L27H29, L28H26;
(Layers 29-31) L29H8, L29H26, L29H27, L30H10, L31H6, L31H15, L31H24

- For Llama3 and $\Delta = 2$, the found SDHs are L0H0, L0H2, L1H21, L7H2, L9H11, L14H8, L14H26.
- For Llama3 and $\Delta = 3$, the found SDH is L0H0.
- For Llama3 and $\Delta = 4$, the found SDH is L0H0.

Qwen2.5-7B-Instruct Model.

- For Qwen2.5 and $\Delta = 0$, the found SDHs are:

(Layers 0-17) L14H9, L14H20, L14H25, L15H8, L15H23, L16H19, L16H26, L17H4;
(Layers 18-31) L18H0, L18H7, L18H17, L18H25, L19H0, L19H11, L19H13, L20H1, L20H2, L20H14, L24H18, L25H2, L25H3, L25H4, L25H20.

- For Qwen2.5 and $\Delta = 1$, the found SDHs are:

(Layer 0) L0H6, L0H7, L0H16, L0H20, L0H23, L0H24, L0H26;
(Layer 1) L1H3, L1H10, L1H11, L1H12, L1H14, L1H15, L1H17, L1H18, L1H19, L1H23, L1H24;
(Layers 2-3) L2H1, L2H2, L2H5, L2H6, L2H27, L3H1, L3H3, L3H10, L3H15, L3H22, L3H24, L3H26, L3H27;
(Layers 4-6) L4H18, L4H23, L5H3, L6H1, L6H2, L6H4, L6H6, L6H15, L6H16, L6H21, L6H25, L6H26;
(Layers 7-14) L7H4, L7H5, L10H4, L10H22, L10H25, L10H27, L11H1, L11H18, L12H21, L13H8, L13H13;
(Layers 15-16) L15H5, L15H8, L15H9, L15H11, L15H13, L15H23, L15H27, L16H21, L16H23;
(Layers 17-20) L17H2, L17H3, L17H9, L17H11, L18H9, L18H20, L19H8, L19H10, L19H12, L19H13, L20H1, L20H5;
(Layers 21-23) L21H15, L21H16, L21H18, L21H19, L21H20, L21H21, L23H24, L23H26, L23H27;
(Layers 24-27) L24H0, L24H1, L24H3, L24H5, L24H6, L24H20, L25H5, L25H14, L25H18, L25H19, L26H11, L26H26, L26H27, L27H8.

- For Qwen2.5 and $\Delta = 2$, the found SDHs are:

(Layer 0) L0H1, L0H5, L0H7, L0H16, L0H23, L0H24, L0H26;
(Layer 1) L1H3, L1H12, L1H14, L1H15, L1H16, L1H17, L1H19, L1H24;
(Layers 2-7) L2H27, L3H1, L3H9, L3H10, L3H22, L3H24, L3H25, L3H26, L3H27, L6H15, L7H4;
(Layers 8-15) L10H4, L10H22, L13H8, L13H10, L13H13, L15H5, L15H8, L15H9, L15H11, L15H12;
(Layers 16-27) L16H21, L17H2, L18H9, L19H8, L20H1, L21H16, L21H20.

- For Qwen2.5 and $\Delta = 3$, the found SDHs are L0H1, L0H5, L1H0, L1H3, L1H17, L3H9, L6H15, L13H10.
- For Qwen2.5 and $\Delta = 4$, the found SDHs are L1H0, L1H17, L6H15.

H.1.2. Attention Scores and Ranks of Q, K, and H

In this section, we present per-head statistics for SDHs with small Δ. For conciseness, we report at most three heads for each Δ in each model. Specifically, for each Δ, we rank the SDHs by their average slash score and list the three heads with the largest values in the following table.

Model	Head	$\tilde{r}_1(Q)$	$\tilde{R}_{0.95}(Q)$	$\tilde{r}_1(K)$	$\tilde{R}_{0.95}(K)$
Gemma	L0H4	0.895	2	0.941	2
Gemma	L0H7	0.953	1	0.978	1
Gemma	L0H1	0.954	1	0.994	1
Llama3	L0H2	0.860	6	0.738	16
Llama3	L0H0	0.753	7	0.738	16
Qwen2.5	L0H6	0.264	59	0.165	73
Qwen2.5	L0H5	0.2751	67	0.165	73

Table 2. The values of $\tilde{r}_1$ and $\tilde{R}_{0.95}$ for the SDHs located in the 0-th layer of Gemma-7B, Llama3-8B-Instruct, and Qwen2.5-7B-Instruct.

Model	Δ	Head	$\mathbb{E}[S_{i,i-\Delta}]$	OOD $\mathbb{E}[S_{i,i-\Delta}]$	$r_1(Q)$	$R_{0.95}(Q)$	$r_1(K)$	$R_{0.95}(K)$	$r_1(H)$	$R_{0.95}(H)$
Gemma	0	L0H4	1.000	1.000	0.822	2	0.854	2	0.343	180
Gemma	0	L26H3	1.000	1.000	0.775	14	0.726	17	0.283	348
Gemma	0	L27H8	1.000	1.000	0.890	3	0.918	2	0.363	241
Gemma	1	L0H7	0.904	0.999	0.888	2	0.942	2	0.343	180
Gemma	1	L19H13	0.364	0.741	0.895	20	0.896	18	0.225	408
Gemma	1	L13H3	0.333	0.752	0.885	16	0.914	8	0.274	375
Gemma	2	L0H1	0.302	0.352	0.918	3	0.991	1	0.343	180
Gemma	2	L3H9	0.148	0.278	0.824	36	0.891	11	0.273	364
Gemma	2	L2H2	0.127	0.298	0.801	52	0.884	23	0.241	334
Gemma	3	L0H1	0.197	0.172	0.918	3	0.991	1	0.343	180
Gemma	3	L2H9	0.115	0.144	0.875	18	0.944	2	0.241	334
Gemma	4	L0H1	0.101	0.072	0.918	3	0.991	1	0.343	180
Llama3	0	L31H14	0.961	0.991	0.876	4	0.801	26	0.243	424
Llama3	0	L12H12	0.698	0.667	0.946	2	0.759	26	0.265	448
Llama3	0	L16H30	0.539	0.538	0.892	9	0.848	15	0.254	388
Llama3	1	L0H2	0.560	0.656	0.957	1	0.860	6	0.310	191
Llama3	1	L14H26	0.516	0.701	0.965	1	0.788	23	0.271	400
Llama3	1	L1H20	0.333	0.749	0.978	1	0.838	19	0.388	160
Llama3	2	L0H0	0.180	0.160	0.956	1	0.860	6	0.310	191
Llama3	2	L0H2	0.169	0.141	0.957	1	0.860	6	0.310	191
Llama3	2	L1H21	0.165	0.410	0.977	1	0.838	19	0.388	160
Llama3	3	L0H0	0.159	0.125	0.957	1	0.860	6	0.310	191
Llama3	4	L0H0	0.109	0.080	0.957	1	0.860	6	0.310	191
Qwen2.5	0	L18H7	1.000	1.000	0.986	1	0.703	26	0.215	419
Qwen2.5	0	L19H0	1.000	1.000	0.982	1	0.952	1	0.232	374
Qwen2.5	0	L14H25	0.742	0.800	0.875	12	0.589	34	0.225	324
Qwen2.5	1	L21H15	0.699	0.840	0.963	1	0.712	33	0.235	360
Qwen2.5	1	L13H13	0.664	0.904	0.952	1	0.699	32	0.244	366
Qwen2.5	1	L0H6	0.655	0.572	0.912	3	0.999	1	0.171	278
Qwen2.5	2	L1H3	0.297	0.327	0.920	3	0.999	1	0.458	155
Qwen2.5	2	L1H17	0.200	0.231	0.869	5	0.975	1	0.458	155
Qwen2.5	2	L0H5	0.192	0.100	0.813	23	0.999	1	0.171	278
Qwen2.5	3	L0H5	0.161	0.121	0.813	23	0.999	1	0.171	278
Qwen2.5	3	L1H17	0.156	0.165	0.869	5	0.975	1	0.458	155
Qwen2.5	3	L1H3	0.154	0.089	0.920	3	0.999	1	0.458	155
Qwen2.5	4	L1H0	0.174	0.205	0.920	3	0.999	1	0.458	155
Qwen2.5	4	L6H15	0.114	0.049	0.898	6	0.919	5	0.331	317
Qwen2.5	4	L1H17	0.107	0.098	0.869	5	0.975	1	0.458	155

Table 3. This table lists the average attention scores of prompts in LongBench and OOD prompts, the rank information of Q, K, and H.

Model	Head	$r_1(W_Q)$	$R_{0.95}(W_Q)$	$r_1(W_K)$	$R_{0.95}(W_K)$
Gemma	L0H4	0.479	10	0.534	10
Gemma	L26H3	0.091	148	0.081	160
Gemma	L27H8	0.255	47	0.244	47
Gemma	L0H7	0.182	122	0.219	135
Gemma	L19H13	0.052	221	0.049	203
Gemma	L13H3	0.044	203	0.039	203
Gemma	L0H1	0.054	210	0.362	194
Gemma	L3H9	0.026	217	0.045	215
Gemma	L2H2	0.022	218	0.029	218
Gemma	L2H9	0.039	218	0.067	215
Llama3	L31H14	0.157	109	0.040	107
Llama3	L12H12	0.141	101	0.040	99
Llama3	L16H30	0.075	107	0.051	102
Llama3	L0H2	0.370	61	0.187	62
Llama3	L14H26	0.178	80	0.037	98
Llama3	L1H20	0.114	69	0.047	71
Llama3	L0H0	0.417	60	0.187	62
Llama3	L1H21	0.109	68	0.047	71
Qwen2.5	L18H7	0.197	77	0.031	98
Qwen2.5	L19H0	0.178	81	0.048	90
Qwen2.5	L14H25	0.041	106	0.024	95
Qwen2.5	L21H15	0.060	109	0.024	105
Qwen2.5	L13H13	0.097	106	0.025	102
Qwen2.5	L0H6	0.061	90	0.021	107
Qwen2.5	L1H3	0.086	82	0.072	82
Qwen2.5	L1H17	0.084	73	0.042	80
Qwen2.5	L0H5	0.029	107	0.020	107
Qwen2.5	L1H17	0.084	73	0.042	80
Qwen2.5	L1H0	0.094	80	0.072	82
Qwen2.5	L6H15	0.044	105	0.032	98

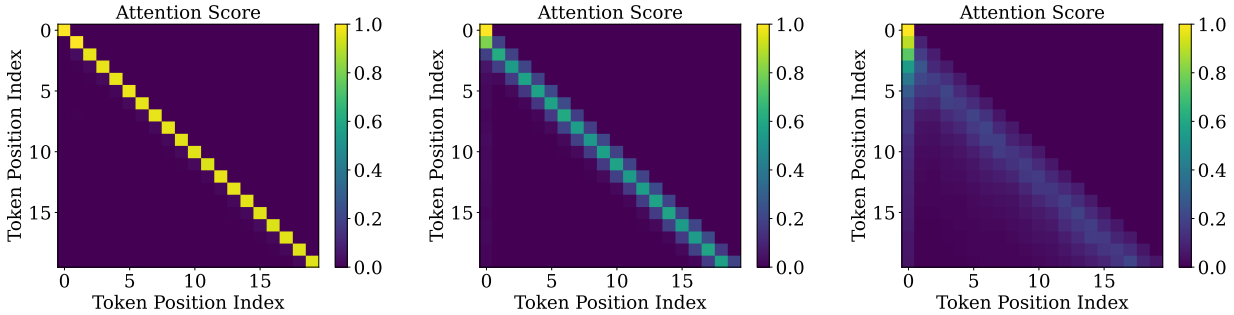
Table 4. The table reports r_1 and $R_{0.95}$ for the query and key projection matrices, W_Q and W_K, across attention heads in LLMs.

Model	Δ	Head	$\mathbb{E}[S_{i,i-\Delta}]$	$\mathbb{E}[S_{i,i-\Delta}]$ w/o high freqs	$\mathbb{E}[S_{i,i-\Delta}]$ w/o med freqs	$\mathbb{E}[S_{i,i-\Delta}]$ w/o low freqs
Gemma	0	L0H4	1.000	0.884	1.000	1.000
Gemma	0	L26H3	1.000	0.807	0.999	1.000
Gemma	0	L27H8	1.000	0.680	1.000	1.000
Gemma	1	L0H7	0.904	0.043	0.902	0.902
Gemma	1	L19H13	0.364	0.282	0.282	0.282
Gemma	1	L13H3	0.333	0.274	0.274	0.274
Gemma	2	L0H1	0.302	0.085	0.072	0.072
Gemma	2	L3H9	0.148	0.103	0.103	0.104
Gemma	2	L2H2	0.127	0.045	0.046	0.046
Gemma	3	L0H1	0.197	0.143	0.125	0.125
Gemma	3	L2H9	0.115	0.054	0.055	0.042
Gemma	4	L0H1	0.101	0.170	0.172	0.173
Llama3	0	L31H14	0.961	0.653	0.908	0.956
Llama3	0	L12H12	0.698	0.431	0.691	0.701
Llama3	0	L16H30	0.539	0.306	0.493	0.542
Llama3	1	L0H2	0.560	0.114	0.530	0.557
Llama3	1	L14H26	0.516	0.077	0.509	0.515
Llama3	1	L1H20	0.333	0.035	0.331	0.341
Llama3	2	L0H0	0.180	0.066	0.155	0.179
Llama3	2	L0H2	0.169	0.121	0.164	0.170
Llama3	2	L1H21	0.165	0.006	0.164	0.168
Llama3	3	L0H0	0.159	0.076	0.140	0.159
Llama3	4	L0H0	0.109	0.083	0.098	0.109
Qwen2.5	0	L18H7	1.000	0.967	0.967	1.000
Qwen2.5	0	L19H0	1.000	0.943	0.993	0.999
Qwen2.5	0	L14H25	0.742	0.473	0.704	0.741
Qwen2.5	1	L21H15	0.699	0.114	0.653	0.693
Qwen2.5	1	L13H13	0.664	0.108	0.629	0.654
Qwen2.5	1	L0H6	0.655	0.120	0.558	0.647
Qwen2.5	2	L1H3	0.297	0.137	0.295	0.296
Qwen2.5	2	L1H17	0.200	0.078	0.175	0.199
Qwen2.5	2	L0H5	0.192	0.062	0.149	0.188
Qwen2.5	3	L0H5	0.161	0.074	0.127	0.161
Qwen2.5	3	L1H17	0.156	0.089	0.138	0.156
Qwen2.5	3	L1H3	0.154	0.120	0.152	0.153
Qwen2.5	4	L1H0	0.174	0.109	0.173	0.174
Qwen2.5	4	L6H15	0.114	0.076	0.095	0.114
Qwen2.5	4	L1H17	0.107	0.094	0.095	0.106

Table 5. This table quantifies the effect of low-, medium-, and high-frequency components on SDHs by reporting, for each band, the average slash score after removing that band.

H.2. Figures of Slash-dominant Heads with Small Δ and Large Average Slash Scores

H.2.1. Results of Llama3-8B-Instruct

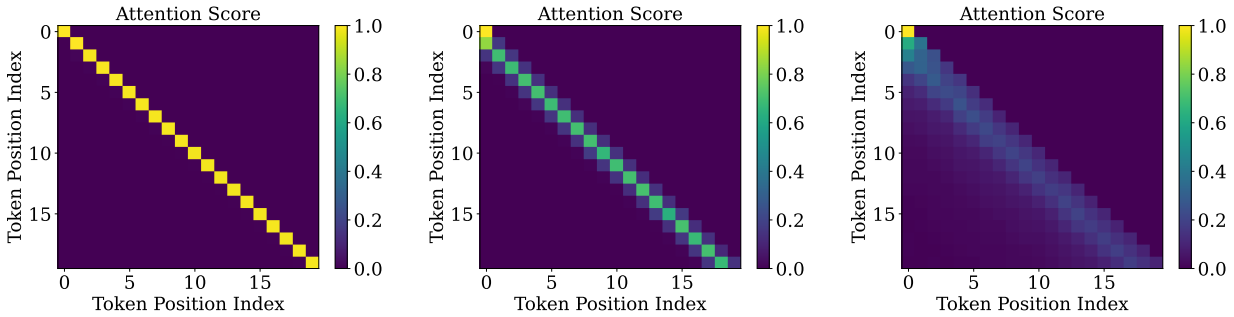


(a) Average attention score matrix of L31H14.

(b) Average attention score matrix of L0H2.

(c) Average attention score matrix of L0H0.

Figure 24. Average of attention score matrices in Llama3-8B-Instruct with prompts from LongBench V2.

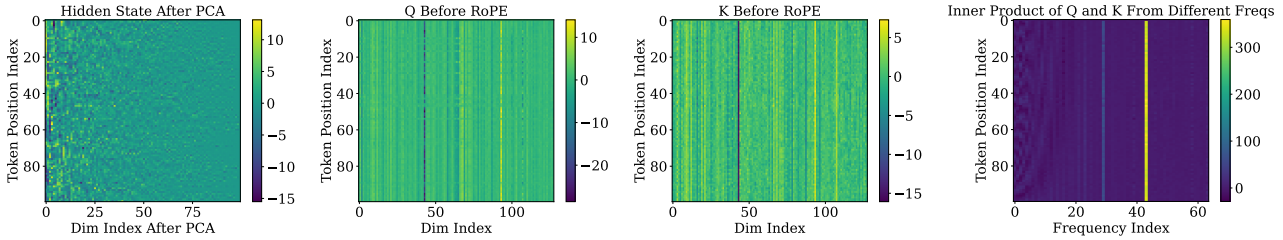


(a) Average attention score matrix of L31H14.

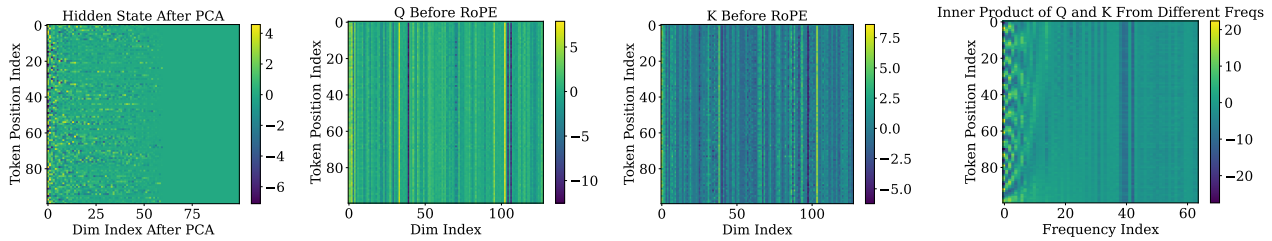
(b) Average attention score matrix of L0H2.

(c) Average attention score matrix of L0H0.

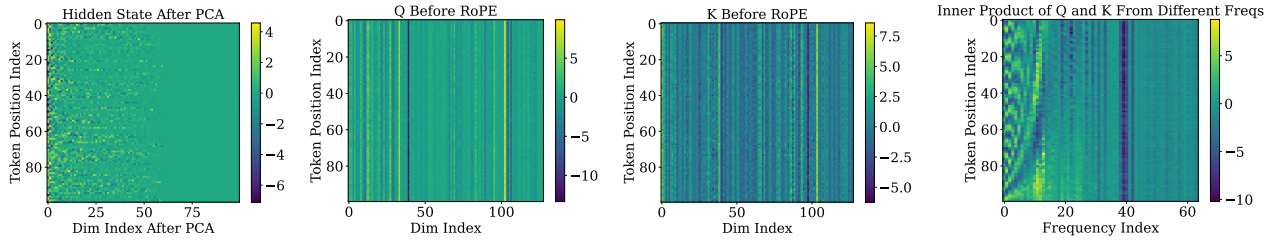
Figure 25. Average of attention score matrices in Llama3-8B-Instruct with prompts whose tokens are i.i.d. sampled from the uniform distribution on the alphabet.



(a) Hidden state H of L31H14 after PCA. (b) Queries Q of L31H14. (c) Keys K of L31H14. (d) $\mathrm{InP}(100, j, l)$ of L31H14.



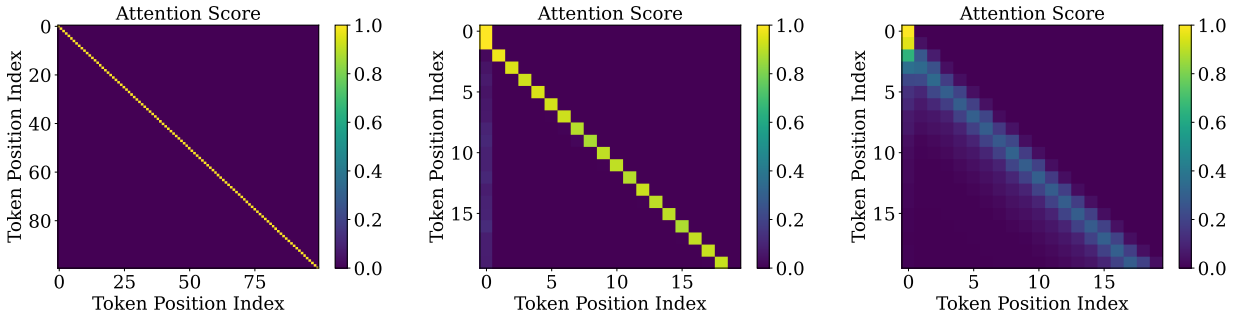
(e) Hidden state H of L0H2 after PCA. (f) Queries Q of L0H2. (g) Keys K of L0H2. (h) $\mathrm{InP}(100, j, l)$ of L0H2.



(i) Hidden state H of L0H0 after PCA. (j) Queries Q of L0H0. (k) Keys K of L0H0. (l) $\mathrm{InP}(100, j, l)$ of L0H1.

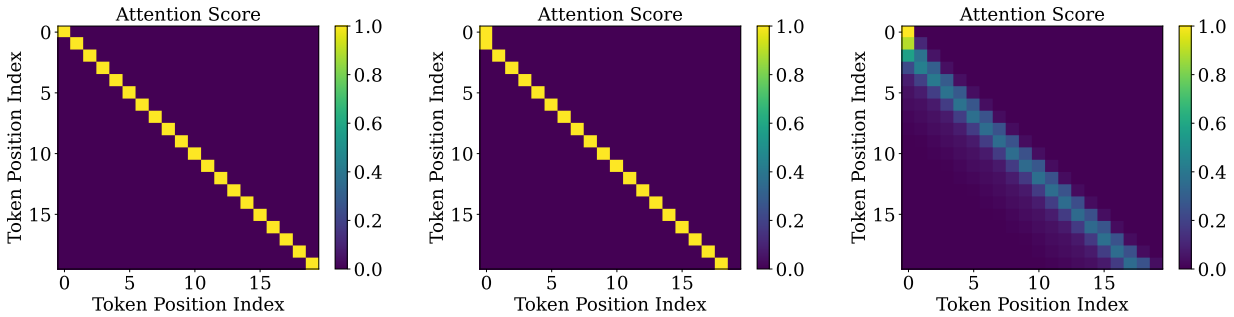
Figure 26. This figure shows the hidden states, queries, keys, and $\mathrm{InP}(100, j, l)$ for $j \in [100], l \in [64]$ for Llama3-8B-Instruct. Here, the example prompt contains 100 tokens. The dimensions of queries and keys are both 128.

H.2.2. Results of Gemma-7B



(a) Average attention score matrix of L0H4. (b) Average attention score matrix of L0H7. (c) Average attention score matrix of L0H1.

Figure 27. Average of attention score matrices in Gemma-7B with prompts from LongBench V2.

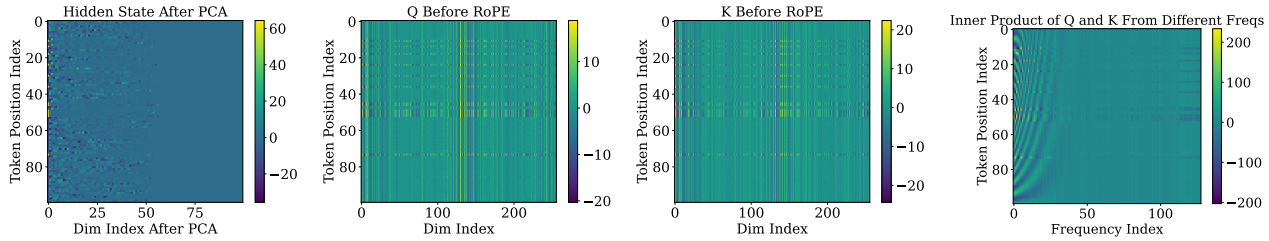


(a) Average attention score matrix of L0H4.

(b) Average attention score matrix of L0H7.

(c) Average attention score matrix of L0H1.

Figure 28. Average of attention score matrices in Gemma-7B with prompts whose tokens are i.i.d. sampled from the uniform distribution on the alphabet.

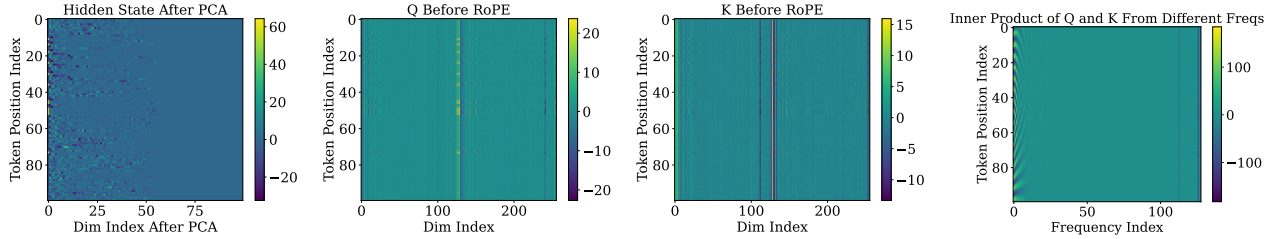


(a) Hidden state H of L0H4 after PCA.

(b) Queries Q of L0H4.

(c) Keys K of L0H4.

(d) $\mathrm{InP}(100, j, l)$ of L0H4.

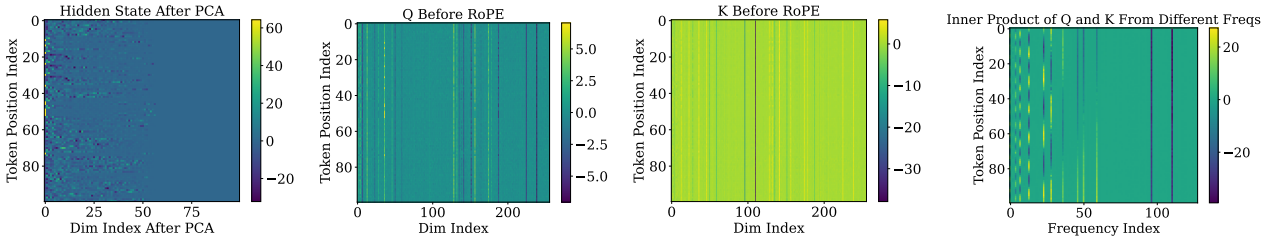


(e) Hidden state H of L0H7 after PCA.

(f) Queries Q of L0H7.

(g) Keys K of L0H7.

(h) $\mathrm{InP}(100, j, l)$ of L0H7.



(i) Hidden state H of L0H1 after PCA.

(j) Queries Q of L0H1.

(k) Keys K of L0H1.

(l) $\mathrm{InP}(100, j, l)$ of L0H1.

Figure 29. This figure shows the hidden states, queries, keys, and $\mathrm{InP}(100, j, l)$ for $j \in [100], l \in [128]$ for Gemma-7B. Here, the example prompt contains 100 tokens. The dimensions of queries and keys are both 256. We implement PCA for the Hidden state H to reduce its dimension to 100.

H.3. More Results of Slash-dominant Heads with Large Δ

In this section, we present additional results on SDHs with large Δ. For notational convenience, we slightly abuse $\mathbb{E}[S_{i,i-\Delta}]$ to represent the average slash score reported in the tables. In the following, we will abbreviate Llama3-8B-Instruct, and Qwen2.5-7B-Instruct as Llama3, and Qwen2.5, respectively.

H.3.1. Full List of SDHs with Large Δ

In the following, we report each head and its corresponding Δ in the form $(\mathrm{L}a\mathrm{H}b, \Delta)$. As mentioned in Section C, we set $\kappa = 10^{-3}$ and do not aim to find all the slash patterns for large Δ but select a subset of them.

Llama3-8B-Instruct Model.

(Layer 0) $(\mathrm{L0H29}, 563)$, $(\mathrm{L0H29}, 564)$, $(\mathrm{L0H29}, 565)$, $(\mathrm{L0H29}, 566)$, $(\mathrm{L0H29}, 619)$, $(\mathrm{L0H29}, 620)$, $(\mathrm{L0H29}, 621)$, $(\mathrm{L0H29}, 689)$, $(\mathrm{L0H29}, 690)$, $(\mathrm{L0H29}, 849)$, $(\mathrm{L0H29}, 850)$, $(\mathrm{L0H29}, 851)$, $(\mathrm{L0H29}, 852)$, $(\mathrm{L0H29}, 853)$, $(\mathrm{L0H29}, 920)$, $(\mathrm{L0H29}, 921)$, $(\mathrm{L0H29}, 922)$, $(\mathrm{L0H29}, 934)$, $(\mathrm{L0H29}, 935)$, $(\mathrm{L0H29}, 1037)$, $(\mathrm{L0H29}, 1044)$, $(\mathrm{L0H29}, 1045)$, $(\mathrm{L0H29}, 1046)$, $(\mathrm{L0H29}, 1283)$, $(\mathrm{L0H29}, 1575)$, $(\mathrm{L0H29}, 1576)$, $(\mathrm{L0H29}, 2117)$, $(\mathrm{L0H29}, 2247)$, $(\mathrm{L0H29}, 2248)$, $(\mathrm{L0H29}, 2354)$, $(\mathrm{L0H29}, 2355)$, $(\mathrm{L0H29}, 2362)$, $(\mathrm{L0H29}, 2363)$, $(\mathrm{L0H29}, 2364)$, $(\mathrm{L0H29}, 2371)$, $(\mathrm{L0H29}, 2372)$, $(\mathrm{L0H29}, 2373)$, $(\mathrm{L0H29}, 2374)$, $(\mathrm{L0H29}, 2409)$, $(\mathrm{L0H29}, 2410)$, $(\mathrm{L0H29}, 2902)$, $(\mathrm{L0H29}, 2910)$, $(\mathrm{L0H29}, 2911)$, $(\mathrm{L0H29}, 2912)$, $(\mathrm{L0H29}, 3073)$, $(\mathrm{L0H29}, 3574)$, $(\mathrm{L0H30}, 599)$, $(\mathrm{L0H30}, 600)$, $(\mathrm{L0H30}, 601)$, $(\mathrm{L0H30}, 685)$, $(\mathrm{L0H30}, 686)$, $(\mathrm{L0H30}, 687)$, $(\mathrm{L0H30}, 692)$, $(\mathrm{L0H30}, 693)$, $(\mathrm{L0H30}, 1270)$, $(\mathrm{L0H30}, 1271)$, $(\mathrm{L0H30}, 1272)$, $(\mathrm{L0H30}, 1273)$, $(\mathrm{L0H30}, 2112)$, $(\mathrm{L0H30}, 2113)$, $(\mathrm{L0H30}, 2114)$, $(\mathrm{L0H30}, 2899)$, $(\mathrm{L0H30}, 2900)$, $(\mathrm{L0H30}, 3909)$, $(\mathrm{L0H30}, 3910)$, $(\mathrm{L0H30}, 3911)$, $(\mathrm{L0H30}, 3912)$, $(\mathrm{L0H30}, 3913)$, $(\mathrm{L0H30}, 3914)$, $(\mathrm{L0H30}, 3915)$.

(Layer 1) $(\mathrm{L1H16}, 3827)$, $(\mathrm{L1H18}, 3827)$.

Qwen2.5-7B-Instruct Model.

(Layer 0) $(\mathrm{L0H0}, 683)$, $(\mathrm{L0H0}, 684)$, $(\mathrm{L0H0}, 731)$, $(\mathrm{L0H0}, 732)$, $(\mathrm{L0H0}, 733)$, $(\mathrm{L0H0}, 740)$, $(\mathrm{L0H0}, 741)$, $(\mathrm{L0H0}, 1487)$, $(\mathrm{L0H0}, 1488)$, $(\mathrm{L0H1}, 673)$, $(\mathrm{L0H2}, 657)$, $(\mathrm{L0H2}, 658)$, $(\mathrm{L0H2}, 659)$, $(\mathrm{L0H2}, 660)$, $(\mathrm{L0H2}, 731)$, $(\mathrm{L0H2}, 732)$, $(\mathrm{L0H2}, 733)$, $(\mathrm{L0H2}, 734)$, $(\mathrm{L0H2}, 739)$, $(\mathrm{L0H2}, 740)$, $(\mathrm{L0H2}, 741)$, $(\mathrm{L0H2}, 742)$, $(\mathrm{L0H2}, 743)$, $(\mathrm{L0H2}, 744)$, $(\mathrm{L0H2}, 745)$, $(\mathrm{L0H2}, 746)$, $(\mathrm{L0H2}, 753)$, $(\mathrm{L0H2}, 754)$, $(\mathrm{L0H2}, 755)$, $(\mathrm{L0H7}, 912)$, $(\mathrm{L0H7}, 924)$, $(\mathrm{L0H7}, 925)$, $(\mathrm{L0H7}, 936)$, $(\mathrm{L0H7}, 937)$, $(\mathrm{L0H7}, 938)$, $(\mathrm{L0H7}, 939)$, $(\mathrm{L0H8}, 922)$, $(\mathrm{L0H8}, 923)$, $(\mathrm{L0H8}, 924)$, $(\mathrm{L0H8}, 925)$, $(\mathrm{L0H8}, 935)$, $(\mathrm{L0H8}, 936)$, $(\mathrm{L0H8}, 937)$, $(\mathrm{L0H8}, 938)$, $(\mathrm{L0H8}, 939)$, $(\mathrm{L0H8}, 940)$, $(\mathrm{L0H8}, 1030)$, $(\mathrm{L0H8}, 1031)$, $(\mathrm{L0H9}, 912)$, $(\mathrm{L0H9}, 913)$, $(\mathrm{L0H9}, 914)$, $(\mathrm{L0H9}, 915)$, $(\mathrm{L0H9}, 916)$, $(\mathrm{L0H9}, 917)$, $(\mathrm{L0H9}, 918)$, $(\mathrm{L0H9}, 919)$, $(\mathrm{L0H9}, 920)$, $(\mathrm{L0H9}, 921)$, $(\mathrm{L0H9}, 922)$, $(\mathrm{L0H9}, 923)$, $(\mathrm{L0H9}, 924)$, $(\mathrm{L0H9}, 925)$, $(\mathrm{L0H9}, 926)$, $(\mathrm{L0H9}, 927)$, $(\mathrm{L0H9}, 928)$, $(\mathrm{L0H9}, 929)$, $(\mathrm{L0H9}, 930)$, $(\mathrm{L0H9}, 936)$, $(\mathrm{L0H9}, 937)$, $(\mathrm{L0H11}, 923)$, $(\mathrm{L0H11}, 924)$, $(\mathrm{L0H11}, 925)$, $(\mathrm{L0H11}, 926)$, $(\mathrm{L0H11}, 927)$, $(\mathrm{L0H11}, 928)$, $(\mathrm{L0H11}, 929)$, $(\mathrm{L0H11}, 930)$, $(\mathrm{L0H11}, 937)$, $(\mathrm{L0H11}, 938)$, $(\mathrm{L0H11}, 939)$, $(\mathrm{L0H11}, 940)$, $(\mathrm{L0H11}, 941)$, $(\mathrm{L0H15}, 917)$, $(\mathrm{L0H15}, 918)$, $(\mathrm{L0H15}, 936)$, $(\mathrm{L0H15}, 1513)$, $(\mathrm{L0H15}, 1514)$, $(\mathrm{L0H15}, 1515)$, $(\mathrm{L0H15}, 1531)$, $(\mathrm{L0H15}, 1532)$, $(\mathrm{L0H15}, 1533)$, $(\mathrm{L0H15}, 1534)$, $(\mathrm{L0H15}, 1535)$, $(\mathrm{L0H15}, 3094)$, $(\mathrm{L0H15}, 3095)$, $(\mathrm{L0H15}, 3096)$, $(\mathrm{L0H15}, 3097)$, $(\mathrm{L0H15}, 3098)$, $(\mathrm{L0H15}, 3140)$, $(\mathrm{L0H15}, 3141)$, $(\mathrm{L0H15}, 3142)$, $(\mathrm{L0H15}, 3143)$, $(\mathrm{L0H15}, 3303)$, $(\mathrm{L0H15}, 3304)$, $(\mathrm{L0H15}, 3305)$, $(\mathrm{L0H15}, 3306)$, $(\mathrm{L0H15}, 3307)$, $(\mathrm{L0H15}, 3323)$, $(\mathrm{L0H15}, 3324)$, $(\mathrm{L0H15}, 3362)$, $(\mathrm{L0H15}, 3363)$, $(\mathrm{L0H15}, 3364)$, $(\mathrm{L0H15}, 3365)$, $(\mathrm{L0H15}, 3366)$, $(\mathrm{L0H15}, 3367)$, $(\mathrm{L0H15}, 3368)$, $(\mathrm{L0H15}, 3412)$, $(\mathrm{L0H15}, 3413)$, $(\mathrm{L0H15}, 3586)$, $(\mathrm{L0H15}, 3587)$, $(\mathrm{L0H15}, 3588)$, $(\mathrm{L0H15}, 3617)$, $(\mathrm{L0H15}, 3618)$, $(\mathrm{L0H15}, 3619)$, $(\mathrm{L0H15}, 3620)$, $(\mathrm{L0H15}, 3644)$, $(\mathrm{L0H15}, 3645)$, $(\mathrm{L0H15}, 3646)$, $(\mathrm{L0H15}, 3647)$, $(\mathrm{L0H15}, 3648)$, $(\mathrm{L0H15}, 3693)$, $(\mathrm{L0H15}, 3694)$, $(\mathrm{L0H15}, 3695)$, $(\mathrm{L0H16}, 1521)$, $(\mathrm{L0H16}, 1538)$, $(\mathrm{L0H16}, 1539)$, $(\mathrm{L0H16}, 1540)$, $(\mathrm{L0H23}, 500)$, $(\mathrm{L0H23}, 501)$, $(\mathrm{L0H23}, 502)$, $(\mathrm{L0H23}, 503)$, $(\mathrm{L0H23}, 504)$, $(\mathrm{L0H23}, 505)$, $(\mathrm{L0H24}, 500)$, $(\mathrm{L0H24}, 501)$, $(\mathrm{L0H24}, 502)$, $(\mathrm{L0H24}, 503)$, $(\mathrm{L0H24}, 504)$, $(\mathrm{L0H24}, 505)$, $(\mathrm{L0H25}, 500)$, $(\mathrm{L0H25}, 501)$, $(\mathrm{L0H25}, 503)$, $(\mathrm{L0H25}, 504)$, $(\mathrm{L0H25}, 505)$, $(\mathrm{L0H25}, 506)$, $(\mathrm{L0H25}, 507)$, $(\mathrm{L0H25}, 508)$, $(\mathrm{L0H25}, 509)$, $(\mathrm{L0H25}, 510)$, $(\mathrm{L0H25}, 511)$, $(\mathrm{L0H26}, 500)$, $(\mathrm{L0H26}, 501)$, $(\mathrm{L0H26}, 502)$, $(\mathrm{L0H26}, 503)$, $(\mathrm{L0H26}, 504)$, $(\mathrm{L0H26}, 505)$.

(Layer 1) $(\mathrm{L1H6}, 502)$, $(\mathrm{L1H6}, 503)$, $(\mathrm{L1H6}, 504)$, $(\mathrm{L1H6}, 505)$, $(\mathrm{L1H6}, 506)$, $(\mathrm{L1H6}, 507)$, $(\mathrm{L1H6}, 508)$, $(\mathrm{L1H6}, 514)$, $(\mathrm{L1H6}, 515)$, $(\mathrm{L1H6}, 516)$, $(\mathrm{L1H6}, 517)$, $(\mathrm{L1H6}, 520)$, $(\mathrm{L1H6}, 521)$, $(\mathrm{L1H6}, 522)$, $(\mathrm{L1H6}, 523)$, $(\mathrm{L1H6}, 524)$, $(\mathrm{L1H6}, 525)$, $(\mathrm{L1H6}, 526)$, $(\mathrm{L1H6}, 527)$, $(\mathrm{L1H6}, 528)$, $(\mathrm{L1H6}, 529)$, $(\mathrm{L1H6}, 533)$, $(\mathrm{L1H6}, 534)$, $(\mathrm{L1H6}, 535)$, $(\mathrm{L1H6}, 536)$, $(\mathrm{L1H6}, 537)$, $(\mathrm{L1H6}, 538)$, $(\mathrm{L1H6}, 539)$, $(\mathrm{L1H6}, 540)$, $(\mathrm{L1H6}, 541)$, $(\mathrm{L1H6}, 542)$, $(\mathrm{L1H6}, 543)$, $(\mathrm{L1H6}, 553)$, $(\mathrm{L1H11}, 553)$, $(\mathrm{L1H11}, 554)$, $(\mathrm{L1H11}, 609)$, $(\mathrm{L1H11}, 610)$, $(\mathrm{L1H11}, 641)$, $(\mathrm{L1H11}, 710)$, $(\mathrm{L1H11}, 717)$, $(\mathrm{L1H11}, 718)$, $(\mathrm{L1H11}, 804)$, $(\mathrm{L1H11}, 805)$, $(\mathrm{L1H11}, 835)$, $(\mathrm{L1H11}, 1239)$, $(\mathrm{L1H11}, 1521)$, $(\mathrm{L1H11}, 3532)$, $(\mathrm{L1H12}, 546)$, $(\mathrm{L1H12}, 602)$, $(\mathrm{L1H12}, 640)$, $(\mathrm{L1H12}, 641)$, $(\mathrm{L1H13}, 640)$, $(\mathrm{L1H13}, 641)$, $(\mathrm{L1H13}, 647)$, $(\mathrm{L1H13}, 648)$, $(\mathrm{L1H15}, 516)$, $(\mathrm{L1H15}, 523)$, $(\mathrm{L1H15}, 554)$, $(\mathrm{L1H15}, 2294)$, $(\mathrm{L1H16}, 541)$, $(\mathrm{L1H16}, 554)$, $(\mathrm{L1H16}, 555)$, $(\mathrm{L1H16}, 556)$, $(\mathrm{L1H16}, 611)$, $(\mathrm{L1H16}, 671)$, $(\mathrm{L1H16}, 672)$, $(\mathrm{L1H16}, 673)$, $(\mathrm{L1H16}, 2778)$, $(\mathrm{L1H19}, 553)$, $(\mathrm{L1H19}, 554)$, $(\mathrm{L1H19}, 610)$, $(\mathrm{L1H19}, 611)$, $(\mathrm{L1H19}, 648)$, $(\mathrm{L1H20}, 554)$,

(L1H23, 553), (L1H23, 554), (L1H23, 678), (L1H23, 679), (L1H23, 685), (L1H23, 686), (L1H23, 687), (L1H23, 709), (L1H23, 710), (L1H23, 717), (L1H23, 718), (L1H23, 998), (L1H23, 999), (L1H23, 1036), (L1H23, 1037), (L1H23, 1162), (L1H23, 1163), (L1H23, 1201), (L1H23, 1848), (L1H23, 1849), (L1H23, 2129), (L1H23, 2130), (L1H23, 2293), (L1H23, 2294), (L1H23, 2295), (L1H23, 2613), (L1H23, 2620), (L1H23, 2621), (L1H23, 2622), (L1H23, 2777), (L1H23, 2932), (L1H23, 2933), (L1H23, 2934), (L1H23, 3097), (L1H23, 3438), (L1H24, 671), (L1H24, 834), (L1H24, 835), (L1H24, 836), (L1H24, 1145), (L1H24, 1146), (L1H24, 1147), (L1H24, 1148), (L1H24, 1149), (L1H24, 2238), (L1H24, 2683), (L1H24, 2684), (L1H24, 2721), (L1H24, 3368), (L1H24, 3369).

(Layer 2) (L2H6, 514), (L2H6, 515), (L2H6, 741), (L2H6, 936), (L2H6, 998), (L2H6, 1933), (L2H6, 1934), (L2H6, 1935), (L2H6, 2159), (L2H6, 2160), (L2H6, 2161), (L2H6, 2185), (L2H6, 2186), (L2H6, 2198), (L2H6, 2273), (L2H6, 2274), (L2H6, 2275), (L2H6, 2362), (L2H6, 2363), (L2H6, 2758), (L2H6, 2845), (L2H6, 2846), (L2H6, 3695), (L2H7, 1277), (L2H7, 1278), (L2H7, 1551), (L2H7, 1552), (L2H7, 1553), (L2H7, 1554), (L2H7, 1557), (L2H7, 1558), (L2H7, 1559), (L2H7, 1560), (L2H7, 1561), (L2H7, 2439), (L2H7, 2440), (L2H7, 2721), (L2H8, 1275), (L2H8, 1276), (L2H8, 1277), (L2H8, 1278), (L2H8, 1279), (L2H8, 1551), (L2H8, 1552), (L2H8, 1553), (L2H8, 1554), (L2H8, 1555), (L2H8, 1556), (L2H8, 1557), (L2H8, 1558), (L2H8, 1559), (L2H8, 1560), (L2H8, 1561), (L2H8, 2439), (L2H8, 2440), (L2H8, 2441), (L2H8, 2716), (L2H8, 2717), (L2H8, 2718), (L2H8, 2720), (L2H8, 2721), (L2H9, 1561), (L2H10, 1552), (L2H10, 1553), (L2H10, 1554), (L2H10, 1555), (L2H10, 1556), (L2H10, 1557), (L2H10, 1558), (L2H10, 1559), (L2H10, 1560), (L2H10, 1561), (L2H11, 1559), (L2H11, 1560), (L2H12, 1559), (L2H12, 1560), (L2H13, 1552), (L2H13, 1553), (L2H13, 1554), (L2H13, 1558), (L2H13, 1559), (L2H13, 1560), (L2H13, 1561).

(Layer 3) (L3H1, 534), (L3H1, 535), (L3H1, 666), (L3H1, 667), (L3H1, 668), (L3H1, 669), (L3H1, 736), (L3H1, 737), (L3H1, 738), (L3H1, 739), (L3H1, 740), (L3H1, 741), (L3H1, 742), (L3H1, 743), (L3H1, 744), (L3H2, 736), (L3H2, 737), (L3H2, 738), (L3H2, 739), (L3H2, 740), (L3H2, 741), (L3H2, 742), (L3H2, 743), (L3H2, 744), (L3H2, 745), (L3H2, 746), (L3H2, 747), (L3H3, 531), (L3H3, 532), (L3H3, 533), (L3H3, 534), (L3H3, 665), (L3H3, 666), (L3H3, 667), (L3H3, 734), (L3H3, 735), (L3H3, 736), (L3H3, 737), (L3H3, 738), (L3H3, 739), (L3H3, 740), (L3H3, 741), (L3H3, 742), (L3H3, 743), (L3H3, 744), (L3H3, 745), (L3H3, 2217), (L3H4, 532), (L3H4, 533), (L3H4, 534), (L3H4, 535), (L3H4, 667), (L3H4, 740), (L3H4, 741), (L3H4, 742), (L3H4, 743), (L3H5, 668), (L3H5, 737), (L3H5, 738), (L3H5, 739), (L3H5, 740), (L3H5, 741), (L3H5, 742), (L3H5, 743), (L3H5, 744), (L3H5, 745), (L3H5, 746), (L3H5, 747), (L3H5, 748), (L3H6, 736), (L3H6, 737), (L3H6, 738), (L3H6, 739), (L3H6, 740), (L3H6, 741), (L3H6, 742), (L3H6, 743), (L3H6, 744), (L3H6, 745), (L3H6, 746), (L3H6, 747), (L3H6, 748), (L3H6, 749), (L3H6, 2218), (L3H6, 2219), (L3H6, 2220), (L3H6, 2221), (L3H6, 2222), (L3H6, 2223), (L3H6, 2224), (L3H6, 2225), (L3H6, 2226), (L3H15, 3368), (L3H15, 3369), (L3H15, 3370), (L3H21, 922), (L3H21, 923), (L3H22, 1030), (L3H22, 1031), (L3H22, 1032), (L3H23, 920), (L3H23, 921), (L3H23, 922), (L3H23, 923), (L3H23, 924), (L3H23, 925), (L3H23, 1016), (L3H23, 1017), (L3H23, 1025), (L3H23, 1026), (L3H23, 1027), (L3H23, 1028), (L3H23, 1029), (L3H23, 1030), (L3H23, 1031), (L3H24, 735), (L3H24, 736), (L3H24, 742), (L3H24, 743), (L3H24, 804), (L3H24, 805), (L3H24, 806), (L3H24, 943), (L3H24, 944), (L3H24, 954), (L3H24, 955), (L3H24, 986), (L3H24, 987), (L3H24, 1014), (L3H24, 1015), (L3H24, 2684), (L3H24, 3712), (L3H24, 3713), (L3H24, 3714), (L3H24, 3715), (L3H25, 3714), (L3H25, 3715), (L3H25, 3716), (L3H26, 743), (L3H26, 922), (L3H26, 923), (L3H26, 924), (L3H26, 1013), (L3H26, 1014), (L3H26, 1015), (L3H26, 1016), (L3H26, 1017), (L3H26, 3713), (L3H26, 3714), (L3H27, 1015), (L3H27, 1016), (L3H27, 1017), (L3H27, 1028), (L3H27, 1029), (L3H27, 1030), (L3H27, 1031), (L3H27, 3712), (L3H27, 3713), (L3H27, 3714), (L3H27, 3715),

(Layer 4) (L4H10, 3839), (L4H10, 3923), (L4H10, 3956), (L4H10, 3957), (L4H10, 3980), (L4H10, 3987), (L4H18, 1094), (L4H18, 1754), (L4H18, 1755), (L4H23, 1258), (L4H23, 1684), (L4H23, 1685), (L4H23, 1741), (L4H23, 1742), (L4H23, 2018), (L4H23, 2276), (L4H23, 3532), (L4H23, 3960), (L4H25, 1740).

(Layer 6) (L6H0, 923), (L6H0, 1037), (L6H0, 1038), (L6H0, 2294), (L6H0, 2295), (L6H0, 2828), (L6H0, 3329), (L6H0, 3330), (L6H0, 3331), (L6H0, 3332), (L6H0, 3650), (L6H0, 3651), (L6H0, 3745), (L6H0, 3814), (L6H1, 1037), (L6H1, 1038), (L6H1, 2294), (L6H1, 2295), (L6H1, 3330), (L6H1, 3331), (L6H2, 922), (L6H2, 923), (L6H2, 924), (L6H2, 1037), (L6H2, 1038), (L6H2, 1452), (L6H2, 1905), (L6H2, 1974), (L6H2, 2293), (L6H2, 2294), (L6H2, 2295), (L6H2, 2827), (L6H2, 2828), (L6H2, 2941), (L6H2, 3330), (L6H2, 3331), (L6H2, 3425), (L6H2, 3745), (L6H2, 3746), (L6H2, 3814), (L6H2, 3853), (L6H3, 923), (L6H3, 924), (L6H3, 925), (L6H3, 931), (L6H3, 1037), (L6H3, 1038), (L6H3, 1039), (L6H3, 1358), (L6H3, 1359),

(L6H3, 1910), (L6H3, 1911), (L6H3, 2295), (L6H3, 2296), (L6H3, 2828), (L6H3, 2829), (L6H3, 2942), (L6H3, 2943), (L6H3, 3330), (L6H3, 3331), (L6H3, 3332), (L6H3, 3650), (L6H3, 3651), (L6H3, 3652), (L6H4, 1037), (L6H4, 2293), (L6H4, 2294), (L6H4, 2295), (L6H4, 2828), (L6H4, 2941), (L6H4, 3329), (L6H4, 3330), (L6H4, 3331), (L6H4, 3425), (L6H4, 3532), (L6H4, 3533), (L6H5, 922), (L6H5, 923), (L6H5, 2294), (L6H5, 3329), (L6H5, 3330), (L6H5, 3331), (L6H6, 919), (L6H6, 1038), (L6H6, 1239), (L6H6, 1974), (L6H6, 2294), (L6H6, 2295), (L6H6, 3212), (L6H6, 3331), (L6H6, 3425), (L6H6, 3426), (L6H6, 3532), (L6H6, 3533), (L6H6, 3746), (L6H6, 3853), (L6H25, 1239), (L6H25, 1268), (L6H25, 1269), (L6H25, 1270), (L6H25, 1445), (L6H25, 1446), (L6H26, 1269), (L6H26, 1270),

(Layers $10-11$) (L10H1, 967), (L10H1, 3204), (L10H1, 3217), (L10H1, 3694), (L10H1, 3695), (L10H22, 1094), (L10H22, 1095), (L11H3, 2161), (L11H3, 2162), (L11H14, 2089), (L11H14, 2090), (L11H14, 2091), (L11H18, 2090), (L11H18, 2091),

(Layers $21-14$) (L21H15, 2175), (L24H0, 880), (L24H0, 881), (L24H0, 1025), (L24H0, 1026), (L24H0, 1414), (L24H0, 1415), (L24H0, 2294), (L24H0, 2295), (L24H0, 3249), (L24H0, 3250), (L24H0, 3368), (L24H0, 3369), (L24H1, 880), (L24H1, 881), (L24H1, 1163), (L24H1, 1415), (L24H1, 2294), (L24H1, 2295), (L24H1, 3249), (L24H1, 3250), (L24H1, 3356), (L24H1, 3368), (L24H1, 3369), (L24H2, 879), (L24H2, 880), (L24H2, 1413), (L24H2, 1414), (L24H2, 2293), (L24H2, 2294), (L24H2, 3248), (L24H2, 3367), (L24H2, 3368), (L24H4, 880), (L24H4, 2294), (L24H4, 3249), (L24H4, 3250), (L24H5, 880), (L24H5, 881), (L24H5, 2294), (L24H5, 2295), (L24H5, 3248), (L24H5, 3249), (L24H5, 3250), (L24H6, 880), (L24H6, 881), (L24H6, 2294), (L24H6, 3249), (L24H6, 3250).

(Layers $25-26$) (L25H14, 2701), (L25H14, 2702), (L25H18, 2701), (L25H18, 2702), (L25H19, 2701), (L25H19, 2702), (L26H8, 610), (L26H8, 2294), (L26H8, 3494), (L26H8, 3532)

H.3.2. Attention Scores and Ranks of Q, K, and H

In this section, we present per-head statistics for SDHs with large Δ. The previous section shows that a single head can exhibit multiple slash patterns. For conciseness, we report at most one values of Δ for each head in each model, since slash patterns within the same head share the same Q, K, and H. Specifically, for each head, we rank all Δ by their average slash score and list the largest values in the following table.

Model	Δ	Head	$\mathbb{E}[S_{i,i-\Delta}](\times 10^{-3})$	OOD $\mathbb{E}[S_{i,i-\Delta}](\times 10^{-3})$	$r_1(Q)$	$R_{0.95}(Q)$	$r_1(K)$	$R_{0.95}(K)$	$r_1(H)$	$R_{0.95}(H)$
Llama3	563	L0H29	1.321	0.725	0.837	3	0.719	4	0.304	294
Llama3	685	L0H30	1.161	0.561	0.939	2	0.719	4	0.304	294
Llama3	3827	L1H16	1.021	4.490	0.922	3	0.847	11	0.386	227
Llama3	3827	L1H18	1.002	7.999	0.895	6	0.847	11	0.386	227
Qwen2.5	937	L0H7	3.064	6.664	0.748	46	0.999	1	0.001	3584
Qwen2.5	503	L0H23	2.049	1.326	0.941	3	0.996	1	0.160	590
Qwen2.5	503	L0H24	2.035	1.398	0.913	13	0.996	1	0.160	590
Qwen2.5	503	L0H26	2.077	1.436	0.910	14	0.996	1	0.160	590
Qwen2.5	804	L1H11	2.644	2.873	0.824	6	0.946	2	0.507	227
Qwen2.5	554	L1H15	2.207	2.106	0.875	6	0.975	1	0.002	3584
Qwen2.5	554	L1H16	2.969	3.222	0.778	9	0.975	1	0.507	227
Qwen2.5	610	L1H19	2.206	2.868	0.741	10	0.975	1	0.507	227
Qwen2.5	2293	L1H23	3.866	10.523	0.909	3	0.934	2	0.002	3584
Qwen2.5	835	L1H24	2.599	12.972	0.812	7	0.934	2	0.002	3584
Qwen2.5	1934	L2H6	2.651	8.729	0.906	9	0.645	51	0.001	3584
Qwen2.5	3713	L3H24	2.940	1.070	0.770	27	0.959	1	0.187	554
Qwen2.5	1685	L4H23	3.636	2.229	0.950	1	0.680	44	0.163	382
Qwen2.5	3331	L6H1	2.217	0.992	0.961	1	0.721	51	0.333	669
Qwen2.5	2294	L6H2	7.480	23.225	0.961	1	0.721	51	0.333	669
Qwen2.5	3651	L6H3	2.573	2.864	0.933	3	0.721	51	0.333	669
Qwen2.5	2294	L6H4	5.650	12.231	0.942	2	0.721	51	0.333	669
Qwen2.5	1974	L6H6	2.241	2.678	0.977	1	0.721	51	0.333	669
Qwen2.5	3694	L10H1	2.275	4.549	0.943	2	0.761	27	0.317	716
Qwen2.5	3369	L24H1	2.650	2.249	0.921	7	0.755	51	0.240	645

Table 6. This table lists the average attention scores of prompts in LongBench and OOD prompts for Llama3-8B-Instruct and Qwen2.5-7B-Instruct, the rank information of Q, K, and H. The unit of the attention scores is 10^{-3}.

Model	Head	$r_1(W_Q)$	$R_{0.95}(W_Q)$	$r_1(W_K)$	$R_{0.95}(W_K)$
Llama3	L0H29	0.339	24	0.344	25
Llama3	L0H30	0.487	26	0.344	25
Llama3	L1H16	0.044	69	0.033	72
Llama3	L1H18	0.035	66	0.033	72
Qwen2.5	L0H7	0.018	98	0.016	100
Qwen2.5	L0H23	0.031	118	0.022	116
Qwen2.5	L0H24	0.027	117	0.022	116
Qwen2.5	L0H26	0.025	117	0.022	116
Qwen2.5	L1H11	0.066	80	0.042	86
Qwen2.5	L1H15	0.058	73	0.042	80
Qwen2.5	L1H16	0.087	75	0.042	80
Qwen2.5	L1H19	0.080	73	0.042	80
Qwen2.5	L1H23	0.066	66	0.049	74
Qwen2.5	L1H24	0.095	65	0.049	74
Qwen2.5	L2H6	0.043	99	0.019	97
Qwen2.5	L3H24	0.030	105	0.021	101
Qwen2.5	L4H23	0.050	99	0.018	101
Qwen2.5	L6H1	0.021	108	0.015	103
Qwen2.5	L6H2	0.033	108	0.015	103
Qwen2.5	L6H3	0.017	108	0.015	103
Qwen2.5	L6H4	0.034	108	0.015	103
Qwen2.5	L6H6	0.023	107	0.015	103
Qwen2.5	L10H1	0.028	101	0.025	102
Qwen2.5	L24H1	0.032	112	0.015	111

Table 7. The table reports r_1 and $R_{0.95}$ for the query and key projection matrices, W_Q and W_K, across attention heads in LLMs.

Model	Head	$\tilde{r}_1(Q)$	$\tilde{R}_{0.95}(W_Q)$	$\tilde{r}_1(K)$	$\tilde{R}_{0.95}(K)$
Llama3	L0H29	0.904	3	0.877	4
Llama3	L0H30	.943	2	0.877	4
Qwen2.5	L0H7	0.221	68	0.079	73
Qwen2.5	L0H23	0.250	72	0.155	75
Qwen2.5	L0H24	0.218	73	0.155	75
Qwen2.5	L0H26	0.215	73	0.155	75

Table 8. The values of $\tilde{r}_1$ and $\tilde{R}_{0.95}$ for the SDHs located in the 0-th layer of Llama3-8B-Instruct, and Qwen2.5-7B-Instruct.

Model	Δ	Head	$\mathbb{E}[S_{i,i-\Delta}]$	$\mathbb{E}[S_{i,i-\Delta}]$ w/o high freqs	$\mathbb{E}[S_{i,i-\Delta}]$ w/o med freqs	$\mathbb{E}[S_{i,i-\Delta}]$ w/o low freqs
Llama3	563	L0H29	1.321	0.598	1.742	1.657
Llama3	685	L0H30	1.161	0.872	1.417	1.223
Llama3	3827	L1H16	1.021	0.710	0.974	2.466
Llama3	3827	L1H18	1.002	0.774	0.970	2.764
Qwen2.5	937	L0H7	3.093	1.523	1.522	3.164
Qwen2.5	503	L0H23	2.063	1.446	1.263	2.073
Qwen2.5	503	L0H24	2.063	1.456	1.230	2.060
Qwen2.5	503	L0H26	2.079	1.409	1.241	2.104
Qwen2.5	804	L1H11	2.717	0.831	2.808	2.800
Qwen2.5	554	L1H15	2.340	0.422	1.924	2.220
Qwen2.5	554	L1H16	3.018	0.975	2.712	2.975
Qwen2.5	610	L1H19	2.245	0.605	2.407	2.207
Qwen2.5	2293	L1H23	3.884	0.257	4.143	5.779
Qwen2.5	835	L1H24	2.773	0.169	6.733	2.754
Qwen2.5	1934	L2H6	2.617	1.348	2.842	2.670
Qwen2.5	3713	L3H24	2.957	1.448	1.974	2.916
Qwen2.5	1685	L4H23	3.638	0.651	10.284	3.637
Qwen2.5	3331	L6H1	2.253	0.710	2.934	2.226
Qwen2.5	2294	L6H2	7.417	1.360	13.771	7.481
Qwen2.5	3651	L6H3	2.666	0.832	7.652	2.583
Qwen2.5	2294	L6H4	5.698	1.466	9.087	5.638
Qwen2.5	1974	L6H6	2.264	0.422	4.368	2.241
Qwen2.5	3694	L10H1	2.179	0.555	4.799	4.768
Qwen2.5	3369	L24H1	2.566	0.613	2.914	2.655

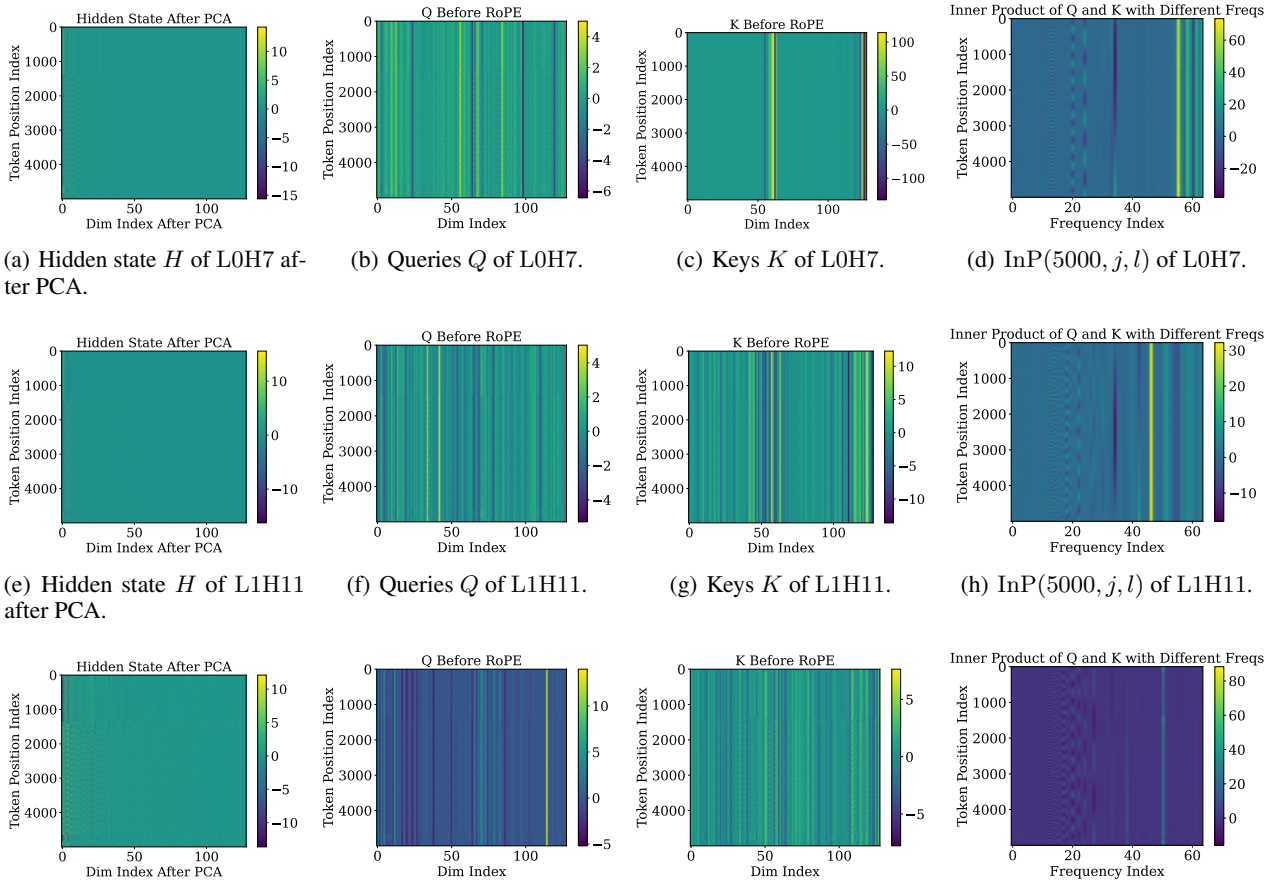
Table 9. This table quantifies the effect of low-, medium-, and high-frequency components on SDHs by reporting, for each band, the average slash score after removing that band. The unit of attention scores is 10^{-3}.

Head	Avg. $\|W_Q^\top \mathbf{x}\|$	$\|\boldsymbol{b}_Q\|$	Avg. $\|W_K^\top \mathbf{x}\|$	$\|\boldsymbol{b}_K\|$
L0H7	7.403	9.343	10.039	281.761
L0H23	5.632	17.083	9.173	108.957
L0H24	5.782	16.097	9.173	108.957
L0H26	5.810	15.828	9.173	108.957

Table 10. This table lists the average norms of $W_Q^\top \mathbf{x}$ and $W_K^\top \mathbf{x}$ across all the token embeddings and the norms of biases $\boldsymbol{b}_Q$ and $\boldsymbol{b}_K$

H.4. Figures of Slash-dominant Heads with Large Δ and Small Average Slash Score

H.4.1. Results of Qwen2.5



(a) Hidden state H of L0H7 after PCA.

(b) Queries Q of L0H7.

(c) Keys K of L0H7.

(d) $\mathrm{InP}(5000, j, l)$ of L0H7.

(e) Hidden state H of L1H11 after PCA.

(f) Queries Q of L1H11.

(g) Keys K of L1H11.

(h) $\mathrm{InP}(5000, j, l)$ of L1H11.

(i) Hidden state H of L2H6 after PCA.

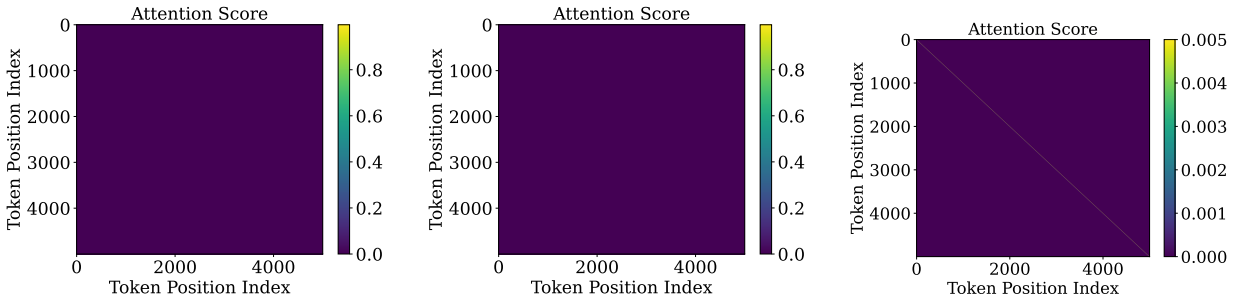
(j) Queries Q of L2H6.

(k) Keys K of L2H6.

(l) $\mathrm{InP}(5000, j, l)$ of L2H6.

Figure 30. This figure shows the hidden states, queries, keys, and $\mathrm{InP}(5000, j, l)$ for $j \in [5000], l \in [64]$ for Qwen2.5-7B-Instruct. Here, the example prompt contains 5000 tokens. The dimensions of queries and keys are both 128.

H.4.2. Results of Llama3-8B-Instruct

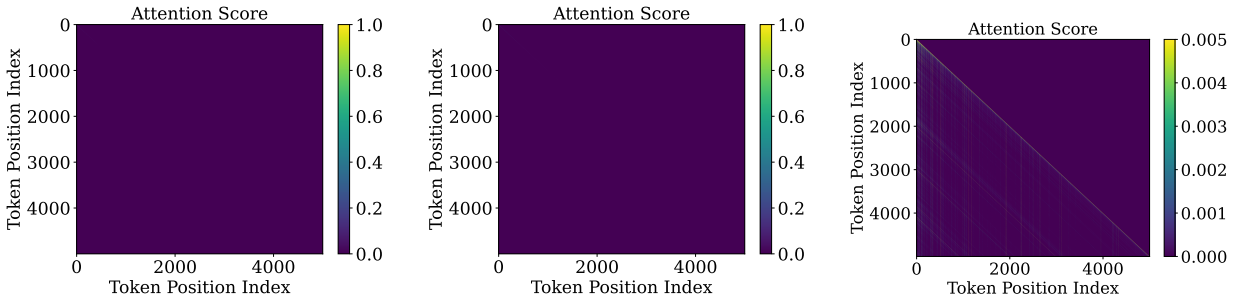


(a) Average attention score matrix of L0H29.

(b) Average attention score matrix of L0H30.

(c) Average attention score matrix of L1H16.

Figure 31. Average of attention score matrices in Llama3-8B-Instruct with prompts from LongBench V2.

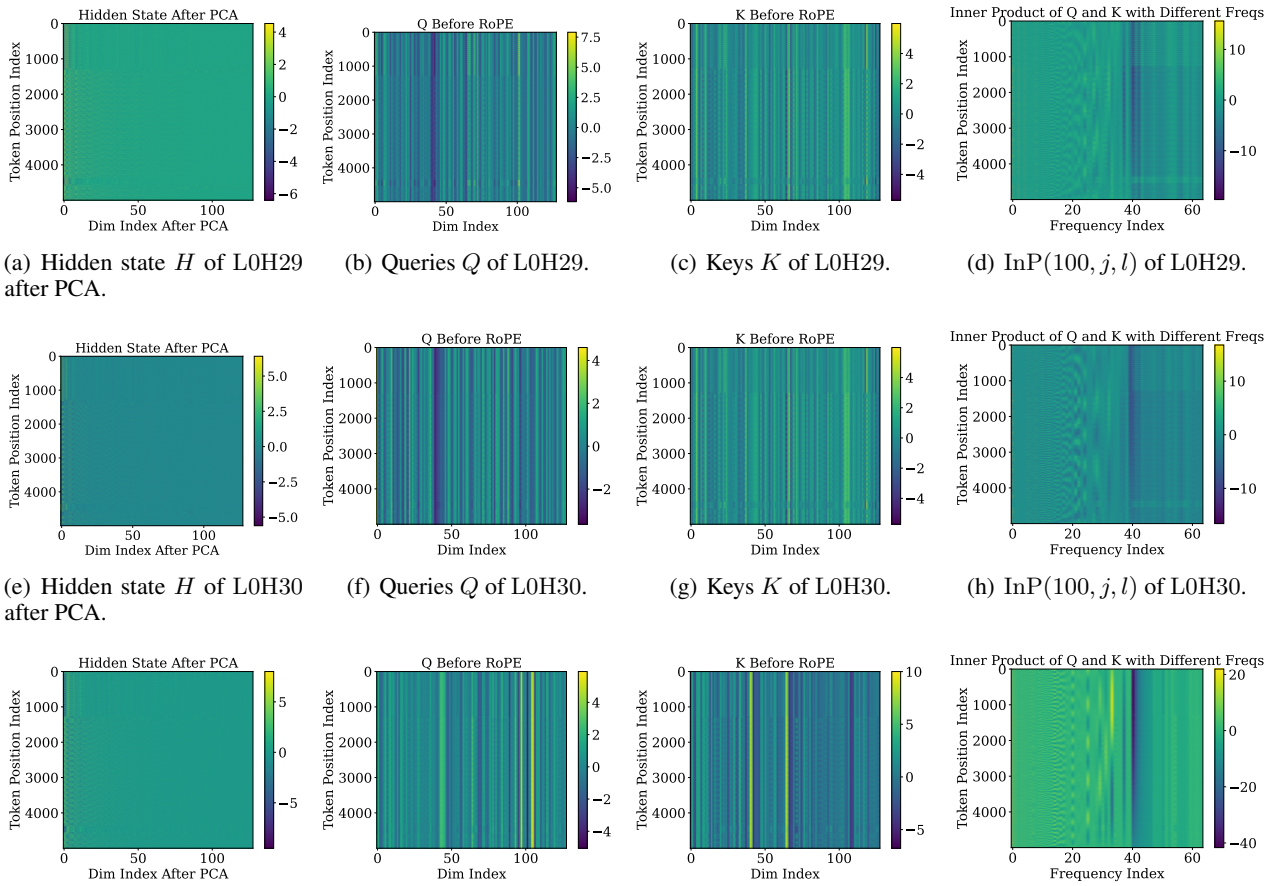


(a) Average attention score matrix of L0H29.

(b) Average attention score matrix of L0H30.

(c) Average attention score matrix of L1H16.

Figure 32. Average of attention score matrices in Llama3-8B-Instruct with prompts whose tokens are i.i.d. sampled from the uniform distribution on the alphabet.



(a) Hidden state H of L0H29 after PCA.

(b) Queries Q of L0H29.

(c) Keys K of L0H29.

(d) $\mathrm{InP}(100, j, l)$ of L0H29.

(e) Hidden state H of L0H30 after PCA.

(f) Queries Q of L0H30.

(g) Keys K of L0H30.

(h) $\mathrm{InP}(100, j, l)$ of L0H30.

(i) Hidden state H of L1H16 after PCA.

(j) Queries Q of L1H16.

(k) Keys K of L1H16.

(l) $\mathrm{InP}(100, j, l)$ of L0H1.

Figure 33. This figure shows the hidden states, queries, keys, and $\mathrm{InP}(100, j, l)$ for $j \in [100], l \in [64]$ for Llama3-8B-Instruct. Here, the example prompt contains 100 tokens. The dimensions of queries and keys are both 128.

Model	Head	RV($\mathbf{v}_Q$)	RV($\mathbf{v}_K$)	RV($\mathbf{v}_{\text{rand}}$)
Gemma	L0H4	8.0%	6.6%	953%
Gemma	L0H7	6.9%	6.2%	953%
Gemma	L0H1	4.3%	6.0%	953%
Llama3	L0H0	34.1%	24.8%	1312%
Llama3	L0H2	20.1%	24.8%	1312%
Llama3	L0H29	9.8%	14.6%	1312%
Llama3	L0H30	8.2%	14.6%	1312%

Table 11. This table reports the relative variation of token embeddings projected onto the dominant subspace of the SDHs in the 0-th layer of Gemma-7B and Llama3-8B-Instruct.

Head	$\|W_Q^\top \mathbf{h}_i\|$	$\|\mathbf{b}_Q\|$	$\|W_K^\top \mathbf{h}_i\|$	$\|\mathbf{b}_K\|$
L0H5	6.324	8.518	7.092	466.607
L0H6	6.439	13.522	7.092	466.607
L0H7	7.403	9.343	10.039	281.761
L0H15	12.112	160.007	7.469	237.963
L0H23	5.632	17.083	9.173	108.957
L0H24	5.782	16.097	9.173	108.957
L0H26	5.810	15.828	9.173	108.957

Table 12. This table lists the average norms of $\|W_Q^\top \mathbf{h}_i\|$ and $\|W_K^\top \mathbf{h}_i\|$ over the alphabet, together with the norms of $\|\mathbf{b}_Q\|$ and $\|\mathbf{b}_K\|$ for SDH in the 0-th layer of Qwen2.5-7B-Instruct.

H.5. Tables related to token embeddings in the 0-th layer

I. Notations in Theory Sections and Expression of Reduced Model

Notations	Descriptions
$A_{i,j}$	The logit from position j to i before softmax in *Layer 1* ((22)).
$S_{i,j}^{(1)}$	The attention score from position j to i in *Layer 1*.
$A, S^{(1)} \in \mathbb{R}^{N\times N}$	The attention logit and attention score matrix in *Layer 1*.
$S_{i,\mathbf{x}}^{(1)}$	Attention scores on $\mathbf{x}$, i.e., odd position: $\sum_{j\le i} \mathbb{1}\{j \equiv 1 (\text{mod } 2)\} S_{i,j}^{(1)}$.
$S_{i,y}^{(1)}$	Attention scores on y, i.e., even position: $\sum_{j\le i} \mathbb{1}\{j \equiv 0 (\text{mod } 2)\} S_{i,j}^{(1)}$.
$S_i^{(2)}, \mathcal{S}_k^{(2)}$	Attention scores to i-token and k-th feature in *Layer 2* (Equation (17)).
$\overline{S}_i^{(2)}, \overline{\mathcal{S}}_k^{(2)}$	Approximate expectation version of $S_i^{(2)}, \mathcal{S}_k^{(2)}$ in Stage I (Theorem J.6).
$\mathbf{u}, S^{(2)} \in \mathbb{R}^{1\times N}$	The attention logit and score vector of the question E_{q} in *Layer 2*.
$\mathbf{u}_i$	The attention logit from E_{q} to E_i in *Layer 2* ((24) and (48)).
$\tilde{\mathbf{c}}$	Constant vector $(1, 0, \cdots, 1, 0)^\top$.

Table 14. Summary of frequently used notations in proof. We omit (t) here for simplicity.

Notations. For two functions f and g of n, we write $f(n) = O(g(n))$ or $f(n) \lesssim g(n)$ if $g(n) \ge 0$, and there exist constants $C, n_0 > 0$ such that for all $n \ge n_0$, $f(n) \le C\, g(n)$, and $f(n) = -O(g(n))$ or $f(n) \gtrsim -g(n)$ if $g(n) \ge 0$, and there exist constants $C > 0$ and n_0 such that for all $n \ge n_0$, $f(n) \ge -C\, g(n)$. Similarly, we write $f(n) = \Omega(g(n))$ or $f(n) \gtrsim g(n)$, if $g(n) \ge 0$, and there exist constants $C > 0$ and n_0 such that for all $n \ge n_0$, $f(n) \ge C\, g(n)$, and we write $f(n) = -\Omega(g(n))$ or $f(n) \lesssim -g(n)$ if $g(n) \le 0$, and there exist constants $C > 0$ and n_0 such that for all $n \ge n_0$, $f(n) \le -C\, g(n)$. We write $f(n) = \Theta(g(n))$ or $f(n) \simeq g(n)$ if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$.

In addition, we abuse the notation $\mathbb{1}\{j = 2r-1\}$ to indicate that for a given j, if there exists $r \in \mathbb{Z}$, s.t. $j = 2r-1$, then it takes 1, and 0 otherwise. For simplicity, we further denote different sub-matrices of E as $E^{\mathbf{c}} = E_{:,1:d_{\mathbf{c}}} \in \mathbb{R}^{N\times d_{\mathbf{c}}}$, $E^{\mathbf{x},y} = E_{:,d_{\mathbf{c}}+1:d} \in \mathbb{R}^{N\times(d_{\mathcal{X}}+2)}$, $E^y = E_{:,d} \in \mathbb{R}^N$. For each i, we write $E_i^{\mathbf{x},y} = E_{i,:}^{\mathbf{x},y} \in \mathbb{R}^{1\times(d+2)}$, and we denote the last row by $E_{\text{q}}^{\mathbf{x},y} = E_{N,:}^{\mathbf{x},y} \in \mathbb{R}^{1\times(d+2)}$. We summarize additional frequently used notations related to transformers in Table 14, and we omit the timestep (t) for simplicity when there is no ambiguity.

Expression of Reduced Model. We first visualize the reduction of Layer 1 and 2 weights in Figures 34 and 35.

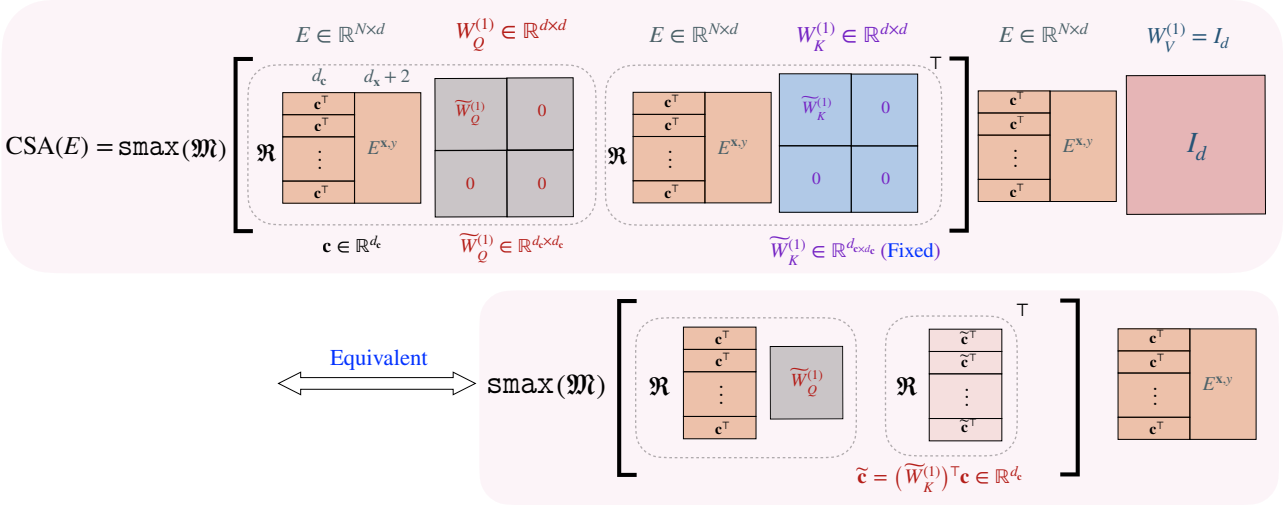


Figure 34. Illustration of the Reduction of Layer 1. In Figures 34 and 35, the softmax operator is abbreviated as sf. The output of Layer 1 is $H^{(1)} = [E, \mathrm{CSA}(E)]$. In the figure, we show the reduction of Layer 1 parameters $W^{(1)}_{\{Q,K,V\}}$ and how $\mathrm{CSA}(E)$ is generated by the reduced parameters. In particular, the weight matrices are only non-zero in the semantically independent subspace associated with $\mathbf{c}$, and the key and value weight matrices $W^{(1)}_{\{K,V\}}$ are fixed during training. Hence, $\mathrm{CSA}(E)$ only depends on $\widetilde{W}_Q^{(1)}$.

Then we finally write the expression of the prediction under the reduced model as follows:

$$\hat{y}_{\mathrm{q}}(E;\tilde{\theta}) = \mathrm{softmax}\left(\Re_{\boldsymbol{\vartheta}_{d_{\mathbf{c}}+1:d}}(E_{\mathrm{q}}^{\mathbf{x},y}\widetilde{W}_Q^{(2)})\Re_{\boldsymbol{\vartheta}_{d_{\mathbf{c}}+1:d}}\left(\mathrm{softmax}\left(A\right)E^{\mathbf{x},y}\right)^\top\right)E^y, \tag{22}$$

where $A_{i,j} = \mathbf{c}^\top \widetilde{W}_Q^{(1)} R_{\boldsymbol{\vartheta}_{1:d_{\mathbf{c}}},j-i}\tilde{\mathbf{c}}$ for $j \le i$, and $-\infty$ otherwise, and the RoPE operator $\Re_{\boldsymbol{\vartheta}_{d_{\mathbf{c}}+1:d}}$ and $R_{\boldsymbol{\vartheta}_{1:d_{\mathbf{c}}},j-i}$ are defined in (3).

J. Proof of Stage I of Theorem D.3

J.1. Roadmap of the Proof

We analyze the emergence of slash-dominance in Stage I via three phases of dynamics (Sections J.4 to J.6). In each phase, we formulate an induction hypothesis and derive several lemmas describing the values and evolution of key statistics, which collectively govern the attention scores. We prove the induction hypothesis and the lemmas by induction.

The main idea of the proof lies in tracking the update dynamics of the attention logit in layer 1 $A_{l,r}(t)$, which decides the Layer 1 attention score $S^{(1)}_{l,r}(t)$. From Theorem J.2, we find that for any $0 \le r \le l \le N$, the update of $A_{l,r}(t)$: $\Delta A_{l,r}(t) = A_{l,r}(t+1) - A_{l,r}(t)$ can be written as:

$$\Delta A_{l,r}(t) = \eta_1 \cdot \Big(C_1 \sum_{i=l-r+1}^{N} a_{i,i+r-l}(t) + C_2 \sum_{1\le j\le i\le N} a_{i,j}(t) \pm \sum_{1\le j\le i\le N} a_{i,j}(t) O(\epsilon_{\mathrm{FN}})\Big),$$

where we denote

$$a_{i,j}(t) = \mathbb{E}\Big[\left(\hat{y}_{\mathrm{q}} - \langle \mathbf{w}, \mathbf{x}_{\mathrm{q}}\rangle\right) S_i^{(2)}(t)(\hat{y}_{\mathrm{q}} - E_i^y) S^{(1)}_{i,j}(t)(I(i)_j - \sum_{\ell\le i} S^{(1)}_{i,\ell}(t) I(i)_\ell)\Big].$$

The value of $\Delta A_{l,r}(t)$ largely depends on $a_{i,j}(t)$ for all $i,j \in [N]$. The expression of $a_{i,j}(t)$ can be further simplified by considering different cases of (i,j), as shown in Theorem J.3, where we demonstrate that $a_{i,j}(t)$ is determined by the attention scores in Layer 1 and Layer 2, $S^{(1)}_{i,j}(t)$ and $\mathcal{S}_k^{(2)}(t)$. In addition, from Theorem J.7, $\mathcal{S}_k^{(2)}(t) = \Theta(1/K)$ in stage I. A schematic of the interaction of statistics is shown in Figure 36. In addition, from the interaction above, we observe that for

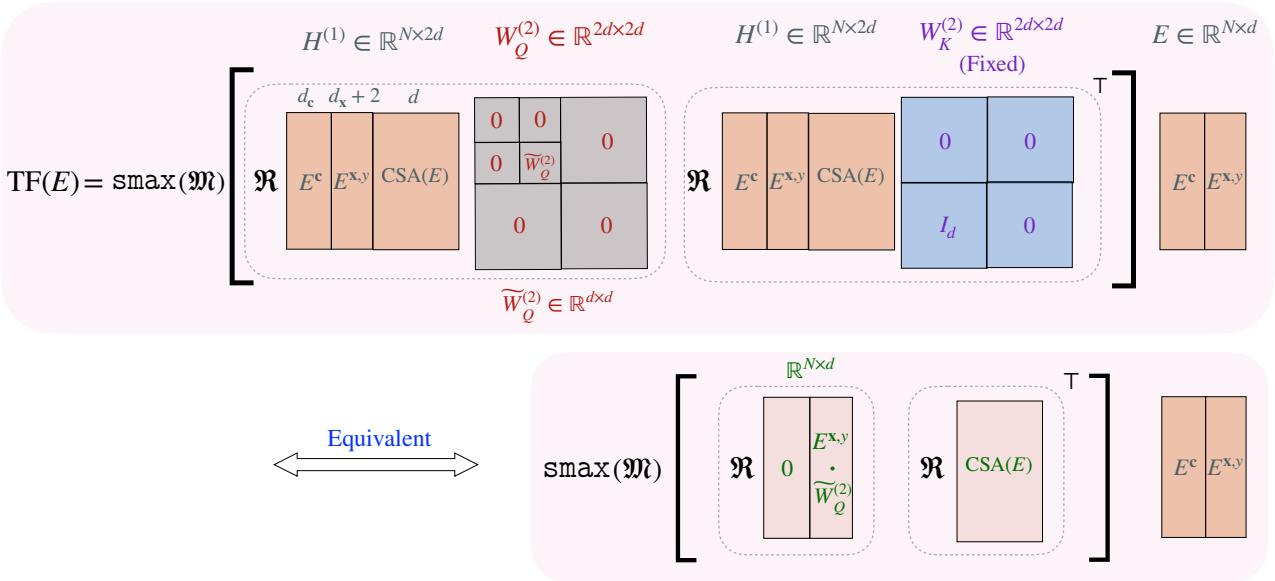


Figure 35. Illustration of the Reduction of Layer 2 and Transformer Output. First, with reduced $W_V^{(2)} = I_{2d}$ and $W_O = [0_{d\times d} \quad 0_{d\times d} \quad I_d \quad 0_{d\times d}]^\top$, the CSA output $\mathrm{CSA}(H^{(1)}) = S^{(2)}H^{(1)}W_V^{(2)} = S^{(2)}H^{(1)} \in \mathbb{R}^{N\times 2d}$, Layer 2 output $H^{(2)} = [E, \mathrm{CSA}(E), \mathrm{CSA}(H^{(1)})] \in \mathbb{R}^{N\times 4d}$, and the transformer output $\mathrm{TF}_\theta(E) = H^{(2)}W_O = S^{(2)}E$. Here, $S^{(2)}$ is the Layer 2 attention score matrix. In this figure, we show the reduction of $S^{(2)}$ under sparse trainable query weight $W_Q^{(2)}$ and fixed key weight $W_K^{(2)}$. Hence, the output $\mathrm{TF}_\theta(E)$ only depends on $\widetilde{W}_Q^{(1)}, \widetilde{W}_Q^{(2)}$.

different $i, j \in [N]$, the attention scores $S_{i,j}^{(1)}(t)$ with the same offset $i-j$ are of the same order. The similar conclusion also holds for $S_{l,r}^{(1)}(t)$, $A_{l,r}^{(1)}(t)$ and the increments $\Delta A_{l,r}^{(1)}(t)$.

The learning process can be divided into three phases. Throughout all three phases, for any $l \in [N]$, the *attention logit* corresponding to the immediately preceding token, $A_{l,l-1}(t)$, keeps growing, although its growth rate, denoted by $\Delta A_{l,l-1}(t)$, varies across the phases. In contrast, other logits $A_{l,r}(t), r \neq l-1$ oscillate, typically with much smaller rate than the growth rate of $A_{l,l-1}(t)$. As a result, $S_{l,l-1}^{(1)}(t)$ keeps increasing and exhibits different orders of magnitude across the three phases as described below.

- **Phase I: Emergence of Slash-Dominance.** ($t \in [0, T_1^{(1)}]$, Section J.4). During phase I, for any $l \in [N]$, the *attention scores* $S_{l,r}^{(1)}$ are relatively uniform across r: $S_{l,l-1}^{(1)}(t) = \Omega\left(l^{-1}\right)$, and $S_{l,r}^{(1)}(t) = O\left(l^{-1}\right), r \neq l-1$. As for the *attention logits*, $A_{l,l-1}(t)$ keeps growing, but $A_{l,r}(t), r \neq l-1$ may oscillate. In addition, the growth rate $\Delta A_{l,l-1}(t)$ is much higher than that of other $A_{l,r}(t)$ with $r \neq l-1$, since for any $i \in [N]$, $a_{i,i-1}$ is always positive and has a much larger order of magnitude than $a_{i,j}$ with $j \neq i$. Specifically, from Theorem J.10, $\Delta A_{l,l-1}(t) - \max_{r\leq l, r\neq l-1} \Delta A_{l,r}(t) \gtrsim \eta_1 C_1 K^{-1} N^{-1}$. Therefore, the increase in $A_{l,l-1}^{(t)}$ dominates the learning dynamics during phase I.

- **Phase II: Rapip Growth of Slash-Dominance.** ($t \in (T_1^{(1)}, T_2^{(1)}]$, Section J.5). After rapid growth of $A_{l,l-1}(t)$ in phase I, $S_{l,l-1}^{(1)}(t) = \Omega(1)$ grows to a constant order. However, for other $r \neq l-1$, $S_{l,r}^{(1)}(t) \leq O\left(N^{-1}\right)$. The *attention scores* $S_{l,r}^{(1)}$ is no longer uniform. The increase in $A_{l,l-1}^{(t)}$ still dominates the learning dynamics during phase II, and the growth rate gets much larger than Phase I. Specifically, from Theorem J.15, $\Delta A_{l,l-1}(t) - \max_{r\leq l, r\neq l-1} \Delta A_{l,r}(t) \gtrsim \eta_1 C_1 K^{-2}$.

- **Phase III: Convergence** ($t \in (T_2^{(1)}, T_3^{(1)}]$, Section J.6). for any $l \in [N]$, $A_{l,l-1}(t)$ keep growing but in a smaller rate. As a result, $S_{l,l-1}^{(1)}(t)$ keeps growing but can not exceed $1-\epsilon_1$. Finally, after the end of phase III, at step $t = T_3^{(1)}+1$, $S_{l,l-1}^{(1)}(t)$ finally exceeds $1-\epsilon_1$.

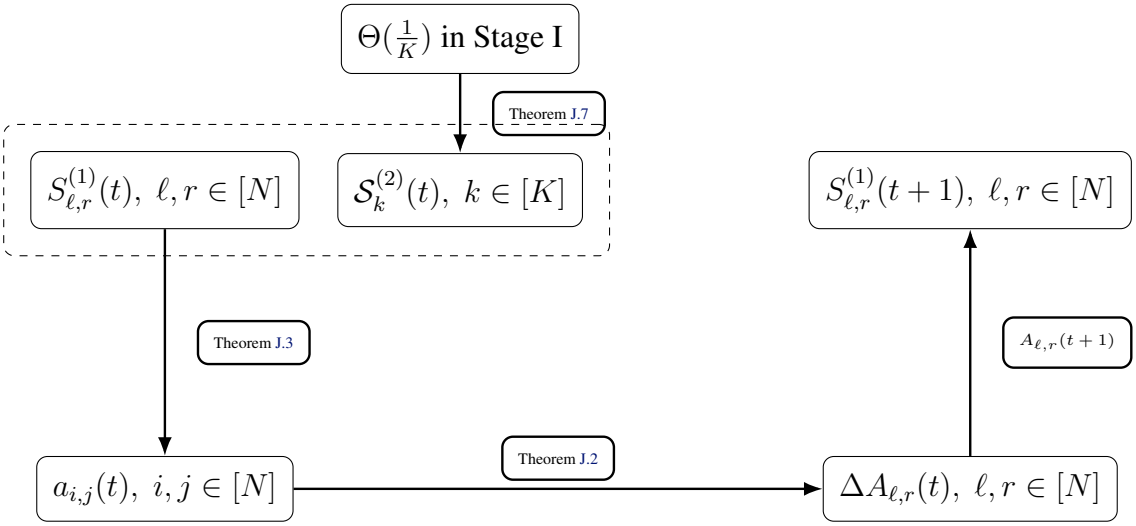


Figure 36. Schematic of dependencies among key statistics and attention scores at timestep t and $t+1$ in Stage I.

We summarize the upcoming sections as follows: In Section J.2, we compute and simplify the gradients to identify the key update variables. In Section J.3, we introduce several useful auxiliary lemmas for Stage I. In Sections J.4 to J.6, we analyze the three phases of the dynamics.

J.2. Stage I: Preliminary Development

Computations of Gradients. We first calculate Stage I gradient with respect to $\widetilde{W}_Q^{(1)}$. We omit (t) in this section when there is no ambiguity and write $L(\tilde{\theta})$ as L here for simplicity.

Lemma J.1 (Layer 1 Gradient)**.** *In Stage I, where $\widetilde{W}_Q^{(2)}$ is kept as I_d. The gradient of the loss function with respect to $\widetilde{W}_Q^{(1)}$ is given by*

$$\nabla_{\widetilde{W}_Q^{(1)}} L = \mathbb{E}\Bigg[(\hat{y}_{\mathrm{q}} - \langle \mathbf{w}, \mathbf{x}_{\mathrm{q}} \rangle) \sum_{j \le i \le N} S_i^{(2)} \left(E_i^y - \hat{y}_{\mathrm{q}}\right) S_{i,j}^{(1)} \Big(I(i)_j - \sum_{\ell \le i} S_{i,\ell}^{(1)} I(i)_\ell \Big) \mathbf{c}\tilde{\mathbf{c}}^\top R_{\boldsymbol{\vartheta},i-j} \Bigg],$$

where $I(i) = E^{\mathbf{x},y} R_{\boldsymbol{\vartheta},N-i} (E_{\mathrm{q}}^{\mathbf{x},y})^\top \in \mathbb{R}^N$, capturing the correlation of the question embedding with the prompt embeddings. The notation $I(i)_\ell$ refers to the ℓ-th entry of $I(i)$, and $E^{\mathbf{x},y}, E_{\mathrm{q}}^{\mathbf{x},y}$ correspond to the semantic dependent subspaces of the token embeddings E and E_{q}, respectively.

Proof We first write out the expression of $\hat{y}_{\mathrm{q}}$. In Stage I, $\widetilde{W}_Q^{(2)}$ is fixed as I_d, then

$$\hat{y}_{\mathrm{q}}(\widetilde{W}_Q^{(1)}, \widetilde{W}_Q^{(2)}) = \sum_{i=1}^{N_{\mathrm{in}}} S_{2i}^{(2)} \cdot y_i, \tag{23}$$

where $S^{(2)} = \mathrm{softmax}(\mathbf{u}) \in \mathbb{R}^{1 \times N}$, and

$$\mathbf{u}_i = S_{i,:}^{(1)} I(i) = (S_{i,:}^{(1)}) E_{\mathrm{q}}^{\mathbf{x},y} R_{\boldsymbol{\vartheta},i-N} (E^{\mathbf{x},y})^\top. \tag{24}$$

Here $S_{i,:}^{(1)}$ is the i-th row of the attention score matrix. Then we can compute the gradient. We first obtain:

$$\nabla_{\widetilde{W}_Q^{(1)}} L = \mathbb{E}[(\hat{y}_{\mathrm{q}} - \langle \mathbf{w}, \mathbf{x}_{\mathrm{q}} \rangle) \nabla_{\widetilde{W}_Q^{(1)}} \hat{y}_{\mathrm{q}}] = \mathbb{E}\Big[(\hat{y}_{\mathrm{q}} - \langle \mathbf{w}, \mathbf{x}_{\mathrm{q}} \rangle) \sum_{\ell \in [N_{\mathrm{in}}]} (\nabla_{\widetilde{W}_Q^{(1)}} S_{2\ell}^{(2)}) y_\ell \Big], \tag{25}$$

where the last equation follows from (23). We next compute $\nabla_{\widetilde{W}_Q^{(1)}} S_{2\ell}^{(2)}$ by chain rule. Recall that for any $U \in \mathbb{R}^{1 \times d}$,

$\nabla_U \mathrm{softmax}(U) = \mathrm{diag}(\mathrm{softmax}(U)) - \mathrm{softmax}(U)^\top \mathrm{softmax}(U)$, we have

$$\frac{\partial S_{2\ell}^{(2)}}{\partial S_{i,n}^{(1)}} = \sum_{m=1}^{N} \frac{\partial S_{2\ell}^{(2)}}{\partial \mathbf{u}_m} \frac{\partial \mathbf{u}_m}{\partial S_{i,n}^{(1)}} \overset{(i)}{=} \frac{\partial S_{2\ell}^{(2)}}{\partial \mathbf{u}_i} \frac{\partial \mathbf{u}_i}{\partial S_{i,n}^{(1)}} \overset{(ii)}{=} (S_i^{(2)} \mathbb{1}\{2\ell = i\} - S_i^{(2)} S_{2\ell}^{(2)}) \cdot I(i)_n, \tag{26}$$

where (i) follows from that $\partial \mathbf{u}_m / \partial S_{i,n}^{(1)} = 0$ if $m \neq i$, and (ii) follows from the definition of $\mathbf{u}_i$ in (24). In addition,

$$\frac{\partial S_{i,n}^{(1)}}{\partial A_{i,j}} = S_{i,j}^{(1)} \mathbb{1}\{n = j\} - S_{i,j}^{(1)} S_{i,n}^{(1)}. \tag{27}$$

Then combine (26) and (27), we have

$$\frac{\partial S_{2\ell}^{(2)}}{\partial A_{i,j}} \overset{(i)}{=} \sum_{n \leq i} \frac{\partial S_{2\ell}^{(2)}}{\partial S_{i,n}^{(1)}} \cdot \frac{\partial S_{i,n}^{(1)}}{\partial A_{i,j}} = S_i^{(2)} (\mathbb{1}\{2\ell = i\} - S_{2\ell}^{(2)}) S_{i,j}^{(1)} \Big(I(i)_j - \sum_{\ell \leq i} S_{i,\ell}^{(1)} I(i)_\ell \Big), \tag{28}$$

where (i) follows from the chain rule and the definition of causal mask. In addition, recall, if $j \leq i$, then $A_{i,j} = \mathbf{c}^\top \widetilde{W}_Q^{(1)} R_{\boldsymbol{\vartheta}, j-i} \tilde{\mathbf{c}}$, and we have

$$\nabla_{\widetilde{W}_Q^{(1)}} A_{i,j} = \mathbf{c} \tilde{\mathbf{c}}^\top (R_{\boldsymbol{\vartheta}, j-i})^\top = \mathbf{c} \tilde{\mathbf{c}}^\top R_{\boldsymbol{\vartheta}, i-j}. \tag{29}$$

As a result, combine (28) and (29), we have

$$\nabla_{\widetilde{W}_Q^{(1)}} S_{2\ell}^{(2)} = \sum_{j \leq i \leq N} S_i^{(2)} (\mathbb{1}\{2\ell = i\} - S_{2\ell}^{(2)}) S_{i,j}^{(1)} \Big(I(i)_j - \sum_{\ell \leq i} S_{i,\ell}^{(1)} I(i)_\ell \Big) \mathbf{c} \tilde{\mathbf{c}}^\top R_{\boldsymbol{\vartheta}, i-j}.$$

Then, by direct calculation, we have

$$\sum_{\ell \in [N_{\mathrm{in}}]} (\nabla_{\widetilde{W}_Q^{(1)}} S_{2\ell}^{(2)}) \cdot y_\ell = \sum_{j \leq i \leq N} S_i^{(2)} (E_i^y - \hat{y}_{\mathrm{q}}) S_{i,j}^{(1)} \Big(I(i)_j - \sum_{\ell \leq i} S_{i,\ell}^{(1)} I(i)_\ell \Big) \mathbf{c} \tilde{\mathbf{c}}^\top R_{\boldsymbol{\vartheta}, i-j}. \tag{30}$$

As a result, plugging (30) into (25) proves Theorem J.1. Thus, we conclude the proof of Lemma J.1. ■

Computations of Logits Update. In stage I, the logits $A_{l,r}(t)$ determine the attention scores. Hence, to get the dynamics of attention scores, we only need to track the update of $A_{l,r}(t) : \Delta A_{l,r}(t) = A_{l,r}(t+1) - A_{l,r}(t)$ as follows.

Lemma J.2 (Track $A_{l,r}(t)$ Update)**.** *Suppose Assumption 4.1 holds. In Stage I, for any $0 \leq r \leq l \leq N$, we have*

$$\Delta A_{l,r}(t) = \eta_1 \Big(C_1 \sum_{i=l-r+1}^{N} a_{i,i+r-l}(t) + C_2 \sum_{1 \leq j \leq i \leq N} a_{i,j}(t) + \sum_{1 \leq j \leq i \leq N} (\pm a_{i,j}(t)) O(\epsilon_{\mathrm{FN}}) \Big),$$

where $C_1, C_2, \epsilon_{\mathrm{FN}}$ are parameters in Assumption 4.1, and

$$a_{i,j}(t) = \mathbb{E}\Big[(\hat{y}_{\mathrm{q}} - \langle \mathbf{w}, \mathbf{x}_{\mathrm{q}} \rangle) S_i^{(2)}(t) (\hat{y}_{\mathrm{q}} - E_i^y) S_{i,j}^{(1)}(t) (I(i)_j - \sum_{\ell \leq i} S_{i,\ell}^{(1)}(t) I(i)_\ell) \Big].$$

Proof For any $1 \leq r \leq l \leq N$, $A_{l,r}(t) = \mathbf{c}^\top \widetilde{W}_Q^{(1)}(t) R_{\boldsymbol{\vartheta}, r-l} \tilde{\mathbf{c}}$. Considering the GD update rule and applying Theorem J.1, we have

$$\Delta A_{l,r}(t) = -\eta_1 \mathbf{c}^\top (\nabla_{\widetilde{W}_Q^{(1)}} L(t)) R_{\boldsymbol{\vartheta}, r-l} \tilde{\mathbf{c}} = -\eta_1 \sum_{j \leq i \leq N} a_{i,j}(t) \left(\tilde{\mathbf{c}}^\top R_{\boldsymbol{\vartheta}, i-j+r-l} \tilde{\mathbf{c}} \right). \tag{31}$$

Following from Assumption 4.1, we have

$$\tilde{\mathbf{c}}^\top R_{\boldsymbol{\vartheta}, i-j+r-l} \tilde{\mathbf{c}} = \sum_{s=1}^{d_{\mathbf{c}}/2} \cos\left(\theta_s (i - j + r - l)\right) = C_1 \delta_0 (i - j + r - l) + C_2 \pm O(\epsilon_{\mathrm{FN}}). \tag{32}$$

By plugging (32) into (31), we conclude the proof of Theorem J.2. ■

Simplification of Logits Update. From Theorem J.2, we find that the update $\Delta A_{r,l}(t)$ depends significantly on $a_{i,j}(t)$. However, the expression of $a_{i,j}(t)$ is quite complicated and depends on i, j. To simplify it, we next characterize $a_{i,j}(t)$ by cases in the next lemma. We omit (t) in Theorem J.3 for abbreviation. The key auxiliary lemmas used in the characterization of $a_{i,j}(t)$ are deferred to Section J.3.

Lemma J.3 (Characterization of $a_{i,j}$)**.** *In Stage I, consider $\overline{S}_i^{(2)}$ and $\overline{\mathcal{S}}_k^{(2)}$ are approximate expectation versions of $S_i^{(2)}$ and $\mathcal{S}_k^{(2)}$ for any $i \in [N]$ and $k \in [K]$. For different cases of $i \in [N], j \le i$, the following holds.*

1. *If $i = 2n$ and $j = 2n-1$ for some $n \in [N_{\text{in}}]$, then we have*

$$a_{i,j} \simeq \frac{S_{i,j}^{(1)}}{KN} \sum_{k=1}^{K} \Big\{ \mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\text{q}} = \mathbf{v}_k\} \Big\{ \Big(\frac{K-1}{K}(1-\overline{\mathcal{S}}_k^{(2)}) + \frac{1}{K}\sum_{o\neq k}^{K} \overline{\mathcal{S}}_o^{(2)} \Big) \Big(S_{i,\mathbf{x}}^{(1)} - S_{i,i-1}^{(1)} \Big) + \Big(\sum_{o=1}^{K} (\overline{\mathcal{S}}_o^{(2)})^2 - 2\overline{\mathcal{S}}_k^{(2)} + 1 \Big) S_{i,y}^{(1)} \Big\} \Big] \pm O\Big(\frac{1 - S_{i,i-1}^{(1)}}{K} \Big(\frac{\log N}{N^2} + \frac{1}{N^{\alpha-1}} \Big) \Big) \Big\} \pm O\Big(\frac{1}{N^3}\Big).$$

2. *If $i = 2n$ and $j = 2m-1$ for some $m < n \le N_{\text{in}}$, then we have*

$$a_{i,j} \simeq \frac{S_{i,j}^{(1)}}{KN} \sum_{k=1}^{K} \Big\{ \mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\text{q}} = \mathbf{v}_k\} \Big\{ \Big(\sum_{o=1}^{K} (\overline{\mathcal{S}}_o^{(2)})^2 - \overline{\mathcal{S}}_k^{(2)} - \frac{\sum_{o=1}^{K} \overline{\mathcal{S}}_o^{(2)}}{K} + \frac{1}{K} \Big) \Big(S_{i,y}^{(1)} + S_{i,i-1}^{(1)} \Big) - \Big(\sum_{o=1}^{K} (\overline{\mathcal{S}}_o^{(2)})^2 - 2\overline{\mathcal{S}}_k^{(2)} + 1 \Big) S_{i,i-1}^{(1)} \Big\} \Big] \pm O\Big(\frac{(1-S_{i,j}^{(1)})\log(N)}{N^2} \Big) \pm O\Big(\frac{1-S_{i,j}^{(1)}}{KN^{\alpha-1}} \Big) \Big\} \pm O\Big(\frac{1}{N^3}\Big).$$

3. *If $i = 2n$ and $j = 2m$ for some $m \le n \le N_{\text{in}}$, then we have*

$$a_{i,j} \simeq \frac{S_{i,j}^{(1)}}{KN} \sum_{k=1}^{K} \Big\{ \mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\text{q}} = \mathbf{v}_k\} \Big\{ \Big(\sum_{o=1}^{K} (\overline{\mathcal{S}}_o^{(2)})^2 - \overline{\mathcal{S}}_k^{(2)} - \frac{\sum_{o=1}^{K} \overline{\mathcal{S}}_o^{(2)}}{K} + \frac{1}{K} \Big) \Big(-\Big(S_{i,\mathbf{x}}^{(1)} - S_{i,i-1}^{(1)} \Big) \Big) \Big(\sum_{o=1}^{K} (\overline{\mathcal{S}}_o^{(2)})^2 - 2\overline{\mathcal{S}}_k^{(2)} + 1 \Big) \Big(-S_{i,i-1}^{(1)} \Big) \Big\} \Big] \pm O\Big(\frac{S_{i,\mathbf{x}}^{(1)} \log(N)}{N^2} \Big) \pm O\Big(\frac{S_{i,\mathbf{x}}^{(1)}}{KN^{\alpha-1}} \Big) \Big\} \pm O\Big(\frac{1}{N^3}\Big).$$

4. *If $i = 2n-1$ and $j = 2m$ for some $m \le n-1 \le N_{\text{in}} - 1$, then we have*

$$a_{i,j} \simeq \frac{S_{i,j}^{(1)}}{KN} \sum_{k=1}^{K} \Big\{ \mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\text{q}} = \mathbf{v}_k\} \Big(\sum_{o=1}^{K} (\overline{\mathcal{S}}_o^{(2)})^2 - \overline{\mathcal{S}}_k^{(2)} \Big) \Big(-S_{i,\mathbf{x}}^{(1)} \Big) \Big] \pm O\Big(\frac{S_{i,\mathbf{x}}^{(1)} \sqrt{\log N}}{N} \Big) \pm O\Big(\frac{S_{i,\mathbf{x}}^{(1)}}{KN^{\alpha-1}} \Big) \Big\} \pm O\Big(\frac{1}{N^3}\Big).$$

5. *If $i = 2n-1$ and $j = 2m-1$ for some $m \le n \le N_{\text{in}}$, then we have*

$$a_{i,j} \simeq \frac{S_{i,j}^{(1)}}{KN} \sum_{k=1}^{K} \Big\{ \mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\text{q}} = \mathbf{v}_k\} \Big(\sum_{o=1}^{K} (\overline{\mathcal{S}}_o^{(2)})^2 - \overline{\mathcal{S}}_k^{(2)} \Big) S_{i,y}^{(1)} \Big] \pm O\Big(\frac{(1-S_{i,j}^{(1)})\sqrt{\log N}}{N} \Big) \pm O\Big(\frac{1-S_{i,j}^{(1)}}{N^{\alpha-1}} \Big) \Big\} \pm O\Big(\frac{1}{N^3}\Big).$$

Proof First, we notice that

$$a_{i,j} = \mathbb{E}\Big[\Big(\sum_{o=1}^{K} \mathcal{S}_o^{(2)} \mathbf{v}_o - \mathbf{x}_{\text{q}} \Big)^\top \mathbf{w}\mathbf{w}^\top \Big(\sum_{o=1}^{K} \mathcal{S}_o^{(2)} \mathbf{v}_o - \mathbb{1}\{i \equiv 0 \pmod 2\} \mathbf{x}_{\frac{i}{2}} \Big) S_i^{(2)} S_{i,j}^{(1)} (I(i)_j - \sum_{\ell \le i} S_{i,\ell}^{(1)} I(i)_\ell) \Big]$$

$$\overset{(i)}{=} \mathbb{E}\Big[\Big(\sum_{o=1}^{K} \mathcal{S}_o^{(2)} \mathbf{v}_o - \mathbf{x}_{\text{q}} \Big)^\top \Big(\sum_{o=1}^{K} \mathcal{S}_o^{(2)} \mathbf{v}_o - \mathbb{1}\{i \equiv 0 \pmod 2\} \mathbf{x}_{\frac{i}{2}} \Big) S_i^{(2)} S_{i,j}^{(1)} (I(i)_j - \sum_{\ell \le i} S_{i,\ell}^{(1)} I(i)_\ell) \Big]$$

$$\overset{(ii)}{\simeq}\mathbb{E}\Big[\Big(\sum_{o=1}^{K}\mathcal{S}_o^{(2)}\mathbf{v}_o-\mathbf{x}_{\mathrm{q}}\Big)^{\top}\Big(\sum_{o=1}^{K}\mathcal{S}_o^{(2)}\mathbf{v}_o-\mathbb{1}\{i\equiv 0 \pmod 2\}\mathbf{x}_{\frac{i}{2}}\Big)S_i^{(2)}S_{i,j}^{(1)}\Big(I(i)_j-\sum_{\ell\le i}S_{i,\ell}^{(1)}I(i)_\ell\Big)\mathbb{1}\{\mathcal{E}_S^*\}\Big]$$
$$+2\mathbb{P}(\mathcal{E}_S^{*c}),$$

where (i) follows from taking the conditional expectation over $\mathbf{w}$, and (ii) follows from the fact that $\Big|\Big(\sum_{o=1}^{K}\mathcal{S}_o^{(2)}\mathbf{v}_o-\mathbf{x}_{\mathrm{q}}\Big)^{\top}\Big(\sum_{o=1}^{K}\mathcal{S}_o^{(2)}\mathbf{v}_o-\mathbb{1}\{i\equiv 0\,(\mathrm{mod}\ 2)\}\mathbf{x}_{\frac{i}{2}}\Big)\,S_i^{(2)}S_{i,j}^{(1)}(I(i)_j-\sum_{\ell\le i}S_{i,\ell}^{(1)}I(i)_\ell)\Big|\le 2.$

By Theorem J.6, and taking $\delta = N^{-3}$, we can replace the random variables $\mathcal{S}_o^{(2)}$ and $S_i^{(2)}$ with their deterministic approximation, which enables a simpler form of $a_{i,j}$. Then we have

$$a_{i,j}\simeq\mathbb{E}\Big[\Big(\sum_{o=1}^{K}\Big(\overline{\mathcal{S}}_o^{(2)}\pm O\Big(\frac{\sqrt{\log(N)}}{N}\Big)\Big)\mathbf{v}_o-\mathbf{x}_{\mathrm{q}}\Big)^{\top}\Big(\sum_{o=1}^{K}\Big(\overline{\mathcal{S}}_o^{(2)}\pm O\Big(\frac{\sqrt{\log(N)}}{N}\Big)\Big)\mathbf{v}_o-\mathbb{1}\{i\equiv 0\,(\mathrm{mod}\ 2)\}\mathbf{x}_{\frac{i}{2}}\Big)$$
$$\overline{S}_i^{(2)}\big(1\pm O\big(\min\{1,i^{-1}\sqrt{\log(2N/\delta)}\}\big)\big)S_{i,j}^{(1)}\Big(I(i)_j-\sum_{\ell\le i}S_{i,\ell}^{(1)}I(i)_\ell\Big)\Big]\pm O\Big(\frac{1}{N^3}\Big)$$
$$\overset{(i)}{\simeq}\mathbb{E}\Big[\Big\{\Big(\sum_{o=1}^{K}\overline{\mathcal{S}}_o^{(2)}\mathbf{v}_o-\mathbf{x}_{\mathrm{q}}\Big)^{\top}\Big(\sum_{o=1}^{K}\overline{\mathcal{S}}_o^{(2)}\mathbf{v}_o-\mathbb{1}\{i\equiv 0\,(\mathrm{mod}\ 2)\}\mathbf{x}_{\frac{i}{2}}\Big)$$
$$\pm O\Big(\Big(1-\mathbb{1}\{i\equiv 0\,(\mathrm{mod}\ 2)\}\frac{\sqrt{\log(N)}}{N}\Big)\frac{\sqrt{\log(N)}}{N}\Big)\Big\}\Big(I(i)_j-\sum_{\ell\le i}S_{i,\ell}^{(1)}I(i)_\ell\Big)\Big]\frac{S_{i,j}^{(1)}}{N}\pm O\Big(\frac{1}{N^3}\Big),\tag{33}$$

where (i) follows from that we only care about the order, and $\overline{S}_i^{(2)}=\Theta(1/N)$. Notice that the choice of N^{-3} here ensures that it remains a negligible term in the later analysis. Starting from (33), we consider by cases. We only prove Case (1), and the other cases follow by the same argument as Case (1).

Case (1): $i=2n, j=2n-1$. Because $i=2n$ is even, by direct calculation, we have

$$a_{i,j}=\mathbb{E}\Big[\Big(\Big(\sum_{o=1}^{K}\overline{\mathcal{S}}_o^{(2)}\mathbf{v}_o-\mathbf{x}_{\mathrm{q}}\Big)^{\top}\Big(\sum_{o=1}^{K}\overline{\mathcal{S}}_o^{(2)}\mathbf{v}_o-\mathbf{x}_n\Big)\pm O\Big(\frac{\log(N)}{N^2}\Big)\Big)\Big(I(i)_j-\sum_{\ell\le i}S_{i,\ell}^{(1)}I(i)_\ell\Big)\Big]\cdot\frac{S_{i,j}^{(1)}}{N}$$
$$\pm O\Big(\frac{1}{N^3}\Big)$$
$$=\sum_{k=1}^{K}\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\Big(\sum_{o=1}^{K}(\overline{\mathcal{S}}_o^{(2)})^2-\overline{\mathcal{S}}_k^{(2)}-\sum_{o=1}^{K}\overline{\mathcal{S}}_o^{(2)}\mathbb{1}\{\mathbf{x}_n=\mathbf{v}_o\}+\mathbb{1}\{\mathbf{x}_n=\mathbf{v}_k\}\pm O\Big(\frac{\log(N)}{N^2}\Big)\Big)$$
$$\Big(I(i)_j-\sum_{\ell\le i}S_{i,\ell}^{(1)}I(i)_\ell\Big)\Big]\frac{S_{i,j}^{(1)}}{N}\pm O\Big(\frac{1}{N^3}\Big)$$
$$=\frac{S_{i,j}^{(1)}}{KN}\sum_{k=1}^{K}(I)+O\Big(\frac{1}{N^3}\Big).\tag{34}$$

Here, the dominant term (I) is defined and further simplified as follows

$$(I)=\sum_{k'=1}^{K}\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\Big(\sum_{o=1}^{K}(\overline{\mathcal{S}}_o^{(2)})^2-\overline{\mathcal{S}}_k^{(2)}-\overline{\mathcal{S}}_{k'}^{(2)}+\mathbb{1}\{k'=k\}\pm O\Big(\frac{\log(N)}{N^2}\Big)\Big)\Big(I(i)_j-\sum_{\ell\le i}S_{i,\ell}^{(1)}I(i)_\ell\Big)\Big]$$
$$\overset{(i)}{=}\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\sum_{k'=1}^{K}\Big(\sum_{o=1}^{K}(\overline{\mathcal{S}}_o^{(2)})^2-\overline{\mathcal{S}}_k^{(2)}-\overline{\mathcal{S}}_{k'}^{(2)}\Big)\Big(-\frac{1}{K}\sum_{\ell\le i-2}S_{i,\ell}^{(1)}\mathbb{1}\{\ell=2r-1\}\Big)\Big]$$
$$+\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\Big(-\sum_{\ell\le i-2}\frac{1}{K}S_{i,\ell}^{(1)}\mathbb{1}\{\ell=2r-1\}+\Big(\sum_{o=1}^{K}(\overline{\mathcal{S}}_o^{(2)})^2-2\overline{\mathcal{S}}_k^{(2)}+1\Big)(1-S_{i,i-1}^{(1)})\Big)\Big]$$

$$
\begin{aligned}
&\pm O\Big(\frac{(1-S^{(1)}_{i,i-1})\log(N)}{KN^2}\Big) \pm O\Big(\frac{1-S^{(1)}_{i,i-1}}{KN^{\alpha-1}}\Big) \\
&= \mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\Big\{\Big(\frac{K-1}{K}(1-\overline{\mathcal{S}}^{(2)}_k)+\frac{1}{K}\sum_{o\neq k}^{K}\overline{\mathcal{S}}^{(2)}_o\Big)(S^{(1)}_{i,\mathbf{x}}-S^{(1)}_{i,i-1}) \\
&\quad+\Big(\sum_{o=1}^{K}(\overline{\mathcal{S}}^{(2)}_o)^2-2\overline{\mathcal{S}}^{(2)}_k+1\Big)S^{(1)}_{i,y}\Big\}\Big] \pm O\Big(\frac{(1-S^{(1)}_{i,i-1})}{K}\Big(\frac{\log N}{N^2}+\frac{1}{N^{\alpha-1}}\Big)\Big),
\end{aligned} \tag{35}
$$

where (i) follows from Theorem J.5, which states that $|I(i)_j - \mathbb{1}\{j=2r-1, \mathbf{x}_r=\mathbf{v}_k\}| \le 5N^{1-\alpha}$ when $\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k$. It also uses the fact that $\{\mathbf{x}_r\}_{r=1}^{\lfloor\frac{i-1}{2}\rfloor}$ is independent with $\{\overline{\mathcal{S}}^{(2)}_o\}_{o=1}^{K}, \{S^{(1)}_{i,j}\}_{j=1}^{i-1}$, and taking partial expectation over $\{\mathbf{x}_r\}_{r=1}^{\lfloor\frac{i-1}{2}\rfloor}$.

Plugging (35) into (34) proves the result of Case (1). The analyses for the other cases proceed analogously to Case (1), and we therefore omit the details. This concludes the proof of Theorem J.3.

■

J.3. Stage I: Auxiliary Lemmas

The next lemma characterizes the influence of the low-frequency components of RoPE on the tokens, which will be useful for simplifying $I(i) = E^{\mathbf{x},y}R_{\boldsymbol{\vartheta},N-i}(E^{\mathbf{x},y}_{\mathrm{q}})^\top$.

Lemma J.4. *Suppose Assumption 4.1 holds. Let $E^{\mathbf{x},y}, E^{\mathbf{x},y}_{\mathrm{q}}$ denote the semantic dependent subspaces of the token embeddings E and E_{q}. Then, for any $1 \le j \le i \le N$, the partial token embedding $E^{\mathbf{x},y}_j$ at position j satisfies:*

1. *If j is odd, and $E^{\mathbf{x},y}_j$ has the same feature with $E^{\mathbf{x},y}_{\mathrm{q}}$, i.e. $E^{\mathbf{x},y}_j = E^{\mathbf{x},y}_{\mathrm{q}}$, then*

$$
1 - E^{\mathbf{x},y}_j R_{\boldsymbol{\vartheta},N-i}(E^{\mathbf{x},y}_{\mathrm{q}})^\top \le \frac{2}{N^{2(\alpha-1)}}.
$$

2. *If j is odd, and $E^{\mathbf{x},y}_j$ has a different feature with $E^{\mathbf{x},y}_{\mathrm{q}}$, i.e. $E^{\mathbf{x},y}_j \perp E^{\mathbf{x},y}_{\mathrm{q}}$, then*

$$
\big|E^{\mathbf{x},y}_j R_{\boldsymbol{\vartheta},N-i}(E^{\mathbf{x},y}_{\mathrm{q}})^\top\big| \le \frac{5}{N^{\alpha-1}}.
$$

3. *If j is even, then $E^{\mathbf{x},y}_j R_{\boldsymbol{\vartheta},N-i}(E^{\mathbf{x},y}_{\mathrm{q}})^\top = 0$.*

Proof We prove by cases.

Case (1). If j is odd, and $E^{\mathbf{x},y}_j$ has the same feature with $E^{\mathbf{x},y}_{\mathrm{q}}$, i.e. $E^{\mathbf{x},y}_j = E^{\mathbf{x},y}_{\mathrm{q}}$, then

$$
\begin{aligned}
&1 - E^{\mathbf{x},y}_j R_{\boldsymbol{\vartheta},N-i}(E^{\mathbf{x},y}_{\mathrm{q}})^\top \\
&= 1 - \sum_{\ell=d_{\mathbf{c}}/2+1}^{d/2}\big\{(E^{\mathbf{x},y}_{\mathrm{q}})^2_{2\ell-1}+(E^{\mathbf{x},y}_{\mathrm{q}})^2_{2\ell}\big\}\cos(\theta_\ell(N-i)) \\
&\le 1 - \sum_{\ell=d_{\mathbf{c}}/2+1}^{d/2}\big\{(E^{\mathbf{x},y}_{\mathrm{q}})^2_{2\ell-1}+(E^{\mathbf{x},y}_{\mathrm{q}})^2_{2\ell}\big\}\Big\{1-\frac{1}{2}(\theta_\ell(N-i))^2\Big\} \\
&\le \frac{2}{N^{2(\alpha-1)}} \\
&\le \frac{2}{N^{\alpha-1}},
\end{aligned}
$$

where the first inequality follows from that $1 - x^2 \le \cos(x)$, and the second inequality follows from the that $\|E^{\mathbf{x},y}_{\mathrm{q}}\|_2 = 1$.

Case (2). If j is odd, and $E_j^{\mathbf{x},y}$ has a different feature with $E_{\mathrm{q}}^{\mathbf{x},y}$, i.e. $E_j^{\mathbf{x},y} \perp E_{\mathrm{q}}^{\mathbf{x},y}$,

$$\begin{aligned}
&E_j^{\mathbf{x},y} R_{\boldsymbol{\vartheta},N-i} (E_{\mathrm{q}}^{\mathbf{x},y})^\top \\
&\overset{(i)}{\leq} \left|(E_{\mathrm{q}}^{\mathbf{x},y})^\top E_j^{\mathbf{x},y} \cos(\theta_{d/2}(N-i))\right| \\
&\quad + \left|\sum_{\ell=1}^{\frac{d}{2}} \left\{(E_{\mathrm{q}}^{\mathbf{x},y})_{2\ell-1}^2 + (E_j^{\mathbf{x},y})_{2\ell-1}^2 + (E_{\mathrm{q}}^{\mathbf{x},y})_{2\ell}^2 + (E_j^{\mathbf{x},y})_{2\ell}^2\right\} \left\{(\theta_\ell - \theta_{d/2})(N-i) \cdot \frac{2}{N^{\alpha-1}}\right\}\right| \\
&\quad + \left|\sum_{\ell=1}^{\frac{d}{2}} \left\{(E_{\mathrm{q}}^{\mathbf{x},y})_{2\ell-1}^2 + (E_j^{\mathbf{x},y})_{2\ell-1}^2 + (E_{\mathrm{q}}^{\mathbf{x},y})_{2\ell}^2 + (E_j^{\mathbf{x},y})_{2\ell}^2\right\}\right| |\theta_\ell(N-i)| \\
&\leq \frac{8}{N^{2(\alpha-1)}} + \frac{4}{N^{\alpha-1}} \leq \frac{5}{N^{\alpha-1}},
\end{aligned}$$

where (i) follows from adding and subtracting $\cos(\theta_{d/2}(N-i))$, the inequality $\sin(x) \leq x$ and the fact that for $0 \leq x_1 \leq x_2 \leq 1$, $|\cos(x_1) - \cos(x_2)| \leq |x_1 - x_2| \sin(x_2)$.

Case (3). If j is even, then the sub-vector of $E_j^{\mathbf{x},y}$ corresponding to $\mathbf{x}$ is zero. Consider the sub-vector of $E_{\mathrm{q}}^{\mathbf{x},y}$ corresponding to y is also zero, we immediately have $E_j^{\mathbf{x},y} R_{\boldsymbol{\vartheta},N-i} (E_{\mathrm{q}}^{\mathbf{x},y})^\top = 0$ for all i is even. Hence, we finish the proof of Theorem J.4. ■

As a direct result of Theorem J.4, we have the following corollary to control $I(i)$.

Corollary J.5 (Bounds of $I(i)$)**.** *Suppose Assumption 4.1 holds, for any $1 \leq j \leq i \leq N$,*

$$|I(i)_j - \mathbb{1}\{j \equiv 1 \pmod 2\}\mathbb{1}\{E_j = E_{\mathrm{q}}\}| \leq 5N^{1-\alpha}.$$

In addition, conditioned on the event $\{\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k\}$, we have

$$\left|I(i)_j - \mathbb{1}\{j \equiv 1 \pmod 2\}\mathbb{1}\{\mathbf{x}_{(j+1)/2} = \mathbf{v}_k\}\right| \leq 5N^{1-\alpha}.$$

We omit the proof here, as it follows directly from Theorem J.4. We next show concentration results of $S^{(2)}$ and $\mathcal{S}^{(2)}$, which is important in simplifying $a_{i,j}$. We define their approximate expectation versions $\overline{S}^{(2)}$ and $\overline{\mathcal{S}}^{(2)}$, then control the approximation errors.

Lemma J.6 (Concentration of $S^{(2)}$,$\mathcal{S}^{(2)}$)**.** *Suppose the prompt P is randomly sampled according to the data model in Section D.1. Define $\overline{S}_i^{(2)} = \exp(\mathbb{E}[\mathbf{u}_i])/(\sum_{j=1}^N \exp(\mathbb{E}[\mathbf{u}_j]))$ and $\overline{\mathcal{S}}_k^{(2)} = K^{-1}\sum_{i=1}^{N_{\mathrm{in}}} \overline{S}_{2i}^{(2)}$ as the approximate expectation versions of $S^{(2)}$ and $\mathcal{S}^{(2)}$. For any $i \in [N]$, with probability at least $1-\delta$, we have*

$$\begin{aligned}
S_i^{(2)} &= \overline{S}_i^{(2)}\big(1 \pm O\big(\min\big\{1, i^{-1}\sqrt{\log(2N/\delta)}\big\}\big)\big), i \in [N], \\
\mathcal{S}_m^{(2)} &= \overline{\mathcal{S}}_m^{(2)} \pm O\big(N^{-1}\sqrt{\log(2K/\delta)}\big), m \in [K].
\end{aligned}$$

Furthermore, denote the event above by $\mathcal{E}_S^*$, we have $\mathbb{P}(\mathcal{E}_S^*) \geq 1-\delta$.

Proof We first prove the concentration result of $S^{(2)}$. Recall that from (24), we have

$$S_i^{(2)} = \frac{\exp(\mathbf{u}_i)}{\sum_{j=1}^N \exp(\mathbf{u}_j)}, \mathbf{u}_i = \sum_{j=1}^i S_{i,j}^{(1)} E_j^{\mathbf{x},y} R_{\boldsymbol{\vartheta},N-i} (E_{\mathrm{q}}^{\mathbf{x},y})^\top.$$

From the definition, $S^{(1)}$ is independent with prompt P. Consequently, when j is odd, the terms $S_{i,j}^{(1)} E_j^{\mathbf{x},y} R_{\boldsymbol{\vartheta},N-i} (E_{\mathrm{q}}^{\mathbf{x},y})^\top$ are conditionally independent with each other for different $j \leq i$ given $E_{\mathrm{q}}^{\mathbf{x},y}$. Furthermore, for each $j \leq i$, the term $S_{i,j}^{(1)} E_j^{\mathbf{x},y} R_{\boldsymbol{\vartheta},N-i} (E_{\mathrm{q}}^{\mathbf{x},y})^\top$ is bounded between $-S_{i,j}^{(1)}$ and $S_{i,j}^{(1)}$. Then for all $i \in [N]$, following from the standard Hoeffding inequality techniques (Hoeffding, 1963; Mohri et al., 2018), with probability at least $1-\delta/2$, we have

$$|\mathbf{u}_i - \mathbb{E}[\mathbf{u}_i]| \leq \sqrt{\frac{\sum_{j=1}^i 4(S_{i,j}^{(1)})^2}{2i^2} \cdot \log(2N/\delta)} \leq \frac{\sqrt{2\log(2N/\delta)}}{i}. \tag{36}$$

By Theorem J.5, which shows $\left|\mathbb{E}[\mathbf{u}_i] - \frac{1}{K}\sum_{j\le i, j\equiv 1 \,(\mathrm{mod}\ 2)} S_{i,j}^{(1)}\right| \le 5N^{1-\alpha}$, and (36), with probability at least $1-\delta/2$, we have

$$\begin{aligned} S_i^{(2)} &= \frac{\exp\left(\mathbb{E}[\mathbf{u}_i]\right)(1 + (\mathbf{u}_i - \mathbb{E}[\mathbf{u}_i]) + o(\mathbf{u}_i - \mathbb{E}[\mathbf{u}_i]))}{\sum_{j=1}^N \exp\left(\mathbb{E}[\mathbf{u}_j]\right)(1 + (\mathbf{u}_j - \mathbb{E}[\mathbf{u}_j]) + o(\mathbf{u}_j - \mathbb{E}[\mathbf{u}_j]))} \\ &= \overline{S}_i^{(2)}\left(1 \pm O\left(i^{-1}\sqrt{\log(2N/\delta)}\right)\right), \end{aligned} \tag{37}$$

In addition, by noticing that $\overline{S}_i^{(2)}$ and $S_i^{(2)}$ are in the same order, i.e., $S_i^{(2)} = \Theta(\overline{S}_i^{(2)})$, we have

$$S_i^{(2)} = \overline{S}_i^{(2)}\big(1 \pm O\big(\min\big\{1, i^{-1}\sqrt{\log(2N/\delta)}\big\}\big)\big).$$

We next prove the concentration of $\mathcal{S}^{(2)}$. With probability at least $1-\delta$, for any $m \in [K]$, we have

$$\begin{aligned} \mathcal{S}_m^{(2)} &\overset{(i)}{=} \sum_{i=1}^{N_{\mathrm{in}}} \mathbb{1}\{\mathbf{x}_i = \mathbf{v}_m\}\overline{S}_{2i}^{(2)}\left(1 \pm O\left(\frac{\sqrt{\log(2N/\delta)}}{2i}\right)\right) \\ &\overset{(ii)}{=} \frac{1}{K}\sum_{i=1}^{N_{\mathrm{in}}} \overline{S}_{2i}^{(2)}\left(1 \pm O\left(\frac{\sqrt{\log(2N/\delta)}}{2i}\right)\right) \pm O\left(\frac{\sqrt{\log\left(2K/\delta\right)}}{N}\right) \\ &= \overline{\mathcal{S}}_m^{(2)} \pm O\left(\frac{\sqrt{\log(2K/\delta)}}{N}\right), \end{aligned}$$

where (i) follows from the definitions of $\mathcal{S}_m^{(2)}$, $\overline{\mathcal{S}}_m^{(2)}$ and (37), and (ii) follows from the standard Hoeffding inequality and that $p_k = 1/K$. Hence, we finish the proof of Theorem J.6. ■

Because $-1 \le \mathbf{u}_j \le 1$ for all $j \in [N]$, by the definition of $\overline{S}_i^{(2)}$ and $\overline{\mathcal{S}}_k^{(2)}$, we obtain the following corollary via simple calculations based on Theorem J.6, which provides the order of $\overline{\mathcal{S}}^{(2)}$ and $\overline{S}^{(2)}$.

Corollary J.7. *Through Stage I, for all $i \in [N]$ and $m \in [K]$, we have $\overline{S}_i^{(2)} = \Theta\left(1/N\right)$ and $\overline{\mathcal{S}}_k^{(2)} = \Theta\left(1/K\right)$.*

In the next section, we will analyze the training dynamics of Stage I.

J.4. Stage I: Phase I

In this section, we shall study the initial phase of learning the relationship between any position i and its previous tokens. We define the **Phase I** as all iterations $0 \le t \le T_1^{(1)}$, where

$$T_1^{(1)} \triangleq \max\left\{t : \min_{i\in[N]}\left(A_{i,i-1}(t) - \max_{j\le i, j\ne i-1} A_{i,j}(t)\right) \le \log(N)\right\}.$$

We then state the following induction hypothesis, which holds throughout Phase I. This hypothesis is proved by induction together with the technical lemmas in Section J.4.1.

Induction Hypothesis J.1. *For each $0 \le t \le T_1^{(1)}$ and any $i \in [N]$, the following holds:*

1. *$A_{i,i-1}(t) - \max_{j\le i, j\ne i-1} A_{i,j}(t)$ is monotonically increasing and for all $i \in [N]$, $0 \le A_{i,i-1}(t) - \max_{j\le i, j\ne i-1} A_{i,j}(t) \le \log N\left(1 + O\left(K^{-1}\log N\right)\right)$.*
2. *$\max_{j\le i, j\ne i-1} A_{i,j}(t) - \min_{l\le i, l\ne i-1} A_{i,l}(t) \le O\left(K^{-1}(\log N)^2\right)$.*

Proof We first prove Claim (1), as $A_{i,j}(0) = 0$ for all $j \le i \le N$, we only need to prove that $\Delta A_{i,i-1}(t-1) - \Delta A_{i,j}(t-1) \ge$ for all $j \le i, j \ne i-1$. By Theorem J.10, we have

$$\Delta A_{i,i-1}(t) - \max_{j\le i, j\ne i-1} \Delta A_{i,j}(t) \gtrsim \frac{\eta_1 C_1}{KN},$$

which is larger than 0. As a result,

$$\begin{aligned}
A_{i,i-1}(t) - \max_{j\le i, j\ne i-1} A_{i,j}(t) &\le \sum_{\tau=0}^{t-1} \Delta A_{i,i-1}(\tau) - \min_{j\le i, j\ne i-1} \Delta A_{i,j}(\tau) \\
&\le \sum_{\tau=0}^{t-1} \left(1 + O\left(K^{-1}\log N\right)\right) \min_{l\in[N]} \left(\Delta A_{l,l-1}(\tau) - \max_{r\le l, r\ne l-1} \Delta A_{l,r}(\tau)\right) \\
&= \left(1 + O\left(K^{-1}\log N\right)\right) \min_{l\in[N]} \left(A_{l,l-1}(t) - \max_{r\le l, r\ne l-1} A_{l,r}(t)\right) \\
&\le \left(1 + O\left(K^{-1}\log N\right)\right) \log N,
\end{aligned}$$

where the second inequality follows from that for each i, $\Delta A_{i,i-1}(t) - \min_{j\le i, j\ne i-1} \Delta A_{i,j}(t) = \min_{i\in[N]}\{\Delta A_{i,i-1}(t) - \max_{j\le i, j\ne i-1} \Delta A_{i,j}(t)\}(1+O(\log N/K))$ as in Theorem J.11, and the last inequality follows from the definition of Phase I. Hence, we prove Claim (1).

Then we proceed to prove Claim (2) as follows.

$$\begin{aligned}
\max_{j\le i, j\ne i-1} A_{i,j}(t) - \min_{l\le i, l\ne i-1} A_{i,l}(t) &\le \sum_{\tau=0}^{t-1} \left(\max_{j\le i, j\ne i-1} \Delta A_{i,j}(\tau) - \min_{l\le i, l\ne i-1} \Delta A_{i,l}(\tau)\right) \\
&\overset{(i)}{\le} \sum_{\tau=0}^{t-1} O\Big(\frac{\log N}{K}\Big)\Big(\Delta A_{i,i-1}(\tau) - \max_{j\le i, j\ne i-1} \Delta A_{i,j}(\tau)\Big) \\
&\le O\Big(\frac{\log N}{K}\Big)\Big(A_{i,i-1}(t) - \max_{j\le i, j\ne i-1} A_{i,j}(t)\Big) \\
&\overset{(ii)}{=} O\Big(\frac{(\log N)^2}{K}\Big),
\end{aligned}$$

where (i) follows from Theorem J.11, and (ii) follows from Claim (1). Hence, we finish the proof of Induction Hypothesis J.1. ■

J.4.1. Technical Lemmas

We first introduce several useful technical lemmas, which characterize the important values including $S^{(1)}, \mathcal{S}^{(2)}, a_{i,j}, \Delta A$ across phase I. By induction, for all these lemmas, we only need to show that (1) for $t = 0$, the lemmas hold, and (2) if we assume that for $S^{(1)}(t-1), \mathcal{S}^{(2)}(t-1), A(t-1), a_{i,j}(t-1), \Delta A(t-1)$, the lemmas hold, then they still hold for $S^{(1)}(t), \mathcal{S}^{(2)}(t), A(t), a_{i,j}(t), \Delta A(t)$.

Lemma J.8. *For iteration $0 \le t \le T_1^{(1)}$, if Induction Hypothesis J.1 holds at iteration t, and Theorems J.9 and J.10 hold at iteration $t-1$, then*

1. *For any $i \in [N]$, we have $S_{i,i-1}^{(1)}(t) = \Omega\left(1/i\right)$ and $S_{i,j}^{(1)}(t) = O\left(1/i\right)$ for $j \ne i-1$.*
2. *For any $i \in [N]$,*
   - *(a) $S_{i,\mathbf{x}}^{(1)}(t) - S_{i,i-1}^{(1)}(t) \ge \Omega(i/N)$.*
   - *(b) $S_{i,y}^{(1)}(t) \ge \Omega(i/N)$.*
   - *(c) $1 - S_{i,i-1}^{(1)}(t) \ge \Omega(i/N)$.*

Proof We first prove Claim (1). For any $i \in [N]$

$$S_{i,i-1}^{(1)}(t) \ge \frac{1}{1 + (i-1)\exp\left(\max_{j\le i, j\ne i-1} A_{i,j}(t) - A_{i,i-1}(t)\right)}$$

$$\begin{aligned}
&\overset{(i)}{\geq} \frac{1}{1+(i-1)\exp\left(\max_{j\leq i, j\neq i-1} A_{i,j}(0) - A_{i,i-1}(0)\right)} \\
&= \frac{1}{i},
\end{aligned}$$

where (i) follows from that $1/(1+(i-1)e^x)$ is monotonically decreasing with x and Induction Hypothesis J.1. Furthermore, we have that

$$\begin{aligned}
S_{i,j}^{(1)}(t) &\overset{(i)}{\leq} \frac{1}{1+\sum_{l\leq i, l\neq j}\exp\left(-\left(\max_{j\leq i, j\neq i-1} A_{i,j}(t) - \min_{l\leq i, l\neq i-1} A_{i,l}(t)\right)\right)} \\
&\overset{(ii)}{\leq} \frac{1}{1+(i-1)\exp\left(-O\left(K^{-1}(\log N)^2\right)\right)} \\
&= \frac{1}{1+(i-1)\Omega(1)} \\
&= O\Big(\frac{1}{i}\Big),
\end{aligned}$$

where (i) follows from that $1/(1+e^x)$ decrease with of x, and Induction Hypothesis J.1 that $A_{i,i-1}(t) - \max_{j\leq i, j\neq i-1} A_{i,j}(t) \geq 0$, and (ii) follows from Induction Hypothesis J.1. Hence, we have proved Claim (1).

We then proceed to prove Claim (2) as follows.

$$\begin{aligned}
1 - S_{i,i-1}^{(1)}(t) &\overset{(i)}{\geq} \frac{\sum_{j\leq i, j\neq i-1}\exp\left(\min_{j\leq i, j\neq i-1} A_{i,j}(t) - A_{i,i-1}(t)\right)}{1+\sum_{j\leq i, j\neq i-1}\exp\left(\min_{j\leq i, j\neq i-1} A_{i,j}(t) - A_{i,i-1}(t)\right)} \\
&\overset{(ii)}{\geq} \frac{\sum_{j\leq i, j\neq i-1}\exp\left(-\log(N)\left(1+O\left(K^{-1}\log N\right)\right)\right)\exp\left(-O\left(K^{-1}(\log N)^2\right)\right)}{1+\sum_{j\leq i, j\neq i-1}\exp\left(-\log(N)\left(1+O\left(K^{-1}\log N\right)\right)\right)\exp\left(-O\left(K^{-1}(\log N)^2\right)\right)} \\
&\geq \Omega\Big(\frac{i-1}{N}\Big),
\end{aligned}$$

where (i) follows from that $e^x/(1+e^x)$ increase with of x, and (ii) follows from Theorem J.11 and Induction Hypothesis J.1. Similarly, we can prove

$$S_{i,y}^{(1)}(t) \geq \Omega\Big(\frac{i-1}{N}\Big), S_{i,\mathbf{x}}^{(1)}(t) - S_{i,i-1}^{(1)}(t) \geq \Omega\Big(\frac{i-1}{N}\Big).$$

Hence, we finish the proof of Theorem J.8. ■

In the following lemma, we control the order of $a_{i,j}(t)$.

Lemma J.9 (Order of $a_{i,j}(t)$)**.** *During stage I, suppose Induction Hypothesis J.1 and Theorem J.8 hold at iteration $0 \leq t \leq T_1^{(1)}$. For different cases of i, j, we have the following results.*

1. *For $i = 2n, j = 2n-1$, $a_{i,i-1}(t) \simeq (KN^2)^{-1}$.*
2. *For $i = 2n, j = 2m-1, m \leq n$, $-(iKN)^{-1}\left(K^{-1}+i^{-1}\right) \lesssim a_{i,j}(t) \lesssim (iKN)^{-1}\left(K^{-1}-i^{-1}\right)$.*
3. *For $i = 2n, j = 2m, m \leq n$, $-(iKN)^{-1}\left(K^{-1}+i^{-1}\right) \lesssim a_{i,j}(t) \lesssim (iKN)^{-1}\left(K^{-1}-i^{-1}\right)$.*
4. *For $i = 2n-1, j = 2m, m \leq n-1$, $|a_{i,j}(t)| \lesssim (iK^2N)^{-1}$.*
5. *For $i = 2n-1, j = 2m-1, m \leq n$, $|a_{i,j}(t)| \lesssim (iK^2N)^{-1}$.*

Proof We only need to apply Theorem J.3 and evaluate all $a_{i,j}(t)$s' order for the timestep t given $S^{(1)}(t), S^{(2)}(t), \mathcal{S}^{(2)}(t), a_{i,j}(t-1)$ by cases.

Case (1): $i = 2n, j = 2n-1$. We have

$$\begin{aligned} a_{i,i-1}(t) &\overset{(i)}{\gtrsim} \frac{1}{iKN}\left\{\sum_{k=1}^{K}\mathbb{E}\left[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\left(\left(1-\frac{1}{K}\right)\left(1-\frac{1}{2K}\right)\frac{i-1}{N}+\left(\frac{1}{4K}-\frac{1}{K}+1\right)\frac{i-1}{N}\right)\right]\right. \\ &\quad \left.\pm O\left(\frac{\log N}{N^2}\right)\pm O\left(\frac{1}{N^{\alpha-1}}\right)\right\}\pm O\left(\frac{1}{N^3}\right) \\ &\simeq \frac{1}{KN^2}>0, \end{aligned}$$

where (i) follows from expression of $a_{i,j}(t)$ in Theorem J.3, that $S^{(1)}_{i,i-1}(t) \geq \Omega(1/i)$ as in Theorem J.8, and $\overline{S}^{(2)}_i = \Theta(1/N)$ and $\overline{\mathcal{S}}^{(2)}_k = \Theta(1/K)$ as in Theorem J.7. Hence, we prove Claim (1).

Similar to Case (1), following Theorems J.3, J.7 and J.8, we can compute for the other cases and finish the proof of Theorem J.9. ■

As a result of Theorem J.9, for any $l \in [N]$, we can characterize the relative increments gap of A across different tokens as follows.

Lemma J.10. *Suppose Induction Hypothesis J.1, Theorems J.8 and J.9 hold at iteration $0 \leq t \leq T^{(1)}_1$, for any $l \in [N]$, we have the following results.*

1. ***Upper Bound:*** $\Delta A_{l,l-1}(t) - \max_{r\leq l, r\neq l-1}\Delta A_{l,r}(t) \gtrsim \eta_1 C_1/(KN)$.
2. ***Lower Bound:*** $\max_{j\leq l, j\neq l-1}\Delta A_{l,j}(t) - \min_{r\leq l, r\neq l-1}\Delta A_{l,r}(t) \lesssim \eta_1 C_1 \log(N)/(NK^2)$.

Proof From Theorem J.2, the increment of $A_{l,r}(t)$ can be decomposed as follows

$$\Delta A_{l,r}(t) = \eta_1\Bigg(\underbrace{C_1\sum_{i=l-r+1}^{N} a_{i,i+r-l}(t)}_{J^1_{l,r}(t)} + \underbrace{C_2\sum_{1\leq j\leq i\leq N} a_{i,j}(t)}_{J^2_{l,r}(t)} + \underbrace{\sum_{1\leq j\leq i\leq N}(\pm a_{i,j}(t)\Theta(\epsilon_{\mathrm{FN}}))}_{J^3_{l,r}(t)}\Bigg),$$

which has three parts: $J^1_{l,r}(t)$, $J^2_{l,r}(t)$ and the error $J^3_{l,r}(t)$. The second part $J^2_{l,r}(t)$ is independent with r, so we only need to discuss the first part $J^1_{l,r}(t)$ and the third part $J^3_{l,r}(t)$.

First Part. We discuss $J^1_{l,r}(t)$ by cases,

For Case (1): $r = l-1$ and $J^1_{l,l-1}(t)$, we have

$$\begin{aligned} J^1_{l,l-1}(t) &= C_1\sum_{n=1}^{N_{\mathrm{in}}} a_{2n,2n-1}(t) + \sum_{m=1}^{N_{\mathrm{in}}} a_{2m+1,2m}(t) \\ &\overset{(i)}{\gtrsim} C_1\left(\sum_{n=1}^{N_{\mathrm{in}}}\frac{1}{KN^2} - \sum_{m=1}^{N_{\mathrm{in}}}\frac{1}{(2m+1)K^2N}\right) \\ &\gtrsim \frac{C_1}{KN}>0, \end{aligned} \tag{38}$$

where (i) follows from Theorem J.9. A direct result from the equation above is that for any l, $J^1_{l,l-1}(t)$ is the same, and is lower bounded by the same order $\Omega(C_1/(KN))$.

Similar to Case (1), for Case (2) $r = l$ and $J^1_{l,l}(t)$, we have

$$J^1_{l,l}(t) = C\Big(\sum_{n=1}^{N_{\mathrm{in}}} a_{2n,2n}(t) + \sum_{m=1}^{N_{\mathrm{in}}+1} a_{2m-1,2m-1}(t)\Big) \lesssim \frac{C_1\log N}{K^2N}. \tag{39}$$

For Case (3), $l - r \geq 2$ and $l - r \equiv 0 \pmod 2$, we have

$$J^1_{l,r}(t) = C_1\Big(\sum_{n=\lfloor (l-r+1)/2 \rfloor}^{N_{\text{in}}} a_{2n,2n+r-l}(t) + \sum_{m=\lfloor (l-r+1)/2 \rfloor}^{N_{\text{in}}} a_{2m+1,2m+1+r-l}(t)\Big) \lesssim \frac{C_1 \log N}{2K^2 N}. \tag{40}$$

For Case (4), $l - r \geq 2$ and $l - r \equiv 0 \pmod 2$, we have

$$J^1_{l,r}(t) = C_1\Big(\sum_{n=\lfloor (l-r+1)/2 \rfloor}^{N_{\text{in}}} a_{2n,2n+r-l}(t) + \sum_{m=\lfloor (l-r+1)/2 \rfloor}^{N_{\text{in}}} a_{2m+1,2m+1+r-l}(t)\Big) \lesssim \frac{C_1 \log N}{2K^2 N}. \tag{41}$$

Combine (38) to (41), we have

$$J^1_{l,l-1}(t) - \max_{r \leq l, r \neq l-1} J^1_{l,r}(t) \gtrsim \frac{C_1}{KN}. \tag{42}$$

Third Part. Then for $J^3_{l,r}(t)$, and any $l \in [N]$, we have

$$\begin{aligned}
&J^3_{l,l-1}(t) - \max_{r \leq l, r \neq l-1} J^3_{l,r}(t) \\
&\quad \gtrsim -O\left(\epsilon_{\text{FN}}\right) \sum_{n=0, n\neq 1}^{N} \sum_{i=n+1}^{N} |a_{i,i-n}(t)| \\
&\quad \overset{(i)}{\gtrsim} -O\left(\epsilon_{\text{FN}}\right) \sum_{n=0, n\neq 1}^{N} O\left(\frac{\log N}{K^2 N}\right) \\
&\quad \gtrsim -O\left(\frac{\epsilon_{\text{FN}} \log N}{K^2}\right),
\end{aligned} \tag{43}$$

where (i) follows from Theorem J.9.

Finally, combine (42) and (43), we have

$$\Delta A_{l,l-1}(t) - \max_{r \leq l, r \neq l-1} \Delta A_{l,r}(t) \gtrsim \frac{\eta_1 C_1}{KN} - \frac{\eta_1 \epsilon_{\text{FN}} \log N}{K^2} \gtrsim \frac{\eta_1 C_1}{KN},$$

where the last equation follows from Assumption 4.1 that $\epsilon_{\text{FN}} \leq \Omega(C_1/N)$. Hence, we prove Claim (1).

Then we prove Claim (2). Similar to Claim (1), combine (42) and (43), we have

$$\max_{j \leq l, j \neq l-1} \Delta A_{l,j}(t) - \min_{r \leq l, r \neq l-1} \Delta A_{l,r}(t) \lesssim \frac{\eta_1 \log N}{K^2}\left(\frac{C_1}{N} + \epsilon_{\text{FN}}\right) \lesssim \frac{\eta_1 C_1 \log N}{NK^2}.$$

Hence, we finish the proof of Theorem J.10. ■

A direct corollary of Theorem J.10 is as follows.

Corollary J.11. *Under the same conditions of Theorem J.10, for any $l \in [N]$, we have*

$$\begin{aligned}
&\frac{\max_{j \leq l, j \neq l-1} \Delta A_{l,j}(t) - \min_{r \leq l, r \neq l-1} \Delta A_{l,r}(t)}{\min_{l \in [N]}\{\Delta A_{l,l-1}(t) - \max_{r \leq l, r \neq l-1} \Delta A_{l,r}(t)\}} = O\left(\frac{\log N}{K}\right), \\
&\frac{\Delta A_{l,l-1}(t) - \min_{r \leq l, r \neq l-1} \Delta A_{l,r}(t)}{\min_{l \in [N]}\{\Delta A_{l,l-1}(t) - \max_{r \leq l, r \neq l-1} \Delta A_{l,r}(t)\}} = 1 + O\left(\frac{\log N}{K}\right), \\
&\frac{\Delta A_{l,l-1}(t) - \max_{r \leq l, r \neq l-1} \Delta A_{l,r}(t)}{\min_{l \in [N]}\{\Delta A_{l,l-1}(t) - \max_{r \leq l, r \neq l-1} \Delta A_{l,r}(t)\}} = 1 + O\left(\frac{\log N}{K}\right).
\end{aligned}$$

J.4.2. End of Phase I

Lemma J.12. *With $T_1^{(1)}$ at most $O\left(\eta_1^{-1}C_1^{-1}KN\log N\right)$, for any $i \in [N]$, at iteration $t = T_1^{(1)} + 1$, we have*

1. $A_{i,i-1}(T_1^{(1)}+1) - \max_{j\le i, j\ne i-1} A_{i,j}(T_1^{(1)}+1) \ge \log(N)$,
2. $S_{i,i-1}^{(1)}(T_1^{(1)}+1) = \Omega(1)$.

Proof Claim (1) holds because of the definition of Phase I. Then, we only need to show the upper bound of $T_1^{(1)}$. As Induction Hypothesis J.1 holds, there exists an i for $T_1^{(1)}$ such that

$$\begin{aligned}\log(N) &\ge A_{i,i-1}(T_1^{(1)}) - \max_{j\le i, j\ne i-1} A_{i,j}(T_1^{(1)}) \\ &\ge \sum_{\tau=0}^{T_1^{(1)}-1} \Delta A_{i,i-1}(\tau) - \max_{j\le i, j\ne i-1} \Delta A_{i,j}(\tau) \\ &\overset{(i)}{\ge} \frac{T_1^{(1)}\eta_1 C_1}{KN},\end{aligned} \tag{44}$$

where (i) follows from Theorem J.10. As a result, we must have $T_1^{(1)} = O\left(\eta_1^{-1}C_1^{-1}KN\log N\right)$, otherwise the inequality would fail to hold. Hence, we prove Claim (1).

We then prove Claim (2), for all $i \in [N]$,

$$\begin{aligned}S_{i,i-1}^{(1)}(T_1^{(1)}+1) &= \frac{\exp\left(A_{i,i-1}(T_1^{(1)}+1)\right)}{\exp\left(A_{i,i-1}(T_1^{(1)}+1)\right) + \sum_{j\le i, j\ne i-1}\exp\left(A_{i,j}(T_1^{(1)}+1)\right)} \\ &\ge \frac{1}{1+(i-1)\exp\left(\max_{j\le i, j\ne i-1} A_{i,j}(T_1^{(1)}+1) - A_{i,i-1}(T_1^{(1)}+1)\right)} \\ &\overset{(i)}{\ge} \frac{1}{1+(i-1)e^{-\log(N)}} = \Omega(1),\end{aligned}$$

where (i) follows from Claim (1) of this lemma. Hence, we finish the proof of this lemma. ■

J.5. Stage I: Phase II

In this section, we study the rapid growth phase of learning the relationship between any position i and its previous tokens. We define the **Phase II** as all iterations $T_1^{(1)} + 1 \le t \le T_2^{(1)}$, where

$$T_2^{(1)} \triangleq \max\left\{t \ge T_1^{(1)} : \min_{i\in[N]}\left(A_{i,i-1}(t) - \max_{j\le i, j\ne i-1} A_{i,j}(t)\right) \le \log(KN)\right\}.$$

Then, we state the following induction hypothesis, which holds throughout Phase II. This hypothesis is proved by induction with the technical lemmas in Section J.5.1.

Induction Hypothesis J.2. *For each iteration $T_1^{(1)} + 1 \le t \le T_2^{(1)}$ and any $i \in [N]$, the followings hold:*

1. $A_{i,i-1}(t) - \max_{j\le i, j\ne i-1} A_{i,j}(t)$ *is monotonically increasing and* $A_{i,i-1}(t) - \max_{j\le i, j\ne i-1} A_{i,j}(t) \in \left[\log(N), \left(1 + O\left(K^{-1}\log N\right)\right)\log(KN)\right]$.
2. $\max_{j\le i, j\ne i-1} A_{i,j}(t) - \min_{l\le i, l\ne i-1} A_{i,l}(t) \le O\left(K^{-1}(\log N)^2\right)$.

Proof We first prove Claim (1), the monotonicity can be proved similarly to Induction Hypothesis J.1. As in Theorem J.15,

$$\Delta A_{i,i-1}(t) - \max_{j\le i, j\ne i-1} \Delta A_{i,j}(t) \gtrsim \frac{\eta_1 C_1}{K^2} \ge 0.$$

By direct calculation, we have

$$\begin{aligned}
&A_{i,i-1}(t) - \max_{j\le i, j\ne i-1} A_{i,j}(t) - \Big(A_{i,i-1}(T_1^{(1)}) - \max_{j\le i, j\ne i-1} A_{i,j}(T_1^{(1)})\Big) \\
&\overset{(i)}{\le} \sum_{\tau=T_1^{(1)}}^{t-1} \left(1+O\left(N^{-1}K\right)\right) \min_{l\in[N]} \Big(\Delta A_{l,l-1}(\tau) - \max_{r\le l, r\ne l-1} \Delta A_{l,r}(\tau)\Big) \\
&\le \left(1+O\left(N^{-1}K\right)\right) \min_{l\in[N]} \Big(A_{l,l-1}(t) - A_{l,l-1}(T_1^{(1)}) - \max_{r\le l, r\ne l-1} \Big(A_{l,r}(t) - A_{l,r}(T_1^{(1)})\Big)\Big) \\
&\overset{(ii)}{\le} (1+O(N^{-1}K))\Big\{ \min_{l\in[N]} \Big(A_{l,l-1}(t) - \max_{r\le l, r\ne l-1} A_{l,r}(t)\Big) - \min_{l\in[N]} \Big(A_{l,l-1}(T_1^{(1)}) - \max_{r\le l, r\ne l-1} A_{l,r}(T_1^{(1)})\Big)\Big\},
\end{aligned} \tag{45}$$

where (i) follows from Theorem J.16, and (ii) follows from $-\max(A-B) \le -\max A + \max B$, and $\min(A-B) \le \min A - \min B$. As a result of (45), we have

$$\begin{aligned}
A_{i,i-1}(t) - \max_{j\le i, j\ne i-1} A_{i,j}(t) &\overset{(i)}{\le} \left(1+O\left(N^{-1}K + K^{-1}\log N\right)\right) \min_{l\in[N]} \Big(A_{l,l-1}(t) - \max_{r\le l, r\ne l-1} A_{l,r}(t)\Big) \\
&\le \left(1+O\left(K^{-1}\log N\right)\right) \log\left(KN\right),
\end{aligned} \tag{46}$$

where (i) follows from Induction Hypothesis J.1, and that $A_{l,l-1}(t) - \max_{r\le l, r\ne l-1} A_{l,r}(t)$ is monotonically increasing. Hence, we prove Claim (1).

We next prove Claim (2),

$$\begin{aligned}
&\max_{j\le i, j\ne i-1} A_{i,j}(t) - \min_{l\le i, l\ne i-1} A_{i,l}(t) - \Big(\max_{j\le i, j\ne i-1} A_{i,j}(T_1^{(1)}) - \min_{l\le i, l\ne i-1} A_{i,l}(T_1^{(1)})\Big) \\
&\le \sum_{\tau=T_1^{(1)}}^{t-1} O\left(N^{-1}K\right) \Big(\Delta A_{i,i-1}(\tau) - \max_{j\le i, j\ne i-1} \Delta A_{i,j}(\tau)\Big) \\
&= O\left(N^{-1}K\right)\Big(A_{i,i-1}(t) - \max_{j\le i, j\ne i-1} A_{i,j}(t) - \Big(A_{i,i-1}(T_1^{(1)}) - \max_{j\le i, j\ne i-1} A_{i,j}(T_1^{(1)})\Big)\Big),
\end{aligned}$$

where the first inequality follows from Theorem J.16. By direct calculation, we have

$$\begin{aligned}
\max_{j\le i, j\ne i-1} A_{i,j}(t) - \min_{l\le i, l\ne i-1} A_{i,l}(t) &\overset{(i)}{\le} O\left(K^{-1}(\log N)^2\right) + O\left(N^{-1}K\right)\left(1+O\left(K^{-1}\log N\right)\right)\log\left(KN\right) \\
&\le O\left(K^{-1}(\log N)^2\right),
\end{aligned}$$

where (i) follows from Induction Hypothesis J.1. Hence, we finish the proof of Induction Hypothesis J.2. ■

J.5.1. Technical Lemmas

Similar to phase I, we next introduce several useful technical lemmas. The proofs follow arguments analogous to those in Phase I and are omitted for brevity.

Lemma J.13. *For iteration $T_1^{(1)} + 1 \le t \le T_2^{(1)}$, if Induction Hypothesis J.2 holds at iteration t, and Theorems J.14 and J.15 hold at iteration $t-1$, then the followings hold:*

1. *For any $i \in [N]$,*
   - *(a) $S_{i,i-1}^{(1)}(t) = \Omega\left(1\right)$,*
   - *(b) $\Omega\left(1/(KN)\right) \le S_{i,j}^{(1)}(t) \le O\left(1/N\right), j \ne i-1$.*
2. *For any $i \in [N]$,*
   - *(a) $S_{i,\mathbf{x}}^{(1)}(t) - S_{i,i-1}^{(1)}(t) \ge \Omega(i/(KN))$,*

(b) $S_{i,y}^{(1)}(t) \geq \Omega(i/(2KN))$,

(c) $1 - S_{i,i-1}^{(1)}(t) \geq \Omega(i/(2KN))$.

Similar to Theorem J.9, and following from the expression of $a_{i,j}(t)$ in Theorem J.3, control of attention scores in Theorems J.7 and J.13, we can control the order of $a_{i,j}(t)$ as follows,

Lemma J.14 (Order of $a_{i,j}(t)$)**.** *Suppose Induction Hypothesis J.2 and Theorem J.13 hold at iteration $T_1^{(1)} + 1 \leq t \leq T_2^{(1)}$. For different cases of i, j, we have the following results.*

1. $i = 2n, j = 2n - 1$, $a_{i,i-1}(t) \gtrsim iK^{-2}N^{-2}$.
2. $i = 2n, j = 2m - 1, m \leq n$, $-K^{-1}N^{-2} \lesssim a_{i,j}(t) \lesssim -K^{-2}N^{-2}$.
3. $i = 2n, j = 2m, m \leq n$, $-K^{-1}N^{-2} \lesssim a_{i,j}(t) \lesssim -K^{-2}N^{-2}$.
4. $i = 2n - 1, j = 2m, m \leq n - 1$, $|a_{i,j}(t)| \lesssim K^{-2}N^{-2}$.
5. $i = 2n - 1, j = 2m - 1, m \leq n$, $|a_{i,j}(t)| \lesssim K^{-2}N^{-2}$.

As a result of Theorem J.14, for any $l \in [N]$, we can characterize the relative increments gap of A across different tokens as follows.

Lemma J.15. *Suppose Induction Hypothesis J.2 and Theorems J.13 and J.14 hold at iteration $T_1^{(1)} + 1 \leq t \leq T_2^{(1)}$. For any $l \in [N]$, we have the following results.*

1. ***Upper Bound:*** $\Delta A_{l,l-1}(t) - \max_{r \leq l, r \neq l-1} \Delta A_{l,r}(t) \gtrsim \eta_1 C_1 K^{-2}$.
2. ***Lower Bound:*** $\max_{j \leq l, j \neq l-1} \Delta A_{l,j}(t) - \min_{r \leq l, r \neq l-1} \Delta A_{l,r}(t) \lesssim \eta_1 C_1 (KN)^{-1}$.

As a direct corollary Theorem J.15, we can bound the increment differences' ratios as follows.

Corollary J.16. *Under the same condition of Theorem J.15, for any $l \in [N]$, we have*

$$\frac{\max_{j \leq l, j \neq l-1} \Delta A_{l,j}(t) - \min_{r \leq l, r \neq l-1} \Delta A_{l,r}(t)}{\min_{l \in [N]} \{\Delta A_{l,l-1}(t) - \max_{r \leq l, r \neq l-1} \Delta A_{l,r}(t)\}} = O\Big(\frac{K}{N}\Big),$$
$$\frac{\Delta A_{l,l-1}(t) - \min_{r \leq l, r \neq l-1} \Delta A_{l,r}(t)}{\min_{l \in [N]} \{\Delta A_{l,l-1}(t) - \max_{r \leq l, r \neq l-1} \Delta A_{l,r}(t)\}} = 1 + O\Big(\frac{K}{N}\Big),$$
$$\frac{\Delta A_{l,l-1}(t) - \max_{r \leq l, r \neq l-1} \Delta A_{l,r}(t)}{\min_{l \in [N]} \{\Delta A_{l,l-1}(t) - \max_{r \leq l, r \neq l-1} \Delta A_{l,r}(t)\}} = 1 + O\Big(\frac{K}{N}\Big).$$

We omit the proof here, as it follows directly from Theorem J.15.

J.5.2. End of Phase II

Lemma J.17. *With $T_2^{(1)} - T_1^{(1)}$ at most $O\left(\eta_1^{-1} C_1^{-1} K^2 \log K\right)$, and at iteration $t = T_2^{(1)} + 1$, for any $i \in [N]$, we have*

1. $A_{i,i-1}(T_2^{(1)} + 1) - \max_{j \leq i, j \neq i-1} A_{i,j}(T_2^{(1)} + 1) \geq \log(KN)$,
2. $1 - S_{i,i-1}^{(1)}(T_2^{(1)} + 1) = O(1/K)$.

Proof Claim (1) holds from the definition of Phase II. We only need to show the upper bound of $T_2^{(1)}$. As Induction Hypothesis J.2 holds, there exists an i for $T_2^{(1)}$ such that

$$\log(KN) \geq A_{i,i-1}(T_2^{(1)}) - \max_{j \leq i, j \neq i-1} A_{i,j}(T_2^{(1)})$$

$$\geq \sum_{\tau=T_1^{(1)}+1}^{T_2^{(1)}-1} \Delta A_{i,i-1}(\tau) - \max_{j\leq i, j\neq i-1} \Delta A_{i,j}(\tau) + A_{i,i-1}(T_1^{(1)}+1) - \max_{j\leq i, j\neq i-1} A_{i,j}(T_1^{(1)}+1)$$
$$\overset{(i)}{\geq} (T_2^{(1)} - T_1^{(1)} - 1)\frac{\eta_1 C_1}{K^2} + \log(N),$$

where (i) follows from Theorem J.15 and Induction Hypothesis J.1. By direct calculation, we have $T_2^{(1)} - T_1^{(1)} = O\left(\eta_1^{-1} C_1^{-1} K^2 \log K\right)$. Hence, we prove Claim (1).

We then prove Claim (2), for all $i \in [N]$, we have

$$1 - S_{i,i-1}^{(1)}(T_2^{(1)}+1) = \frac{\sum_{j\leq i, j\neq i-1} \exp\left(A_{i,j}(T_2^{(1)}+1) - A_{i,i-1}(T_2^{(1)}+1)\right)}{1 + \sum_{j\leq i, j\neq i-1} \exp\left(A_{i,j}(T_2^{(1)}+1) - A_{i,i-1}(T_2^{(1)}+1)\right)}$$
$$\overset{(i)}{\leq} \frac{(i-1)\exp\left(-\log(KN)\right)}{1 + (i-1)\exp\left(-\log(KN)\right)} \leq O\Big(\frac{i}{KN}\Big),$$

where (i) follows from that $x/(1+x)$ increase with x and $\max_{j\leq i, j\neq i-1} A_{i,j}(T_2^{(1)}+1) - A_{i,i-1}(T_2^{(1)}+1) \leq -\log(KN)$. Hence, we finish the proof of this lemma. ■

J.6. Stage I: Phase III

In this section, we study the convergence phase. For $\epsilon_1 = O(N^{-\frac{1}{2}})$, we define the **Phase III** as all iterations $T_2^{(1)} + 1 \leq t \leq T_3^{(1)}$, where

$$T_3^{(1)} \triangleq \max\Big\{t \geq T_2^{(1)} : \min_{i\in[N]} \Big(A_{i,i-1}(t) - \max_{j\leq i, j\neq i-1} A_{i,j}(t)\Big) \leq \log\left(N{\epsilon_1}^{-1}\right)\Big\}.$$

We then state the following induction hypothesis, which holds throughout Phase III. This hypothesis is proved by induction with the technical lemmas in Section J.6.1.

Induction Hypothesis J.3. *For each $T_2^{(1)} + 1 \leq t \leq T_3^{(1)}$ and any $i \in [N]$, the following holds:*

1. *$A_{i,i-1}(t) - \max_{j\leq i, j\neq i-1} A_{i,j}(t)$ is monotonically increasing and $A_{i,i-1}(t) - \max_{j\leq i, j\neq i-1} A_{i,j}(t) \in \big[\log(N), \big(1 + O((KN\epsilon_1)^{-1} + K^{-1}\log N)\big)\log\left(N{\epsilon_1}^{-1}\right)\big]$,*
2. *$\max_{j\leq i, j\neq i-1} A_{i,j}(t) - \min_{l\leq i, l\neq i-1} A_{i,l}(t) \leq O\left(K^{-1}(\log N)^2 + (KN\epsilon_1)^{-1}\log\left(N{\epsilon_1}^{-1}\right)\right)$.*

Proof We first prove Claim (1), and the monotonicity can be proved similarly to Induction Hypothesis J.2. Following from Theorem J.20, we have

$$\Delta A_{i,i-1}(t) - \max_{j\leq i, j\neq i-1} \Delta A_{i,j}(t) \gtrsim \frac{\eta_1 C_1 \epsilon_1}{K} \geq 0.$$

Furthermore, following from Theorem J.21, and following the same steps as in (45), except for the change in coefficients, we obtain

$$A_{i,i-1}(t) - \max_{j\leq i, j\neq i-1} A_{i,j}(t) - \big(A_{i,i-1}(T_2^{(1)}) - \max_{j\leq i, j\neq i-1} A_{i,j}(T_2^{(1)})\big)$$
$$\leq \Big(1 + O\Big(\frac{K}{N}\Big)\Big)\Big\{\min_{l\in[N]} \Big(A_{l,l-1}(t) - \max_{r\leq l, r\neq l-1} A_{l,r}(t)\Big) - \min_{l\in[N]} \Big(A_{l,l-1}(T_2^{(1)}) - \max_{r\leq l, r\neq l-1} A_{l,r}(T_2^{(1)})\Big)\Big\}.$$

By direct calculation, we have

$$A_{i,i-1}(t) - \max_{j\leq i, j\neq i-1} A_{i,j}(t)$$

$$
\begin{aligned}
&\overset{(i)}{\leq} \Big(1 + O\Big(\frac{1}{KN\epsilon_1}\Big) + O\Big(\frac{\log N}{K}\Big)\Big) \min_{l\in[N]} \Big(A_{l,l-1}(t) - \max_{r\leq l, r\neq l-1} A_{l,r}(t)\Big) \\
&\leq \Big(1 + O\Big(\frac{1}{KN\epsilon_1}\Big) + O\Big(\frac{\log N}{K}\Big)\Big) \log\left(N{\epsilon_1}^{-1}\right),
\end{aligned}
$$

where (i) follows from Induction Hypothesis J.2 and Equation (46), and that $A_{l,l-1}(t) - \max_{r\leq l, r\neq l-1} A_{l,r}(t)$ is monotonically increasing with t. Hence, we prove Claim (1).

We next prove Claim (2),

$$
\begin{aligned}
&\max_{j\leq i, j\neq i-1} A_{i,j}(t) - \min_{l\leq i, l\neq i-1} A_{i,l}(t) - \Big(\max_{j\leq i, j\neq i-1} A_{i,j}(T_2^{(1)}) - \min_{l\leq i, l\neq i-1} A_{i,l}(T_2^{(1)})\Big) \\
&\leq \sum_{\tau=T_2^{(1)}}^{t-1} \Big(\max_{j\leq i, j\neq i-1} \Delta A_{i,j}(\tau) - \min_{l\leq i, l\neq i-1} \Delta A_{i,l}(\tau)\Big) \\
&\leq \sum_{\tau=T_2^{(1)}}^{t-1} O\Big(\frac{1}{KN\epsilon_1}\Big)\Big(\Delta A_{i,i-1}(\tau) - \max_{j\leq i, j\neq i-1} \Delta A_{i,j}(\tau)\Big) \\
&= O\Big(\frac{1}{KN\epsilon_1}\Big)\Big(A_{i,i-1}(t) - \max_{j\leq i, j\neq i-1} A_{i,j}(t) - \Big(A_{i,i-1}(T_2^{(1)}) - \max_{j\leq i, j\neq i-1} A_{i,j}(T_2^{(1)})\Big)\Big), \qquad (47)
\end{aligned}
$$

where the last inequality follows from Theorem J.21. By direct calculation, we have

$$
\begin{aligned}
&\max_{j\leq i, j\neq i-1} A_{i,j}(t) - \min_{l\leq i, l\neq i-1} A_{i,l}(t) \\
&\overset{(i)}{\leq} O\Big(\frac{(\log N)^2}{K}\Big) + O\Big(\frac{1}{KN\epsilon_1}\Big)\Big(1 + O\Big(\frac{1}{KN\epsilon_1}\Big) + O\Big(\frac{\log N}{K}\Big)\Big) \log\left(N{\epsilon_1}^{-1}\right) \\
&\leq O\Big(\frac{(\log N)^2}{K} + \frac{\log(N{\epsilon_1}^{-1})}{KN\epsilon_1}\Big),
\end{aligned}
$$

where (i) follows from Induction Hypothesis J.2 and Equation (47). Hence, we finish the proof of Induction Hypothesis J.3. ■

J.6.1. Technical Lemmas

Similar to Phases I and II, we next introduce several useful technical lemmas. The proofs follow arguments analogous to those in Phase I and are omitted for brevity.

Lemma J.18. *Suppose Induction Hypothesis J.3 holds at iteration t, and Theorems J.19 and J.20 hold at iteration $t-1$. For iteration $T_2^{(1)} + 1 \leq t \leq T_3^{(1)}$, we have the following results.*

1. *For any $i \in [N]$,*
   - *(a)* $S_{i,i-1}^{(1)}(t) = \Omega(1)$,
   - *(b)* $\Omega(\epsilon_1/N) \leq S_{i,j}^{(1)}(t) \leq O(1/(KN)), j \neq i-1$.
2. *For any $i \in [N]$,*
   - *(a)* $S_{i,\mathbf{x}}^{(1)}(t) - S_{i,i-1}^{(1)}(t) \geq \Omega(i\epsilon_1/N)$,
   - *(b)* $S_{i,y}^{(1)}(t) \geq \Omega(i\epsilon_1/N)$,
   - *(c)* $1 - S_{i,i-1}^{(1)}(t) \geq \Omega(i\epsilon_1/N)$.

Similar to Theorems J.9 and J.14, and following from the expression of $a_{i,j}(t)$ in Theorem J.3, control of attention scores in Theorems J.7 and J.18, we can control the order of $a_{i,j}(t)$.

Lemma J.19 (Order of $a_{i,j}(t)$)**.** *Suppose Induction Hypothesis J.3 and Theorem J.18 hold at iteration $T_2^{(1)}+1 \le t \le T_3^{(1)}$. For different cases of i, j, we have the following results.*

1. *For $i = 2n, j = 2n-1$, $a_{i,i-1}(t) \gtrsim i\epsilon_1/(KN^2)$.*
2. *For $i = 2n, j = 2m-1, m \le n$, $-1/(K^2N^2) \lesssim a_{i,j}(t) \lesssim -\epsilon_1/(KN^2)$.*
3. *For $i = 2n, j = 2m, m \le n$, $-1/(K^2N^2) \lesssim a_{i,j}(t) \lesssim -\epsilon_1/(KN^2)$.*
4. *For $i = 2n-1, j = 2m, m \le n-1$, $|a_{i,j}(t)| \lesssim 1/(K^3N^2)$.*
5. *For $i = 2n-1, j = 2m-1, m \le n$, $|a_{i,j}(t)| \lesssim 1/(K^3N^2)$.*

As a result of Theorem J.19, for any $l \in [N]$, we can characterize the relative increments gap of A across different tokens as follows.

Lemma J.20. *Suppose Induction Hypothesis J.3 and Theorems J.18 and J.19 hold at iteration $T_2^{(1)}+1 \le t \le T_3^{(1)}$, for any $l \in [N]$, the following holds*

1. ***Upper Bound:*** $\Delta A_{l,l-1}(t) - \max_{r \le l, r \ne l-1} \Delta A_{l,r}(t) \gtrsim K^{-1}\eta_1 C_1 \epsilon_1$.
2. ***Lower Bound:*** $\max_{j \le l, j \ne l-1} \Delta A_{l,j}(t) - \min_{r \le l, r \ne l-1} \Delta A_{l,r}(t) \lesssim K^{-2}N^{-1}\eta_1 C_1 \log N$.

As a direct corollary of Theorem J.20, we can bound the increment differences' ratios as follows.

Corollary J.21. *Under the same condition of Theorem J.20, for any $l \in [N]$, we have*

$$\frac{\max_{j \le l, j \ne l-1} \Delta A_{l,j}(t) - \min_{r \le l, r \ne l-1} \Delta A_{l,r}(t)}{\min_{l \in [N]}\{\Delta A_{l,l-1}(t) - \max_{r \le l, r \ne l-1} \Delta A_{l,r}(t)\}} = O\Big(\frac{1}{KN\epsilon_1}\Big),$$
$$\frac{\Delta A_{l,l-1}(t) - \min_{r \le l, r \ne l-1} \Delta A_{l,r}(t)}{\min_{l \in [N]}\{\Delta A_{l,l-1}(t) - \max_{r \le l, r \ne l-1} \Delta A_{l,r}(t)\}} = 1 + O\Big(\frac{1}{KN\epsilon_1}\Big),$$
$$\frac{\Delta A_{l,l-1}(t) - \max_{r \le l, r \ne l-1} \Delta A_{l,r}(t)}{\min_{l \in [N]}\{\Delta A_{l,l-1}(t) - \max_{r \le l, r \ne l-1} \Delta A_{l,r}(t)\}} = 1 + O\Big(\frac{1}{KN\epsilon_1}\Big).$$

We omit the proof here, as it follows directly from Theorem J.20.

J.6.2. End of Phase III

Lemma J.22. *With $T_3^{(1)} - T_2^{(1)}$ at most $O\left(\eta_1^{-1}C_1^{-1}\epsilon_1^{-1}K\log(K^{-1}\epsilon_1^{-1})\right)$, and at iteration $t = T_3^{(1)}+1$, for any $i \in [N]$, we have*

1. $A_{i,i-1}(T_3^{(1)}+1) - \max_{j \le i, j \ne i-1} A_{i,j}(t) \ge \log(N\epsilon_1^{-1})$,
2. $1 - S_{i,i-1}^{(1)}(T_3^{(1)}+1) \le \epsilon_1$.

Proof Claim (1) holds from the definition of Phase III, so we only need to show the upper bound of $T_3^{(1)}$. As Induction Hypothesis J.3 holds, there exists an i for $T_3^{(1)}$ such that,

$$\begin{aligned}
\log\left(N{\epsilon_1}^{-1}\right) &\ge A_{i,i-1}(T_3^{(1)}) - \max_{j \le i, j \ne i-1} A_{i,j}(T_3^{(1)}) \\
&\ge \sum_{\tau=T_2^{(1)}+1}^{T_3^{(1)}-1} \Delta A_{i,i-1}(\tau) - \max_{j \le i, j \ne i-1} \Delta A_{i,j}(\tau) + A_{i,i-1}(T_2^{(1)}+1) - \max_{j \le i, j \ne i-1} A_{i,j}(T_2^{(1)}+1) \\
&\overset{(i)}{\ge} (T_3^{(1)} - T_2^{(1)} - 1)\frac{\eta_1 C_1 \epsilon_1}{K} + \log(KN),
\end{aligned}$$

where (i) follows from Theorem J.20 and Induction Hypothesis J.2. As a result, $T_3^{(1)} - T_2^{(1)} = O\left(\eta_1^{-1}C_1^{-1}\epsilon_1^{-1}K\log(K^{-1}\epsilon_1^{-1})\right)$.

We then prove Claim (2), for all $i \in [N]$, we have

$$\begin{aligned}
1 - S_{i,i-1}^{(1)}(T_3^{(1)}+1) &= \frac{\sum_{j\le i, j\ne i-1}\exp\left(A_{i,j}(T_3^{(1)}+1) - A_{i,i-1}(T_3^{(1)}+1)\right)}{1+\sum_{j\le i, j\ne i-1}\exp\left(A_{i,j}(T_3^{(1)}+1) - A_{i,i-1}(T_3^{(1)}+1)\right)} \\
&\le \frac{(i-1)\exp\left(\max_{j\le i, j\ne i-1} A_{i,j}(T_3^{(1)}+1) - A_{i,i-1}(T_3^{(1)}+1)\right)}{1+(i-1)\exp\left(\max_{j\le i, j\ne i-1} A_{i,j}(T_3^{(1)}+1) - A_{i,i-1}(T_3^{(1)}+1)\right)} \\
&\overset{(i)}{\le} \frac{(i-1)e^{-\log(N\epsilon_1^{-1})}}{1+(i-1)e^{-\log(N\epsilon_1^{-1})}} \\
&\le \epsilon_1,
\end{aligned}$$

where (i) follows from that $x/(1+x)$ increase with x and $\max_{j\le i, j\ne i-1} A_{i,j}(T_2^{(1)}+1) - A_{i,i-1}(T_2^{(1)}+1) \le -\log(N\epsilon_1^{-1})$. Hence, we finish the proof of this lemma. ■

J.7. Proof of Stage I of Theorem D.3

Proof As a direct result of Theorems J.12, J.17 and J.22, with at most

$$\begin{aligned}
\tau_1 &= T_3^{(1)} + 1 \\
&= O\Big(\frac{KN\log N}{\eta_1 C_1}\Big) + O\Big(\frac{K^2\log K}{\eta_1 C_1}\Big) + O\Big(\frac{K\log(K^{-1}\epsilon_1^{-1})}{\eta_1 C_1 \epsilon_1}\Big) \\
&= O\Big(\frac{KN\log N}{\eta_1 C_1} + \frac{K\log(K^{-1}\epsilon_1^{-1})}{\eta_1 C_1 \epsilon_1}\Big),
\end{aligned}$$

we have $1 - S_{i,i-1}^{(1)}(\tau_1) \le \epsilon_1$. Hence, we finish the proof. ■

K. Proof of Stage II of Theorem D.3

In Stage II, $\widetilde{W}_Q^{(1)}$ is fixed as $\widetilde{W}_Q^{(1)}(\tau_1)$. As a result, $S_{i,j}^{(1)}$ is fixed as $S_{i,j}^{(1)}(\tau_1)$. We omit τ_1 for simplicity. We also omit (t) and write $L(\tilde{\theta})$ as L here when there is no ambiguity.

K.1. Roadmap of the Proof

We analyze the feature matching in stage II via two phases of dynamics (Sections K.4 and K.5). In each phase, we formulate an induction hypothesis and derive several lemmas describing the values and evolution of key statistics, which collectively govern the attention scores. We prove the induction hypothesis and the lemmas by induction.

We begin by introducing the key statistic that is tracked throughout the different phases. Notice that after stage I training, Layer 1 attention scores almost concentrate on the immediate prefix, i.e., for each l, $1 - S_{l,l-1}^{(1)}(t) \le \epsilon_1$. Under this condition, and by Theorem K.2, we observe that Layer 2 attention logit $\mathbf{u}_i(t)$ depends strongly on the features associated with $\mathbf{x}_\mathrm{q}$ and $\mathbf{x}_i$. Since the frequencies are sufficiently small, given a fixed $\mathbf{x}_\mathrm{q}$, both $\mathbf{u}_i(t)$ and its update take nearly the same value across all positions i where $\mathbf{x}_i$ share the same feature $\mathbf{v}_m, m \in [K]$. Consequently, we denote this common value of $\mathbf{u}_i(t)$ by $B_m(t)$ and that of its update by $b_m(t)$ for feature $\mathbf{v}_m$. Both quantities are random, as they depend on the random variable $\mathbf{x}_\mathrm{q}$.

The main idea of the proof lies in analyzing the GD dynamics of $B_m(t)$, which decides Layer 2 attention scores $S_i^{(2)}(t)$ and $\mathcal{S}_m^{(2)}(t)$. By tracking $B_m(t)$, we obtain the attention scores on the features, $\mathcal{S}_m^{(2)}(t)$. Note that when $\mathbf{x}_\mathrm{q} = \mathbf{v}_k$, achieving a good prediction only requires that $\mathcal{S}_k^{(2)}(t) \approx 1$.

For any $k \in [K]$, conditioned on $\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k$, the learning process can be divided into two phases.

- **Phase I: Growth** ($t \in [\tau_1 + 1, \tau_1 + T_{1,k}^{(2)}]$, Section K.4). During phase I, $B_k(t)$ keeps growing at a rate of $b_k(t) = \Omega(\eta K^{-2})$, while for $k' \neq k$, $B_{k'}(t)$ oscillates with a smaller rate of $b_{k'}(t)$, which satisfying $|b_{k'}(t)| = O(K^{-1} b_k(t))$. Therefore, the increase in $B_k(t)$ will dominate the learning dynamics during phase I. The attention score $\mathcal{S}_k^{(2)}(t)$ keeps increasing, while still satisfying $1 - \mathcal{S}_k^{(2)}(t) = \Omega(1)$.
- **Phase II: Convergence** ($t \in (\tau_1 + T_{1,k}^{(2)} + 1, \tau_1 + T_{2,k}^{(2)}]$, Section K.5). After the rapid growth of self-attention module parameters in phase I, the question token featuring $\mathbf{v}_k$ is aligned with these input tokens also featuring $\mathbf{v}_k$ effectively and disregards other features. Then the process proceeds to the convergence phase, where $B_k(t)$ monotonically increases and for $k' \neq k$, $B_{k'}^{(t)}$ monotonically decreases, which finally contributes to the convergence of the loss. As a result, $B_k(t)$ keeps increasing and for $k' \neq k$, $B_{k'}(t)$ keeps deceasing. After the end of phase II, $B_k(T_{1,k}^{(2)} + 1) \geq \log(K\epsilon_2^{-1})$, and $1 - \mathcal{S}_k^{(2)}(T_{1,k}^{(2)} + 1) \leq (\epsilon_2/2)^{\frac{1}{2}}$.

We summarize the upcoming sections as follows: In Section K.2, we compute and simplify the gradients to identify the key update variables. In Section K.3, we introduce several useful auxiliary lemmas for Stage II. In Sections K.4 and K.5, we analyze the two phases of the dynamics.

K.2. Stage II: Preliminary Development

Computations of Gradients. We first calculate Stage II gradient with respect to $\widetilde{W}_Q^{(2)}$.

Lemma K.1 (Layer 2 Gradient)**.** *In Stage II, the gradient of the loss function with respect to $\widetilde{W}_Q^{(2)}$ is given by*

$$\nabla_{\widetilde{W}_Q^{(2)}} L = \mathbb{E}\Big[(\hat{y}_{\mathrm{q}} - \langle \mathbf{w}, \mathbf{x}_{\mathrm{q}} \rangle) \sum_{\ell \in [N_{\mathrm{in}}]} y_\ell S_{2\ell}^{(2)} \cdot \sum_{i=1}^{N} (\mathbb{1}\{2\ell = i\} - S_i^{(2)}) \Big(R_{\boldsymbol{\vartheta}, N-i} \Big(S_{i,i-1}^{(1)} (E_{i-1}^{\mathbf{x},y})^\top + \sum_{j \neq i-1} S_{i,j}^{(1)} (E_j^{\mathbf{x},y})^\top \Big) E_{\mathrm{q}}^{\mathbf{x},y} \Big)^\top \Big],$$

where $E_j^{\mathbf{x},y}$ and $E_{\mathrm{q}}^{\mathbf{x},y}$ correspond to the semantic dependent subspaces of the token embeddings E_j and E_{q}, respectively.

Proof We first write out the expression of $\hat{y}_{\mathrm{q}}$. In Stage II, $\widetilde{W}_Q^{(1)}$ is fixed, then

$$\hat{y}_{\mathrm{q}}(W_{\mathbf{q}}(\tau_1), \widetilde{W}_Q^{(2)}) = \sum_{i=1}^{N} S_{2i}^{(2)} y_i = \sum_{k=1}^{K} \mathcal{S}_k^{(2)} \langle \mathbf{w}, \mathbf{v}_k \rangle.$$

Consider$S^{(2)} = \mathrm{softmax}(\mathbf{u}) \in \mathbb{R}^{1 \times N}$, and

$$\begin{aligned} \mathbf{u}_i &= E_{\mathrm{q}}^{\mathbf{x},y} \widetilde{W}_Q^{(2)} R_{\boldsymbol{\vartheta}, N-i} (S^{(1)} E^{\mathbf{x},y})_{i,:}^\top \\ &= E_{\mathrm{q}}^{\mathbf{x},y} \widetilde{W}_Q^{(2)}(t) R_{\boldsymbol{\vartheta}, N-i} \Big(S_{i,i-1}^{(1)} (E_{i-1}^{\mathbf{x},y})^\top + \sum_{j \neq i-1} S_{i,j}^{(1)} (E_j^{\mathbf{x},y})^\top \Big), \end{aligned} \tag{48}$$

where the last inequality follows from that $(S^{(1)} E^{\mathbf{x},y})_{i,:} = S_{i,:}^{(1)}(t) E^{\mathbf{x},y} = S_{i,i-1}^{(1)} E_{i-1,:}^{\mathbf{x},y} + \sum_{j \neq i-1} S_{i,j}^{(1)} E_{j,:}^{\mathbf{x},y}$. Then we can calculate the gradient. We first obtain:

$$\nabla_{\widetilde{W}_Q^{(2)}} L = \mathbb{E}[(\hat{y}_{\mathrm{q}} - \langle \mathbf{w}, \mathbf{x}_{\mathrm{q}} \rangle) \nabla_{\widetilde{W}_Q^{(2)}} \hat{y}_{\mathrm{q}}] = \mathbb{E}\Big[(\hat{y}_{\mathrm{q}} - \langle \mathbf{w}, \mathbf{x}_{\mathrm{q}} \rangle) \sum_{\ell \in [N_{\mathrm{in}}]} (\nabla_{\widetilde{W}_Q^{(2)}} S_{2\ell}^{(2)}) y_\ell \Big]. \tag{49}$$

Recall that for any $U \in \mathbb{R}^{1 \times d}$, $\nabla_U \mathrm{softmax}(U) = \mathrm{diag}(\mathrm{softmax}(U)) - \mathrm{softmax}(U)^\top \mathrm{softmax}(U)$, we have

$$\nabla_{\widetilde{W}_Q^{(2)}} S_{2\ell}^{(2)} = \sum_{i=1}^{N} \frac{\partial S_{2\ell}^{(2)}}{\partial \mathbf{u}_i} \nabla_{\widetilde{W}_Q^{(2)}} \mathbf{u}_i$$

$$= S_{2\ell}^{(2)} \sum_{i=1}^{N} (\mathbb{1}\{2\ell = i\} - S_i^{(2)}) \Big(R_{\boldsymbol{\vartheta},N-i} \Big(S_{i,i-1}^{(1)} (E_{i-1}^{\mathbf{x},y})^\top + \sum_{j \neq i-1} S_{i,j}^{(1)} (E_j^{\mathbf{x},y})^\top \Big) E_{\mathrm{q}}^{\mathbf{x},y} \Big)^\top. \tag{50}$$

Plug (50) into (49), we finish the proof of Theorem K.1. ■

We find that for stage II, $\mathbf{u}_i(t)$ is important in determining the attention scores. As a result, we track the update of $\mathbf{u}_i(t)$ as follows.

Lemma K.2 (Track variable update)**.** *In Stage II, the attention logit at position i can be expressed as*

$$\mathbf{u}_i(t) = E_{\mathrm{q}}^{\mathbf{x},y} \widetilde{W}_Q^{(2)}(t) R_{\boldsymbol{\vartheta},N-i} \Big(S_{i,i-1}^{(1)}(t) (E_{i-1}^{\mathbf{x},y})^\top + \sum_{j \neq i-1} S_{i,j}^{(1)}(t) (E_j^{\mathbf{x},y})^\top \Big).$$

Conditioned on the event that $\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k$, we can characterize the increments of $\mathbf{u}_i(t)$ over t for even i and upper bound $\mathbf{u}_i(t)$ for odd i as follows.

1. *For any $m \in [K]$, if i is even and $\mathbf{x}_{i/2} = \mathbf{v}_m$, we define the following variables conditioned on the features,*

$$B_{i,m}(t) = \mathbf{u}_i(t),$$

where the subscript m specifies the particular feature $\mathbf{v}_m$ taken by $\mathbf{x}_{i/2}$. Then for each $m \in [K]$, the update of $B_{i,m}(t)$ denoted by $b_{i,m}(t) = B_{i,m}(t+1) - B_{i,m}(t)$ can be characterized as follows.

   (a) *If $m = k$, then*

$$b_{i,k}(t) = \eta_2 \mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k\} \mathcal{S}_k^{(2)}(t) \Big((\mathcal{S}_k^{(2)}(t) - 1)^2 + \sum_{m \neq k} (\mathcal{S}_m^{(2)}(t))^2 \Big) \Big] \pm O\Big(\frac{\eta_2 \epsilon_1 d_{\mathcal{X}}}{K} \Big).$$

   (b) *For $m \neq k$, we have*

$$b_{i,m}(t) = \eta_2 \mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k\} \mathcal{S}_m^{(2)}(t) \Big(\sum_{m \in [K]} (\mathcal{S}_m^{(2)}(t))^2 - \mathcal{S}_m^{(2)}(t) - \mathcal{S}_k^{(2)}(t) \Big) \Big] \pm O\Big(\frac{\eta_2 \epsilon_1 d_{\mathcal{X}}}{K} \Big).$$

Since the dependence on i is negligible from RHS above, we omit the position subscripts i in $B_{i,m}(t)$ and $b_{i,m}(t)$ when no ambiguity arises.

2. *If i is odd, then $\mathbf{u}_i(t)$ is always relatively small and can be controlled*

$$\mathbf{u}_i(t) \leq \epsilon_1 \Theta\Big(\max_{m \in [K]} B_m(t) \Big).$$

Proof We first prove Claim (1), where i is even, and $i-1$ is odd. In the following, we omit (t) of $S^{(1)}(t), S^{(2)}(t)$ and $\mathcal{S}^{(2)}(t)$ for abbreviation. From (48), we have

$$\begin{aligned}
&\mathbf{u}_i(t+1) - \mathbf{u}_i(t) \\
&= -\eta_2 E_{\mathrm{q}}^{\mathbf{x},y} \nabla_{\widetilde{W}_Q^{(2)}} L(t) R_{\boldsymbol{\vartheta},N-i} \Big(S_{i,i-1}^{(1)} (E_{i-1}^{\mathbf{x},y})^\top + \sum_{j \neq i-1} S_{i,j}^{(1)} (E_j^{\mathbf{x},y})^\top \Big) \\
&\overset{(i)}{=} -\eta_2 E_{\mathrm{q}}^{\mathbf{x},y} \mathbb{E}\Big[(\hat{y}_{\mathrm{q}} - \langle \mathbf{w}, \mathbf{x}_{\mathrm{q}} \rangle) \sum_{\ell \in [N_{\mathrm{in}}]} y_\ell S_{2\ell}^{(2)} \sum_{m=1}^{N} (\mathbb{1}\{2\ell = m\} - S_m^{(2)}) \\
&\quad \cdot \Big((E_{\mathrm{q}}^{\mathbf{x},y})^\top \Big(S_{m,m-1}^{(1)} E_{m-1}^{\mathbf{x},y} + \sum_{j \neq m-1} S_{m,j}^{(1)} E_j^{\mathbf{x},y} \Big) R_{\boldsymbol{\vartheta},m-i} \Big) \Big] \Big(S_{i,i-1}^{(1)} (E_{i-1}^{\mathbf{x},y})^\top + \sum_{j \neq i-1} S_{i,j}^{(1)} (E_j^{\mathbf{x},y})^\top \Big) \\
&\overset{(ii)}{=} -\eta_2 \mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k\} (\hat{y}_{\mathrm{q}} - \langle \mathbf{w}, \mathbf{x}_{\mathrm{q}} \rangle) \sum_{\ell \in [N_{\mathrm{in}}]} y_\ell S_{2\ell}^{(2)} \sum_{m=1}^{N} (\mathbb{1}\{2\ell = m\} - S_m^{(2)}) \left(E_{m-1}^{\mathbf{x},y} R_{\boldsymbol{\vartheta},m-i} \right) \Big]
\end{aligned}$$

$$\cdot\left(E_{i-1}^{\mathbf{x},y}\right)^\top \pm O\Big(\eta_2\epsilon_1\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\Big|\left(\hat{y}_{\mathrm{q}}-\langle\mathbf{w},\mathbf{x}_{\mathrm{q}}\rangle\right)\sum_{\ell\in[N_{\mathrm{in}}]}y_\ell S_{2\ell}^{(2)}\Big|\Big]\Big)$$

$$\overset{(iii)}{=}-\eta_2\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}(\hat{y}_{\mathrm{q}}-\langle\mathbf{w},\mathbf{x}_{\mathrm{q}}\rangle)\sum_{\ell\in[N_{\mathrm{in}}]}y_\ell S_{2\ell}^{(2)}\sum_{m=1}^{N}S_m^{(2)}\left(E_{2\ell-1}^{\mathbf{x},y}R_{\boldsymbol{\vartheta},2\ell-i}-E_{m-1}^{\mathbf{x},y}R_{\boldsymbol{\vartheta},m-i}\right)\Big]$$

$$\cdot\left(E_{i-1}^{\mathbf{x},y}\right)^\top \pm O\left(K^{-1}\eta_2\epsilon_1 d_{\mathcal{X}}\right), \tag{51}$$

where (i) follows from Theorem K.1 and the property of RoPE operator, (ii) follows from $E_{\mathrm{q}}^{\mathbf{x},y}(E_{\mathrm{q}}^{\mathbf{x},y})^\top=1$, and the property that $1-S_{i,i-1}^{(1)}(\tau_1)\le\epsilon_1$ for all $i\in[N]$ at the end of Stage I, as in Theorem J.22, and (iii) follows from that $\|\mathbf{w}\|_2\le\sqrt{d_{\mathcal{X}}}$ and $\|\mathbf{v}_k\|_2=1$ for all $k\in[K]$.

For the dominant term (first term) in (51), conditioned on the event $\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k$, we can further divide it into two terms and discuss by cases.

Case (a): If the corresponding $\mathbf{x}_{i/2}$ of $E_{i-1}^{\mathbf{x},y}$ is $\mathbf{v}_k$, then with $\hat{y}_{\mathrm{q}}-\langle\mathbf{w},\mathbf{x}_{\mathrm{q}}\rangle=\mathbf{w}^\top(\sum_{\ell\in[N_{\mathrm{in}}]}S_{2\ell}^{(2)}\mathbf{x}_\ell-\mathbf{x}_{\mathrm{q}})$, we have the first term as

$$\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\left(\hat{y}_{\mathrm{q}}-\langle\mathbf{w},\mathbf{x}_{\mathrm{q}}\rangle\right)\sum_{\ell\in[N_{\mathrm{in}}]}y_\ell S_{2\ell}^{(2)}\sum_{m=1}^{N}S_m^{(2)}E_{2\ell-1}^{\mathbf{x},y}R_{\boldsymbol{\vartheta},2\ell-i}\Big]\left(E_{i-1}^{\mathbf{x},y}\right)^\top$$

$$\overset{(i)}{=}\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\Big(\sum_{r\in[N_{\mathrm{in}}]}S_{2r}^{(2)}\mathbf{x}_r-\mathbf{x}_{\mathrm{q}}\Big)^\top\sum_{\ell\in[N_{\mathrm{in}}]}\mathbf{x}_\ell S_{2\ell}^{(2)}\left(\mathbf{x}_\ell\right)^\top R_{\boldsymbol{\vartheta},2\ell-i}\Big]\mathbf{v}_k$$

$$\overset{(ii)}{=}\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\Big(\sum_{r\in[N_{\mathrm{in}}]}S_{2r}^{(2)}\mathbf{x}_r-\mathbf{x}_{\mathrm{q}}\Big)^\top\sum_{\ell\in[N_{\mathrm{in}}]}\mathbf{x}_\ell S_{2\ell}^{(2)}\big(\mathbb{1}\{\mathbf{x}_\ell=\mathbf{v}_k\}\pm\mathbb{1}\{\mathbf{x}_\ell\neq\mathbf{v}_k\}O\left(N^{-\alpha+1}\right)\big)\Big]$$

$$=\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\big(\mathcal{S}_k^{(2)}\big(\mathcal{S}_k^{(2)}-1\big)\pm\sum_{m\neq k}(\mathcal{S}_m^{(2)})^2O\left(N^{-\alpha+1}\right)\big)\Big], \tag{52}$$

where (i) follows from taking the conditional expectation over $\mathbf{w}$ and that the last 2-dimensional subspace of embedding of the input is a null space, and (ii) follows similarly from Theorem J.4. And the second term is

$$\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\left(\hat{y}_{\mathrm{q}}-\langle\mathbf{w},\mathbf{x}_{\mathrm{q}}\rangle\right)\sum_{\ell\in[N_{\mathrm{in}}]}y_\ell S_{2\ell}^{(2)}\sum_{m=1}^{N}S_m^{(2)}\left(-E_{m-1}^{\mathbf{x},y}R_{\boldsymbol{\vartheta},m-i}\right)\Big]\left(E_{i-1}^{\mathbf{x},y}\right)^\top$$

$$\overset{(i)}{=}\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\Big(\sum_{r\in[N_{\mathrm{in}}]}S_{2r}^{(2)}\mathbf{x}_r-\mathbf{x}_{\mathrm{q}}\Big)^\top\Big(\sum_{\ell\in[N_{\mathrm{in}}]}\mathbf{x}_\ell S_{2\ell}^{(2)}\Big)\sum_{m=1}^{N}S_m^{(2)}\left(-E_{m-1}^{\mathbf{x},y}R_{\boldsymbol{\vartheta},m-i}\right)\Big]\left(E_{i-1}^{\mathbf{x},y}\right)^\top$$

$$\overset{(ii)}{=}-\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\Big(\sum_{m\in[K]}(\mathcal{S}_m^{(2)})^2-\mathcal{S}_k^{(2)}\Big)\Big(\mathcal{S}_k^{(2)}(1-\Omega\left(N^{-\alpha+1}\right))\pm\sum_{m\neq k}\mathcal{S}_m^{(2)}O(N^{-\alpha+1})\Big)\Big], \tag{53}$$

where (i) follows from taking the conditional expectation over $\mathbf{w}$ and that the last 2-dimensional subspace of embedding of the input is a null space, and (ii) follows from the fact that the summation over ℓ and r are independent of the summation over m, and Theorem J.4. Sum (52) and (53), we get the dominant term in (51) as follow

$$-\eta_2\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\Big(\mathcal{S}_k^{(2)}\Big(-(\mathcal{S}_k^{(2)}-1)^2-\sum_{m\neq k}(\mathcal{S}_m^{(2)})^2\Big)\pm O\left(N^{-\alpha+1}\right)\Big)\Big].$$

Plugging into (51), we have

$$b_{i,k}(t)=\eta_2\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\mathcal{S}_k^{(2)}\Big((\mathcal{S}_k^{(2)}-1)^2+\sum_{m\neq k}(\mathcal{S}_m^{(2)})^2\Big)\Big]\pm O\left(\frac{\eta_2\epsilon_1 d_{\mathcal{X}}}{K}\right).$$

Case (b): If the corresponding $\mathbf{x}_{i/2}$ of $E_{i-1}^{\mathbf{x},y}$ is $\mathbf{v}_m, m\neq k$, similar to Case (a), we can get

$$b_{i,m}(t)=\eta_2\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\mathcal{S}_m^{(2)}\Big(\mathcal{S}_m^{(2)}+\mathcal{S}_k^{(2)}-\sum_{m\in[K]}(\mathcal{S}_m^{(2)})^2\Big)\Big]\pm O\left(\frac{\eta_2\epsilon_1 d_{\mathcal{X}}}{K}\right).$$

We note that with the low frequency small enough, the dependence of the increments of $\mathbf{u}_i(t)$ on the position i is dominated and can be controlled by a minor term. Since the dependence on i is negligible, we omit the position subscripts i when no ambiguity arises. Hence we prove Claim (1).

We next prove Claim (2), where i is odd, and $i-1$ is even. Then, $E_{\mathrm{q}}^{\mathbf{x},y}\widetilde{W}_Q^{(2)}R_{\boldsymbol{\vartheta},N-i}S_{i,i-1}^{(1)}(E_{i-1}^{\mathbf{x},y})^\top = 0$, as a result,

$$\begin{aligned}\mathbf{u}_i(t) &= E_{\mathrm{q}}^{\mathbf{x},y}\widetilde{W}_Q^{(2)}(t)R_{\boldsymbol{\vartheta},N-i}\Big(\sum_{j\neq i-1} S_{i,j}^{(1)}(E_j^{\mathbf{x},y})^\top\Big)\\ &\overset{(i)}{\leq} \epsilon_1 \max_{m\in[K]} \mathbf{v}_k^\top \widetilde{W}_Q^{(2)}(t)R_{\boldsymbol{\vartheta},N-i}\mathbf{v}_m\\ &\overset{(ii)}{\leq} \epsilon_1\Theta\Big(\max_{m\in[K]} B_m(t)\Big),\end{aligned}$$

where (i) follows from the property that $\sum_{j\neq i-1} S_{i,j}^{(1)}(\tau_1) = 1 - S_{i,i-1}^{(1)}(\tau_1) \leq \epsilon_1$ for all $i\in[N]$ at the end of Stage I, as in Theorem J.22, and that the last 2-dimensional subspace of embedding of the input is a null space, and (ii) follows from the definition of $B_m(t)$. Hence, we prove Theorem K.2. ■

K.3. Stage II: Auxiliary Lemmas

Recall that given a prompt $P = (\mathbf{x}_1, y_1, \ldots, \mathbf{x}_{N_{\mathrm{in}}}, y_{N_{\mathrm{in}}}, \mathbf{x}_{\mathrm{q}})$, we denote P_{in} as the collection of input tokens in the example, i.e., $\{\mathbf{x}_i\}_{i=1}^{N_{\mathrm{in}}}$. Similar to (Huang et al., 2024), it is worth noting that, based on our data distribution, the occurrence count of the k-th feature in P_{in}, denoted as $|\mathcal{V}_k|$, follows a multinomial distribution. Leveraging the concentration property inherent to multinomial distributions, we can identify a high-probability event to which P_{in} belongs. As $N = 2N_{\mathrm{in}}+1$, we can use N instead of N_{in} in the following lemma.

Lemma K.3 (High-probability Event for P_{in})**.** *Suppose that $p_k = \Theta(1/K)$ for any $k\in[K]$ and $K^3 \ll N$. For some constant $c \geqslant \sqrt{20K^3/N}$, define*

$$\mathcal{E}_{\mathrm{P}}^* := \Big\{P_{\mathrm{in}} : |\mathcal{V}_k| \in \Big[p_kN - \frac{cN}{K}, p_kN + \frac{cN}{K}\Big] \text{ for } k\in[K]\Big\}.$$

Then , we have

$$\mathbb{P}(P_{\mathrm{in}} \in \mathcal{E}_{\mathrm{P}}^*) \geq 1 - 3\exp\left(-\frac{c^2N}{25K^2}\right).$$

Let us denote $L_k = p_kK - c$ and $U_k = p_kK + c$. Note that L_k, U_k are at the order of the constant level since $p_k = \Theta(1/K)$. Then for any P_{in} belonging to $\mathcal{E}_{\mathrm{P}}^*$, $|\mathcal{V}_k| \in [L_kN/K, U_kN/K] = \Theta(N/K)$. Note that we can properly choose c to guarantee $L_k > 0$ for $k\in[K]$.

K.4. Stage II: Phase I

In this section, we study the initial phase of learning the relationship between the question token E_{q} and its previous tokens with different features. We define the **Phase I** as all iterations $\tau_1 + 1 \leq t \leq \tau_1 + T_{1,k}^{(2)}$, where

$$T_{1,k}^{(2)} \triangleq \max\{t : B_k(\tau_1 + t) \leq \log K\}.$$

Then, we state the following induction hypothesis, which holds throughout Phase I. This hypothesis is proved by induction with the technical lemmas in Section K.4.1.

Induction Hypothesis K.1. *Suppose Theorems K.4 and K.5 and Induction Hypothesis K.1 hold at iteration $t-1$, where $\tau_1 + 1 \leq t \leq \tau_1 + T_{1,k}^{(2)}$. Given any $k\in[K]$, conditioned on the events that $\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k$ and $P_{\mathrm{in}} \in \mathcal{E}_{\mathrm{P}}^*$, the following holds for iteration t:*

1. *$B_k(t)$ is monotonically increasing, and $B_k(t) \in [(1-\Omega(\epsilon_1 + N^{1-\alpha})), \log(K)]$.*

2. *For any $k' \neq k$, $|B_{k'}(t)| = O\left(B_k(t)/K\right) = O((\log K)/K)$.*

3. *Specially, for the initialization of Stage II, $t = \tau_1 + 1$,*

$$B_k(\tau_1+1) \in \left[1 - \Omega(\epsilon_1 + N^{1-\alpha}), 1 + O(\epsilon_1)\right], |B_{k'}(\tau_1+1)| \leq O\left(\epsilon_1\right).$$

Proof We first prove Claim (3). By the definition of B_m in Theorem K.2, we only need to consider odd positions. For any $i \in [N_{\text{in}}]$, by definition of $\mathbf{u}_{2i}(\tau_1+1)$ and $\widetilde{W}_Q^{(2)}(\tau_1+1) = I_d$, we have

$$\mathbf{u}_{2i}(\tau_1+1) = E_{\text{q}}^{\mathbf{x},y} R_{\boldsymbol{\vartheta},N-2i}\big(S_{2i,2i-1}^{(1)}(E_{2i-1}^{\mathbf{x},y})^\top + \sum_{j\neq i} S_{2i,j}^{(1)}(E_j^{\mathbf{x},y})^\top\big).$$

Then applying Theorem K.2, together with $\mathbf{x}_{\text{q}} = \mathbf{x}_i = \mathbf{v}_k$ and $1 - S_{i,i-1}^{(1)}(\tau_1+1) \leq \epsilon_1$ for all $i \in N_{\text{in}}$, we have

$$B_{2i,k}(\tau_1+1) = \big(S_{i,i-1}^{(1)}(1 - \Omega(N^{1-\alpha})) \pm O(\epsilon_1)\big) \in \left[(1 - \Omega(\epsilon_1 + N^{1-\alpha})), 1 + O(\epsilon_1)\right].$$

Hence, $B_k(\tau_1+1) \in \left[(1 - \Omega(\epsilon_1 + N^{1-\alpha})), 1 + O(\epsilon_1)\right]$. And for $k' \neq k$, we have

$$|B_{2i,k'}(\tau_1+1)| = |\mathbf{u}_{2i}(\tau+1)|\{\mathbf{x}_i = \mathbf{v}_{k'}\}| = O\left(\epsilon_1\right).$$

Hence $B_{k'}(\tau_1+1) = O\left(\epsilon_1\right)$, and we finish the proof of Claim (3).

We next prove Claim (1). For any $\tau_1 + 1 \leq t \leq \tau_1 + T_{1,k}^{(2)}$, we first need to show $b_k(t) \geq 0$, which is obvious from Theorem K.5. The lower bound follows from Claim (3), and the upper bound follows from the definition of Phase I.

We finally prove Claim (2). By Theorem K.5, we have for any $k' \neq k$, $|b_{k'}(t-1)| \leq O\left(b_k(t-1)/K\right)$ at time $t-1$. Then,

$$|B_{k'}(t)| \leq |B_{k'}(t-1)| + |b_{k'}(t-1)| \leq O\left(\frac{B_k(t-1)}{K}\right) + O\left(\frac{b_k(t-1)}{K}\right) \leq O\left(\frac{B_k(t)}{K}\right).$$

Hence, we finish the proof of Induction Hypothesis K.1.

■

K.4.1. Technical Lemmas

Similar to stage I, we next introduce several useful technical lemmas and prove them by induction.

Lemma K.4. *Suppose Induction Hypothesis K.1 holds at iteration $\tau_1 + 1 \leq t \leq \tau_1 + T_{1,k}^{(2)}$. Conditioned on the events $\mathbf{x}_{\text{q}} = \mathbf{v}_k$ and $P_{\text{in}} \in \mathcal{E}_{\text{P}}^*$, the following holds for iteration t.*

1. *For the attention scores related to feature k,*
   - *(a) $\mathcal{S}_k^{(2)}(t) = \Omega(1/K)$,*
   - *(b) $1 - \mathcal{S}_k^{(2)}(t) = \Omega(1)$.*

2. *For the attention scores related to feature $k' \neq k$, $\mathcal{S}_{k'}^{(2)}(t) = \Theta\big((1 - \mathcal{S}_k^{(2)}(t))/K\big) = \Theta(1/K)$.*

Proof We first prove Claim (1)-(a).

$$\begin{aligned}
\mathcal{S}_k^{(2)}(t) &= \frac{\sum_{i=2n,n\in[N_{\text{in}}]} \exp\left(\mathbf{u}_i(t)\right)\mathbb{1}\{\mathbf{x}_i = \mathbf{v}_k\}}{\sum_{m\in[K]}\sum_{i=2n,n\in[N_{\text{in}}]} \exp\left(\mathbf{u}_i(t)\right)\mathbb{1}\{\mathbf{x}_i = \mathbf{v}_m\} + \sum_{i=2n-1,n\in[N_{\text{in}}+1]} \exp\left(\mathbf{u}_i(t)\right)} \\
&= \frac{|\mathcal{V}_k| \exp\left(B_k(t)\right)}{|\mathcal{V}_k| \exp\left(B_k(t)\right) + \sum_{m\neq k} |\mathcal{V}_m| \exp\left(B_m(t)\right) + \sum_{i=2n-1,n\in[N_{\text{in}}+1]} \exp\left(\mathbf{u}_i(t)\right)} \\
&\overset{(i)}{\geq} \frac{1}{1 + \sum_{m\neq k}(|\mathcal{V}_m|/|\mathcal{V}_k|) \exp\left(B_m(t) - B_k(t)\right) + (N_{\text{in}}/|\mathcal{V}_k|) \exp\left(-(1-\Theta(\epsilon_1))B_k(t)\right)}
\end{aligned} \tag{54}$$

$$\overset{(ii)}{\geq} \frac{1}{1+\sum_{m\neq k}(|\mathcal{V}_m|/|\mathcal{V}_k|)\exp\left(B_m(t)-B_k(t)\right)+N_{\text{in}}/|\mathcal{V}_k|},$$

where (i) follows from Theorem K.2, which gives $\mathbf{u}_i(t) \leq \epsilon_1\Theta(\max_{m\in[K]} B_m(t))$ for odd i, together with the fact that $\max_{m\neq k} B_m(t) \leq B_k(t)$ in Induction Hypothesis K.1, and (ii) follows from that $-B_k(t)(1-\Theta(\epsilon_1)) \leq 0$. Furthermore, by Induction Hypothesis K.1, for any $k' \neq k$

$$\exp\left(-\log K - O((\log K)/K)\right) \leq \exp(B_{k'}^{(t)} - B_k^{(t)}) \leq \exp\left(O((\log K)/K)\right).$$

By direct calculation and Theorem K.3, we have

$$\mathcal{S}_k^{(2)}(t) \geq \frac{1}{e^{O((\log K)/K)}(N_{\text{in}}/|\mathcal{V}_k|-1)+1+N_{\text{in}}/|\mathcal{V}_k|} \geq \frac{1}{e^{O((\log K)/K)}(K/L_k-1)+1+K/L_k} = \Omega\Big(\frac{1}{K}\Big).$$

Hence, we prove Claim (1)-(a) and next prove Claim (1)-(b). Similar to Claim (1)-(a), we have

$$\mathcal{S}_k^{(2)}(t) \leq \frac{1}{1+\sum_{m\neq k}(|\mathcal{V}_m|/|\mathcal{V}_k|)\exp\left(B_m(t)-B_k(t)\right)} \leq \frac{1}{e^{-1}(1/U_k-1/K)+1}.$$

Together with $U_k = \Theta(1)$, we have

$$1-\mathcal{S}_k^{(t)} \geq \frac{(1/U_k-1/K)}{(1/U_k-1/K)+e} \geq \Omega(1).$$

Hence, we prove Claim (1)-(b). Finally, we prove Claim (2). For any $k' \neq k$, we have

$$\mathcal{S}_{k'}^{(2)}(t) = \frac{|\mathcal{V}_{k'}|\exp\left(B_{k'}(t)\right)}{|\mathcal{V}_{k'}|\exp\left(B_{k'}(t)\right)+\sum_{m\neq k'}|\mathcal{V}_m|\exp\left(B_m(t)\right)+\sum_{i=2n-1,n\in[N_{\text{in}}+1]}\exp\left(\mathbf{u}_i(t)\right)}.$$

Because $\exp\left(\mathbf{u}_i(t)\right) \geq 0$ for all i, we have

$$\frac{\mathcal{S}_{k'}^{(2)}(t)}{1-\mathcal{S}_k^{(2)}(t)} \leq \frac{1}{1+\sum_{m\neq k',k}(|\mathcal{V}_m|/|\mathcal{V}_{k'}|)\exp\left(B_m(t)-B_{k'}(t)\right)} \leq O\left(\frac{1}{K}\right),$$

where the last inequality follows from Theorem K.3 and $U_m/L_{k'} = \Theta(1)$. For another direction, we have

$$\begin{aligned}
\frac{\mathcal{S}_{k'}^{(2)}(t)}{1-\mathcal{S}_k^{(2)}(t)} &\overset{(i)}{\geq} \frac{1}{1+\sum_{m\neq k',k}(|\mathcal{V}_m|/|\mathcal{V}_{k'}|)\exp\left(B_m(t)-B_{k'}(t)\right)+(N_{\text{in}}/|\mathcal{V}_{k'}|)\exp\left(\epsilon_1\Theta\left(B_k(t)\right)-B_{k'}(t)\right)} \\
&\overset{(ii)}{\geq} \frac{1}{1+\sum_{m\neq k',k}(|\mathcal{V}_m|/|\mathcal{V}_{k'}|)e^{O((\log K)/K)}+(N_{\text{in}}/|\mathcal{V}_{k'}|)e^{\epsilon_1\log(K)+O((\log K)/K)}} \\
&\overset{(iii)}{\geq} \frac{1}{1+(\sum_{m\neq k',k}U_m e/L_{k'})+Ke/L_{k'}} \geq \Omega\left(\frac{1}{K}\right),
\end{aligned} \tag{55}$$

where (i) follows from Theorem K.2, (ii) follows from Induction Hypothesis K.1, and (iii) follows from Theorem K.3. Hence, we finish the proof of Theorem K.4. ■

In the following lemma, we control the order of $b_{k'}(t), k' \in [K]$.

Lemma K.5 (Order of updates $b(t)$)**.** *Suppose Induction Hypothesis K.1 and Theorem K.4 hold at iteration $\tau_1+1 \leq t \leq \tau_1+T_{1,k}^{(2)}$, under the same assumption of Theorem K.4, we have*

1. *For k, $b_k(t) \geq 0$ and $b_k(t) = \Omega\left(\eta_2/K^2\right)$.*
2. *For $k' \neq k$, $|b_{k'}(t)| \leq O(b_k(t)/K)$.*

Proof We first prove Claim (1). By Theorem K.2, we only need to consider the dominated term of $b_k(t)$ as follows:

$$\begin{aligned}
&\eta_2\mathbb{E}\big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\big(\mathcal{S}_k^{(2)}\big((\mathcal{S}_k^{(2)}-1)^2+\sum_{m\neq k}(\mathcal{S}_m^{(2)})^2\big)\big)\big]\\
&\geq \eta_2 p_k\mathbb{P}(P_{\mathrm{in}}\in\mathcal{E}_{\mathrm{P}}^*)\times\mathbb{E}\big[\mathcal{S}_k^{(2)}\left(\mathcal{S}_k^{(2)}-1\right)^2|\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\cap\{P_{\mathrm{in}}\in\mathcal{E}_{\mathrm{P}}^*\}\big]+3\eta_2 p_k\mathbb{P}(P_{\mathrm{in}}\notin\mathcal{E}_{\mathrm{P}}^*)\\
&\overset{(i)}{\geq}\Omega\left(\frac{\eta_2}{K^2}\right),
\end{aligned}$$

where (i) follows from Theorems K.3 and K.4. Hence, we prove Claim (1). We next prove Claim (2). By Theorem K.2, for any $k'\neq k$, we denote the dominated term of $b_{k'}(t)$ as follows

$$(I)=\eta_2\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\mathcal{S}_{k'}^{(2)}\Big(\sum_{m\in[K]}(\mathcal{S}_m^{(2)})^2-(\mathcal{S}_{k'}^{(2)}+\mathcal{S}_k^{(2)})\Big)\Big].$$

On one hand, we have

$$\begin{aligned}
(I)&\leq\eta_2 p_k\mathbb{P}(P_{\mathrm{in}}\in\mathcal{E}_{\mathrm{P}}^*)\mathbb{E}\big[\mathcal{S}_{k'}^{(2)}\big(\max_{m\neq k,k'}\mathcal{S}_m^{(2)}\big)|\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\cap\{P_{\mathrm{in}}\in\mathcal{E}_{\mathrm{P}}^*\}\big]+\eta_2 p_k\mathbb{P}(P_{\mathrm{in}}\notin\mathcal{E}_{\mathrm{P}}^*)\\
&\overset{(i)}{\leq}O\left(\frac{\eta_2}{K^3}\right),
\end{aligned}\tag{56}$$

where (i) follows from Theorem K.3, and Theorem K.4 and $N\gg K^3$. On the other hand, we have

$$\begin{aligned}
-(I)&\overset{(i)}{\leq}\eta_2\big\{2p_k\cdot\mathbb{P}(P_{\mathrm{in}}\notin\mathcal{E}_{\mathrm{P}}^*)+p_k\cdot\mathbb{P}(P_{\mathrm{in}}\in\mathcal{E}_{\mathrm{P}}^*)\\
&\quad\cdot\mathbb{E}\Big[\Theta\Big(\frac{1-\mathcal{S}_k^{(2)}(t)}{K}\Big)\Big(\Theta\Big(\frac{1-\mathcal{S}_k^{(2)}(t)}{K}\Big)+\mathcal{S}_k^{(2)}(1-\mathcal{S}_k^{(2)}(t))\Big)\Big|\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\cap\{P_{\mathrm{in}}\in\mathcal{E}_{\mathrm{P}}^*\}\Big]\big\}\\
&\overset{(ii)}{\leq}O\Big(\frac{b_k(t)}{K}\Big),
\end{aligned}\tag{57}$$

where (i) follows from Theorem K.4, (ii) follows from Theorem K.3, Claim (1), which states $b_k(t)=\Omega\left(\eta_2/K^2\right)$ and the fact that $\eta_2/K^2\geq\exp\left(-c^2N/(25K^2)\right)$. As a result, combining (56) and (57), we prove Claim (2) and this lemma. ■

K.4.2. End of Phase I

Lemma K.6. *Conditioned on the events that $\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k$ and $P_{\mathrm{in}}\in\mathcal{E}_{\mathrm{P}}^*$, with $T_{1,k}^{(2)}$ at most $O\left((K^2\log K)/\eta_2\right)$, and at iteration $t=\tau_1+T_{1,k}^{(2)}+1$, we have*

1. $B_k(\tau_1+T_{1,k}^{(2)}+1)\geq\log(K)$,
2. $\mathcal{S}_k^{(2)}(\tau_1+T_{1,k}^{(2)}+1)=\Omega(1)$.

Proof Claim (1) holds from the definition of Phase I. We only need to show the upper bound of $T_{1,k}^{(2)}$. By direct calculation, we have

$$\begin{aligned}
\log(K)&\geq B_k(\tau_1+T_{1,k}^{(2)})\\
&\geq B_k(\tau_1+1)+\sum_{t=\tau_1+1}^{\tau_1+T_{1,k}^{(2)}-1}B_k(t+1)-B_k(t)\\
&\geq 1-\Omega(\epsilon_1+N^{1-\alpha})+(T_{1,k}^{(2)}-2)\Omega(\eta_2/K^2),
\end{aligned}$$

where the last inequality follows from Theorem K.5 and Induction Hypothesis K.1. As a result, $T_{1,k}^{(2)} \leq O\left((K^2 \log K)/\eta_2\right)$. We next prove Claim (2). Similar to (54) and by $B_k(\tau_1 + T_{1,k}^{(2)} + 1) \geq \log(K)$, we have

$$\mathcal{S}_k^{(2)}(\tau_1 + T_{1,k}^{(2)} + 1) \gtrsim \frac{1}{1 + \sum_{m \neq k} |\mathcal{V}_m|(|\mathcal{V}_k|K) + N_{\text{in}}(|\mathcal{V}_k|K)} \overset{(i)}{\simeq} \Theta(1),$$

where (i) follows from the fact $|\mathcal{V}_k| = \Theta(N/K)$ in Theorem K.3. Hence, we prove Claim (2). ■

K.5. Stage II: Phase II

We define the **Phase II** as all iterations $\tau_1 + T_{1,k}^{(2)} + 1 \leq t \leq \tau_1 + T_{2,k}^{(2)}$, where

$$T_{2,k}^{(2)} \triangleq \max\left\{t : B_k(\tau_1 + t) - \max_{m \neq k} B_m(\tau_1 + t) \leq \log\left(K\epsilon_2^{-\frac{1}{2}}\right)\right\}.$$

We then state the following induction hypothesis, which holds throughout Phase II. This hypothesis is proved by induction with the technical lemmas in Section K.5.1.

Induction Hypothesis K.2. *Suppose Theorems K.7 and K.8 and Induction Hypothesis K.1 hold at iteration $t-1$, where $\tau_1 + T_{1,k}^{(2)} + 1 \leq t \leq \tau_1 + T_{2,k}^{(2)}$. Given any $k \in [K]$, conditioned on the event that $\mathbf{x}_{\text{q}} = \mathbf{v}_k$ and $P_{\text{in}} \in \mathcal{E}_{\text{P}}^*$, the following holds for iteration t:*

1. *$B_k(t)$ is monotonically increasing and $B_k(t) \in [\log(K), O(\log(K/\epsilon_2))]$.*
2. *For any $k' \neq k$, $B_{k'}^{(t)}$ is monotonically decreasing and $|B_{k'}(t)| = O(B_k(t)/K)$.*

Proof We first prove Claim (2). By Theorem K.8, we have for any $k' \neq k$, $|b_{k'}(t-1)| \leq O\left(b_k(t-1)/K\right)$ at time $t-1$, then

$$|B_{k'}(t)| \leq |B_{k'}(t-1)| + |b_{k'}(t-1)| \leq O\left(B_k(t)/K\right).$$

Hence, we prove Claim (2). We next prove Claim (1). For any $\tau_1 + T_{1,k}^{(2)} + 1 \leq t \leq \tau_1 + T_{2,k}^{(2)}$, we first need to show $b_k(t) \geq 0$, which is obvious from Theorem K.8. Then, by the monotonicity and Theorem K.6, we have $\log(K) \leq B_k(t)$. In addition, we have

$$\left(1 - O\left(1/K\right)\right) B_k(t) \leq B_k(t) - \max_{m \neq k} B_m(t) \leq \log(K\epsilon_2^{-\frac{1}{2}}) \leq O\left(\log(K/\epsilon_2)\right),$$

where the first inequality follows from Claim (2), and the second inequality follows from the definition of Phase II. Hence, we prove Claim (1). ■

K.5.1. Technical Lemmas

Similar to phase I, we next introduce several useful technical lemmas and prove them by induction.

Lemma K.7. *Suppose Induction Hypothesis K.2 holds at iteration $\tau_1 + T_{1,k}^{(2)} + 1 \leq t \leq \tau_1 + T_{2,k}^{(2)}$, conditioned on the events that $\mathbf{x}_{\text{q}} = \mathbf{v}_k$ and $P_{\text{in}} \in \mathcal{E}_{\text{P}}^*$, the following holds for iteration t:*

1. *For the attention scores related to feature k, we have that*
   - *(a) $\mathcal{S}_k^{(2)}(t) = \Omega(1)$,*
   - *(b) $(1 - \mathcal{S}_k^{(2)}(t))^2 \geq \Omega(\epsilon_2)$.*
2. *For the attention scores related to feature $k' \neq k$, $\mathcal{S}_{k'}^{(2)}(t) = \Theta\big((1 - \mathcal{S}_k^{(2)}(t))/K\big)$.*

Proof We first prove Claim (1)-(a). Similar to Theorem K.4, we have

$$\begin{aligned}
\mathcal{S}_k^{(2)}(t) &\geq \frac{1}{\sum_{m\in[K]}(|\mathcal{V}_m|/|\mathcal{V}_k|)\exp\left(B_m(t)-B_k(t)\right)+(N_{\text{in}}/|\mathcal{V}_k|)\exp\left(\epsilon_1\Theta\left(\max_{m\in[K]}B_m(t)\right)-B_k(t)\right)} \\
&\overset{(i)}{\geq} \frac{1}{1+(N_{\text{in}}/|\mathcal{V}_k|-1)\exp\left(\max_{m\neq k}B_m(t)-B_k(t)\right)+(N_{\text{in}}/|\mathcal{V}_k|)\exp\left(-B_k(t)(1-\Theta(\epsilon_1))\right)} \\
&\overset{(ii)}{\gtrsim} \frac{1}{1+(K/L_k)\cdot(\epsilon_2^{1/2}/K)+K^{\Theta(\epsilon_1)}/L_k}=\Omega(1),
\end{aligned}$$

where (i) follows from the fact $\max_{m\neq k}B_m(t)\leq B_k(t)$ in Induction Hypothesis K.2, and (ii) follows from the bound of $B_k(t)$ in Induction Hypothesis K.2 and Theorem K.3. Hence, we prove Claim (1)-(a). Similarly, for Claim (1)-(b), we have

$$\begin{aligned}
1-\mathcal{S}_k^{(2)}(t) &= \frac{\sum_{m\neq k}|\mathcal{V}_m|\exp\left(B_m(t)\right)+\sum_{i=2n-1,n\in[N_{\text{in}}+1]}\exp\left(\mathbf{u}_i(t)\right)}{|\mathcal{V}_k|\exp\left(B_k(t)\right)+\sum_{m\neq k}|\mathcal{V}_m|\exp\left(B_m(t)\right)+\sum_{i=2n-1,n\in[N_{\text{in}}+1]}\exp\left(\mathbf{u}_i(t)\right)} \\
&\overset{(i)}{\geq} \frac{\sum_{m\neq k}(|\mathcal{V}_m|/|\mathcal{V}_k|)\exp\left(B_m(t)-B_k(t)\right)}{1+\sum_{m\neq k}(|\mathcal{V}_m|/|\mathcal{V}_k|)\exp\left(B_m(t)-B_k(t)\right)+(N/|\mathcal{V}_k|)\exp\left(-B_k(t)(1-\Theta(\epsilon_1))\right)} \\
&\overset{(ii)}{\geq} \frac{\exp\left(\max_{m\neq k}B_m(t)-B_k(t)-\Delta B(t)\right)(K/U_k-1)}{1+\exp\left(\max_{m\neq k}B_m(t)-B_k(t)-\Delta B(t)\right)(K/U_k-1)+(K/U_k)\exp\left(-B_k(t)(1-\Theta(\epsilon_1))\right)} \\
&\overset{(iii)}{\geq} \frac{K^{-1}\epsilon_2^{1/2}\exp\left(-O(K^{-1}\log(K/\epsilon_2))\right)\cdot(K/U_k-1)}{1+(K^{-1}\epsilon_2^{1/2}\exp -O(K^{-1}\log(K/\epsilon_2)))(K/U_k-1)+K^{\Theta(\epsilon_1)}/L_k}\geq\Omega\left(\epsilon_2^{\frac{1}{2}}\right),
\end{aligned}$$

where (i) follows from Theorem K.2 and Induction Hypothesis K.2, (ii) follows by defining $\Delta B(t)=\max_{m\neq k}B_m(t)-\min_{m\neq k}B_m(t)$, and (iii) follows from Induction Hypothesis K.2 and the fact that $\Delta B(t)=O\left(B_k(t)/K\right)=O\left(K^{-1}\log(K/\epsilon_2)\right)$. Hence, we prove Claim (1)-(b).

We next prove Claim (2). For any $k'\neq k$, we have

$$\mathcal{S}_{k'}^{(2)}(t)=\frac{|\mathcal{V}_{k'}|\exp\left(B_{k'}(t)\right)}{|\mathcal{V}_{k'}|\exp\left(B_{k'}(t)\right)+\sum_{m\neq k'}|\mathcal{V}_m|\exp\left(B_m(t)\right)+\sum_{i=2n-1,n\in[N_{\text{in}}+1]}\exp\left(\mathbf{u}_i(t)\right)}.$$

Then, similar to proof of Claim (2) in Theorem K.4, we have

$$\begin{aligned}
\frac{\mathcal{S}_{k'}^{(2)}(t)}{1-\mathcal{S}_k^{(2)}(t)} &= \frac{|\mathcal{V}_{k'}|\exp\left(B_{k'}(t)\right)}{|\mathcal{V}_{k'}|\exp\left(B_{k'}(t)\right)+\sum_{m\neq k',k}|\mathcal{V}_m|\exp\left(B_m(t)\right)+\sum_{i=2n-1,n\in[N_{\text{in}}+1]}\exp\left(\mathbf{u}_i(t)\right)} \\
&\leq \frac{1}{1+\sum_{m\neq k',k}(|\mathcal{V}_m|/|\mathcal{V}_{k'}|)\exp\left(B_m(t)-B_{k'}(t)\right)} \\
&\leq \frac{1}{1+\sum_{m\neq k',k}(U_m/L_{k'})\exp\left(-O(K^{-1}\log(K/\epsilon_2))\right)}=O\left(\frac{1}{K}\right),
\end{aligned}$$

where the last inequality follows from the fact $U_m/L_{k'}=\Theta(1)$ in Theorem K.3 and Induction Hypothesis K.2.

Similar to (55), we have

$$\begin{aligned}
\frac{\mathcal{S}_{k'}^{(2)}(t)}{1-\mathcal{S}_k^{(2)}(t)} &\geq \frac{1}{1+\sum_{m\neq k',k}(|\mathcal{V}_m|/|\mathcal{V}_{k'}|)\exp\left(B_m(t)-B_{k'}(t)\right)+(N_{\text{in}}/|\mathcal{V}_{k'}|)\exp\left(\epsilon_1\Theta\left(B_k(t)\right)-B_{k'}(t)\right)} \\
&\overset{(i)}{\geq} \frac{1}{1+\sum_{m\neq k',k}U_m e/L_{k'}+Ke/L_{k'}}\geq\Omega\left(\frac{1}{K}\right),
\end{aligned}$$

where (i) follows from Theorem K.2 and Induction Hypothesis K.2. Hence, we prove Theorem K.7. ■

In the following lemma, we control the order of $b_{k'}(t), k'\in[K]$.

Lemma K.8 (Order of updates $b(t)$)**.** *Suppose Induction Hypothesis K.2 and Theorem K.7 hold at iteration $\tau_1 + T_{1,k}^{(2)} + 1 \le t \le \tau_1 + T_{2,k}^{(2)}$, under the same assumption of Theorem K.7, we have that*

1. *For k, $b_k(t) \ge 0$ and $b_k(t) = \Omega(\eta_2\epsilon_2/K)$,*
2. *For $k' \ne k$, $b_{k'}(t) \le 0$, and $|b_{k'}(t)| \le O(b_k(t)/K)$.*

Proof We first prove Claim (1), by Theorem K.2, and $\epsilon_2 \ge \epsilon_1$, we only need to consider the dominated term of $b_k(t)$ as follows:

$$\begin{aligned}
&\eta_2\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k\}\Big(\mathcal{S}_k^{(2)}\Big((\mathcal{S}_k^{(2)} - 1)^2 + \sum_{m\ne k}(\mathcal{S}_m^{(2)})^2\Big)\Big)\Big] \\
&\ge \eta_2 p_k \mathbb{P}(P_{\mathrm{in}} \in \mathcal{E}_{\mathrm{P}}^*) \times \mathbb{E}\big[\mathcal{S}_k^{(2)}\big(\mathcal{S}_k^{(2)} - 1\big)^2 | \{\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k\} \cap \{P_{\mathrm{in}} \in \mathcal{E}_{\mathrm{P}}^*\}\big] + 3\eta_2 p_k \mathbb{P}(P_{\mathrm{in}} \notin \mathcal{E}_{\mathrm{P}}^*) \\
&\overset{(i)}{\ge} \Omega\left(\frac{\eta_2\epsilon_2}{K}\right),
\end{aligned}$$

where (i) follows from Theorems K.3 and K.7. Hence, we prove Claim (1). We next prove Claim (2). By Theorem K.2, for any $k' \ne k$, denote the dominated term of $b_{k'}(t)$ as follows

$$(I) = \eta_2\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k\}\Big(\mathcal{S}_{k'}^{(2)} \sum_{m\in[K]}(\mathcal{S}_m^{(2)})^2 - (\mathcal{S}_{k'}^{(2)} + \mathcal{S}_k^{(2)})\Big)\Big].$$

On one hand, we have

$$\begin{aligned}
(I) &\le \eta_2\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k\} \cap \{P_{\mathrm{in}} \in \mathcal{E}_{\mathrm{P}}^*\}\mathcal{S}_{k'}^{(2)}\Big(\sum_{m\notin k,k'}(\mathcal{S}_m^{(2)})^2 - \mathcal{S}_k^{(2)}(1 - \mathcal{S}_k^{(2)})\Big)\Big] + \eta_2 p_k \mathbb{P}(P_{\mathrm{in}} \notin \mathcal{E}_{\mathrm{P}}^*) \\
&\overset{(i)}{\lesssim} \eta_2\Big(-p_k\mathbb{E}\Big[\frac{(1-\mathcal{S}_k^{(2)})^2}{K}\Big|\{\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k\} \cap \{P_{\mathrm{in}} \in \mathcal{E}_{\mathrm{P}}^*\}\Big] + 3p_k\exp\Big(-\frac{c^2N}{25K^2}\Big)\Big) \\
&\overset{(ii)}{\le} -\Omega\left(\frac{\eta_2\epsilon_2}{K}\right) \le 0,
\end{aligned} \tag{58}$$

where (i) follows from Theorem K.3 and the fact that $\mathcal{S}_{k'}^{(2)}(t) = \Theta\big((1 - \mathcal{S}_k^{(2)}(t))/K\big)$ in Theorem K.7, and (ii) follows from the fact $(1 - \mathcal{S}_k^{(2)}(t))^2 \ge \Omega(\epsilon_2)$ in Theorem K.4 and $N \gg K^3$.

On the other hand, we have

$$\begin{aligned}
-(I) &\overset{(i)}{\le} \eta_2\big\{2p_k \cdot \mathbb{P}(P_{\mathrm{in}} \notin \mathcal{E}_{\mathrm{P}}^*) + p_k \cdot \mathbb{P}(P_{\mathrm{in}} \in \mathcal{E}_{\mathrm{P}}^*) \times \\
&\quad \mathbb{E}\Big[\Theta\Big(\frac{1-\mathcal{S}_k^{(2)}(t)}{K}\Big)\Big(\Theta\Big(\frac{1-\mathcal{S}_k^{(2)}(t)}{K}\Big) + \mathcal{S}_k^{(2)}(1-\mathcal{S}_k^{(2)}(t))\Big)\Big|\{\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k\} \cap \{P_{\mathrm{in}} \in \mathcal{E}_{\mathrm{P}}^*\}\Big]\Big\} \\
&\overset{(ii)}{\le} \eta_2\Big\{p_k\mathbb{E}\Big[O\Big(\frac{\mathcal{S}_k^{(2)}(t)(1-\mathcal{S}_k^{(2)}(t))^2}{K}\Big)|\{\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k\} \cap \{P_{\mathrm{in}} \in \mathcal{E}_{\mathrm{P}}^*\}\Big] + 6p_k\exp\Big(-\frac{c^2N}{25K^2}\Big)\Big\} \\
&\overset{(iii)}{\le} O\Big(\frac{b_k(t)}{K}\Big),
\end{aligned} \tag{59}$$

where (i) follows from Theorem K.7, (ii) follows from Theorem K.3, and (iii) follows from Claim (1). Finally, combining (58) and (59), we prove Claim (2). ■

K.5.2. End of Phase II

Lemma K.9. *Conditioned on the events that $\mathbf{x}_{\mathrm{q}} = \mathbf{v}_k$ and $P_{\mathrm{in}} \in \mathcal{E}_{\mathrm{P}}^*$, with $T_{2,k}^{(2)} - T_{1,k}^{(2)}$ at most $O(\eta_2^{-1}\epsilon_2^{-1}K\log(K\epsilon_2^{-\frac{1}{2}}))$, and at iteration $t = \tau_1 + T_{2,k}^{(2)} + 1$, we have that*

1. $B_k(\tau_1 + T_{2,k}^{(2)} + 1) - \max_{m\neq k} B_m(\tau_1 + T_{2,k}^{(2)} + 1) \geq \log\big(K\epsilon_2^{-\frac{1}{2}}\big)$,

2. $1 - \mathcal{S}_k^{(2)}(\tau_1 + T_{2,k}^{(2)} + 1) \leq \epsilon_2^{\frac{1}{2}}$.

Proof Claim (1) holds from the definition of Phase II. We only need to show the upper bound of $T_{2,k}^{(2)}$. By Induction Hypothesis K.2 and Theorem K.8, for any $\tau_1 + T_{1,k}^{(2)} + 1 \leq t \leq \tau_1 + T_{2,k}^{(2)}$, we have

$$\big(B_k(t+1) - \max_{m\neq k} B_m(t+1)\big) - \big(B_k(t) - \max_{m\neq k} B_m(t)\big) \geq (1 - O(1/K))b_k(t) = \Omega(\eta_2\epsilon_2/K).$$

By direct calculation, we have

$$\begin{aligned}
&\big(B_k(\tau_1 + T_{2,k}^{(2)}) - \max_{m\neq k} B_m(\tau_1 + T_{2,k}^{(2)})\big) - \big(B_k(\tau_1 + T_{1,k}^{(2)} + 1) - \max_{m\neq k} B_m(\tau_1 + T_{1,k}^{(2)} + 1)\big) \\
&= \sum_{t=\tau_1+T_{1,k}^{(2)}+1}^{\tau_1+T_{2,k}^{(2)}-1} \big(B_k(t+1) - \max_{m\neq k} B_m(t+1)\big) - \big(B_k(t) - \max_{m\neq k} B_m(t)\big) \\
&\geq \big(T_{2,k}^{(2)} - T_{1,k}^{(2)} - 1\big)\Omega\big(\eta_2\epsilon_2/K\big).
\end{aligned}$$

Dividing both side by $\Omega\big(\eta_2\epsilon_2/K\big)$, we have

$$T_{2,k}^{(2)} - T_{1,k}^{(2)} \leq O\bigg(\frac{K\log(K\epsilon_2^{-\frac{1}{2}})}{\eta_2\epsilon_2}\bigg).$$

Hence, we prove Claim (1). Then we prove Claim (2). For abbreviation, we omit the iteration $\tau_1 + T_{1,k}^{(2)} + 1$ in the following. By definition, we have

$$\begin{aligned}
1 - \mathcal{S}_k^{(2)} &= \frac{\sum_{m\neq k} |\mathcal{V}_m| \exp(B_m) + \sum_{i=2n-1, n\in[N_{\text{in}}+1]} \exp(\mathbf{u}_i)}{\sum_{m\in[K]} |\mathcal{V}_m| \exp(B_m) + \sum_{i=2n-1, n\in[N_{\text{in}}+1]} \exp(\mathbf{u}_i)} \\
&\overset{(i)}{\leq} \frac{\sum_{m\neq k}(|\mathcal{V}_m|/|\mathcal{V}_k|)\exp(B_m - B_k) + (N_{\text{in}}/|\mathcal{V}_k|)\exp(-B_k(1-\Theta(\epsilon_1)))}{1 + \sum_{m\neq k}(|\mathcal{V}_m|/|\mathcal{V}_k|)\exp(B_m - B_k)} \\
&\overset{(ii)}{\leq} \frac{\exp(\max_{m\neq k} B_m - B_k)(K/L_k - 1) + K/L_k \exp(-B_k(1-\Theta(\epsilon_1)))}{1 + \exp(\max_{m\neq k} B_m(T_{1,k}^{(2)} + 1) - B_k(T_{1,k}^{(2)} + 1))(K/L_k - 1)} \\
&\overset{(iii)}{\leq} \frac{K^{-1}\epsilon_2^{1/2}\cdot(K/L_k - 1) + (K/L_k)\cdot K^{-1}\epsilon_2^{1/2}}{1 + K^{-1}\epsilon_2^{1/2}\cdot(K/L_k - 1)} \leq O\big(\epsilon_2^{\frac{1}{2}}\big),
\end{aligned}$$

where (i) follows from the fact $\mathbf{u}_i(t) \leq \epsilon_1\Theta(\max_{m\in[K]} B_m(t))$ for all t in Theorem K.2, (ii) follows from the bound of $|\mathcal{V}_k|$ in Theorem K.3, and (iii) follows from Claim (1) and the definition of phase II. Hence, we finish the proof of Theorem K.9. ■

L. Proof of Theorem D.3

Proof The Claims about Stage I and Stage II follow directly from Theorems J.22 and K.9, respectively. Then we only need to show the loss convergence. The loss can be rewritten as

$$\begin{aligned}
L(\theta(\tau_1 + \tau_2 + 1)) &= \frac{1}{2}\mathbb{E}\Big[\Big(\sum_{i=1}^{N} \mathcal{S}_i^{(2)}(\tau_1 + \tau_2 + 1))E^y - y_{N+1}\Big)^2\Big] \\
&= \frac{1}{2}\sum_{k=1}^{K} \mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\text{q}} = \mathbf{v}_k\}\Big(\sum_{m=1}^{K} \mathcal{S}_m^{(2)}(\tau_1 + \tau_2 + 1)\mathbf{w}^\top\mathbf{v}_m - \mathbf{w}^\top\mathbf{v}_k\Big)^2\Big]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(i)}{=} \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\Big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\Big(\sum_{m\neq k}(\mathcal{S}_m^{(2)}(\tau_1+\tau_2+1))^2+(1-\mathcal{S}_k^{(2)}(\tau_1+\tau_2+1))^2\Big)\Big]\\
&\leq \sum_{k=1}^{K}\mathbb{E}\big[\mathbb{1}\{\mathbf{x}_{\mathrm{q}}=\mathbf{v}_k\}\big(1-\mathcal{S}_k^{(2)}(\tau_1+\tau_2+1)\big)^2\big]\\
&\overset{(ii)}{\leq} \epsilon_2,
\end{aligned}
$$

where (i) follows from taking the conditional expectation of $\mathbf{w}$ conditioned on P and direct calculation, and (ii) follows from Theorem K.9. Hence, we finish the proof. ■