

# Performance–Efficiency Tradeoffs in Transformers: An Approximation Theory Perspective

Ruoxi Yu<sup>1,\*</sup>  
yuruoxi@stu.pku.edu.cn

Haotian Jiang<sup>2,\*</sup>  
haotian@nus.edu.sg

Jingpu Cheng<sup>3</sup>  
chengjingpu@u.nus.edu

Penghao Yu<sup>3</sup>  
e1353366@u.nus.edu

Qianxiao Li<sup>2,3</sup>  
qianxiao@nus.edu.sg

Zhong Li<sup>4</sup>  
lzhong@microsoft.com

<sup>1</sup>Center for Data Science, Peking University

<sup>2</sup>Institute for Functional Intelligent Materials, National University of Singapore

<sup>3</sup>Department of Mathematics, National University of Singapore

<sup>4</sup>Microsoft Research Asia

\*Equal contribution.

## Abstract

Transformers have achieved remarkable empirical successes across a wide range of applications, yet how to balance model performance and computation efficiency in a principled manner remains underexplored. In this work, we theoretically study the parameter efficiency in multi-head attention Transformers by quantitatively analyzing the relationship between Transformers’ hyper-parameters (i.e. the number of attention heads and head dimensions) and performance. (i) With a fixed total parameters budget, we establish an approximation error estimate for information extraction tasks, revealing an intrinsic trade-off between the number of heads and head dimensions to boost model performance. (ii) With maintained performance, we further mathematically prove a saturation pattern in softmax-based attention, showing that increasing head dimensions leads to diminishing returns in performance (especially for long sequences), implying potential parameter reduction. All these theoretical results are consistently verified through extensive experiments from synthetic to real-world simulations, providing principled guidance for trade-offs between performance and parameter efficiency of Transformers.

## 1 Introduction

Transformers have achieved remarkable successes across numerous applications such as natural language processing ([22, 10, 30]), computer vision ([23, 53, 88]), audio/speech ([29, 96, 54, 44]) and beyond, serving as the backbone of modern large language models (LLMs) and multi-modal systems due to their ability to capture long-range dependencies and scale efficiently ([78, 20, 16, 85, 6]). Their impact spans healthcare ([59, 12]), robotics ([95, 52]), agent ([75, 92]), chemistry ([9, 37]), finance ([98, 89]), and so on.

Despite their success, large language models face a fundamental efficiency challenge, as naive scaling demands rapidly growing data and computation ([76, 5]). A growing body of work has investigated improving Transformer efficiency through parameter allocation and architectural design. In particular, multi-head attention (MHA) has been shown to exhibit substantial redundancy: Many attention heads can be pruned without significant performance drops, and only a small subset consistently captures meaningful patterns across tasks and languages ([62, 79, 57]). Building on these observations, subsequent studies propose automated and stable pruning criteria for model compression while preserving model performance ([15, 45]). Meanwhile, computation efficiency has been further improved through architecture and system-level modifications, including sparse or hierarchical attention mechanisms ([53, 17, 32, 6, 94]), shared key–value projections ([1]), and optimized attention kernels ([20, 33]). Despite that all these prior advances provide strong empirical

evidence of the redundancy in Transformers, and efficiency gains can be achieved via architecture design or system-level optimization, they usually identify removable components in trained models, leading to limited guidance on how architectural resources should be allocated a priori, or why particular model configurations are effective in theory.

To tackle these problems, in this work, we qualitatively study the topic of parameter efficiency (in Transformers), which is in fact two sides of the same coin: (i) With fixed resources/budgets of total parameters, how to configure the model to enhance performance? (ii) To maintain performance, can we and how to reduce resources/budgets of total parameters? Specifically, this parameter-performance trade-off in multi-head attention Transformers are regarding the allocation of attention heads and head dimensions. Particularly in this work, we consider the following two problems: (i) Given fixed model dimensions, how to allocate the number of heads and head dimensions for better model performance? (ii) Without degrading model performance, can we reduce head dimensions with a fixed number of heads?

To this end, for problem (i), we derive an approximation error estimate for information extraction tasks, which implies an intrinsic trade-off between the number of heads and head dimensions aiming at enhancing model performance with the fixed total parameter budget (i.e. fixed model dimension) in Section 3. For problem (ii), we further theoretically prove a saturation pattern in softmax-based attention, showing that continuously increasing head dimensions (with the fixed number of heads) leads to diminishing returns in terms of model capacity, especially for long input sequences (Section 4). Together, these *theoretical results* derived for the *fundamental attention mechanism* provide principled guidance for efficient parameter allocation in Transformers, and are also consistently validated through extensive numerical experiments.

Our main contributions are summarized as follows:

- With the fixed model dimension as total parameter budget, we theoretically characterize the trade-off between the number of heads and head dimensions to improve the model performance on information extraction tasks (Theorem 3.1).
- With maintained performance, we mathematically prove a saturation pattern of softmax-based attention (Theorem 4.1 and Corollary 4.2), showing diminishing returns in model expressiveness when continuously increasing head dimensions (the number of heads fixed).
- Our theory is consistently verified through extensive experiments across multiple tasks ( $n$ -gram, next-token prediction and masked language modeling) and different models (customized Transformers, LLAMA-3-8B and LLAMA-3.2-1B), yielding principled and data-agnostic parameter allocation strategies (Corollary 3.2 and Algorithm 1).

**Notations.** For  $k \in \mathbb{N}_+$ , let  $[k] = \{1, 2, \dots, k\}$ . For a vector  $\mathbf{v}$  and  $1 \leq p \leq \infty$ , we denote the  $\ell_p$ -norm of  $\mathbf{v}$  by  $\|\mathbf{v}\|_{\ell_p}$ . For a matrix  $\mathbf{A} = (a_{ij})$ , we denote the spectral norm of  $\mathbf{A}$  by  $\|\mathbf{A}\|_2$ , and  $\|\mathbf{A}\|_\infty$  denotes the maximum norm, i.e., the maximum absolute entry of  $\mathbf{A}$ ;  $\|\mathbf{A}\|_{2,\infty}$  is defined by first computing the  $\ell_2$  norm of each column of  $\mathbf{A}$ , and then taking the maximum among them. For an event  $S$ , we define the indicator  $\mathbb{I}_{\{S\}}$ , which equals to 1 when  $S$  occurs and 0 otherwise. We use the standard big-O notation  $\mathcal{O}$  to hide absolute positive constants. For a matrix  $\mathbf{X}$ , we denote the  $a$ -th to  $b$ -th columns by  $\mathbf{X}_{a:b}$ .

## 2 Preliminaries

In this section, we introduce the Transformer architecture by formulating each single layer. The Transformer maps input sequences to output sequences in the following order.

**Input sequence.** We assume that the input sequence has the length  $L$ , and each token  $\mathbf{x}_t$  ( $t \in [L]$ ) is in  $\mathbb{R}^d$ , hence the input sequence is  $\mathbf{X} = (\mathbf{x}_t)_{t \in [L]} \in \mathbb{R}^{d \times L}$ .

**Embedding layer.** For each token  $\mathbf{x}_t$ , we transform it into a  $D$ -dimensional vector through a linear mapping  $\mathbf{x}_t^{(0)} = \mathbf{W}_E \mathbf{x}_t + \mathbf{b}_E$  ( $\mathbf{W}_E \in \mathbb{R}^{D \times d}$ ,  $\mathbf{b}_E \in \mathbb{R}^D$ ). We denote the embedded input sequence by  $\mathbf{X}^{(0)} \in \mathbb{R}^{D \times L}$ .

**Transformer layer.** The embedding  $\mathbf{X}^{(0)}$  are then passed through  $N$  Transformer layers, each composed of a multi-head self-attention (MHSA) and a feed forward network (FFN) layer:

$$\begin{aligned}\mathbf{X}^{(n-\frac{1}{2})} &= \mathbf{X}^{(n-1)} + \text{MHSA}^{(n)}(\mathbf{X}^{(n-1)}), \quad n \in [N], \\ \mathbf{X}^{(n)} &= \mathbf{X}^{(n-\frac{1}{2})} + \text{FFN}^{(n)}(\mathbf{X}^{(n-\frac{1}{2})}), \quad n \in [N],\end{aligned}$$

where

$$\begin{aligned}\text{MHSA}^{(n)}(\mathbf{X}^{(n-1)}) &= \mathbf{W}_O^{(n)} \cdot \text{Concat}((\text{Attn}^{(n,h)}(\mathbf{X}^{(n-1)}))_{h=1}^H), \\ \text{Attn}^{(n,h)}(\mathbf{X}^{(n-1)}) &= \mathbf{W}_V^{(n,h)} \mathbf{X}^{(n-1)} \cdot \text{softmax}(\langle \mathbf{W}_Q^{(n,h)} \mathbf{X}^{(n-1)}, \mathbf{W}_K^{(n,h)} \mathbf{X}^{(n-1)} \rangle + \mathbf{R}^{(n,h)}),\end{aligned}$$

where  $\mathbf{W}_K^{(n,h)}, \mathbf{W}_Q^{(n,h)}, \mathbf{W}_V^{(n,h)} \in \mathbb{R}^{d_h^{(n)} \times D}$ , and  $\mathbf{W}_O^{(n)} \in \mathbb{R}^{D \times D}$  are learnable parameters (weight matrices) corresponding to the key, query, value and output matrix of the  $h$ -th head at the  $n$ -th layer, and  $\langle \mathbf{A}, \mathbf{B} \rangle := \mathbf{A}^\top \mathbf{B}$ . The activation function  $\text{softmax}(\cdot)$  is performed in a column-wise sense, and  $\text{Concat}(\cdot)$  acts vertically to concatenate  $H$  matrices;  $\mathbf{R}^{(n,h)}$  denotes the relative positional encoding (RPE) matrix ([78, 22, 72, 19, 67, 77, 66, 35, 91]), which satisfies  $\mathbf{R}_{ij}^{(n,h)} = -\infty$  if  $i < j$ . For instance, in the Alibi RPE ([66]), we have  $\mathbf{R}_{ij}^{(n,h)} = \phi(i - j; p^{(n,h)})$ , where  $p^{(n,h)}$  collects learnable parameters and

$$\phi(z; p) = \begin{cases} -p \cdot z, & z \geq 0, \\ -\infty, & \text{otherwise.} \end{cases}$$

**Feed forward network.** The feed forward network (FFN) is applied token-wisely, which maps each column of the internal sequence  $\mathbf{X}^{(n-\frac{1}{2})}$  from  $\mathbb{R}^D$  to  $\mathbb{R}^D$  in the first  $N - 1$  layers, and from  $\mathbb{R}^D$  back to  $\mathbb{R}^d$  in the  $N$ -th layer. According to [18, 34, 55, 50, 47, 46], the FFN has universal approximation property as the number of parameters increases.

### 3 Improving Performance with Fixed Parameter Budget

In this section, we investigate problem (i) regarding parameter allocation in multi-head attention, aiming to determine the optimal number of heads and head dimensions given a fixed total parameter budget (i.e. a fixed model/embedding dimension) that minimizes the learning error. We present our approximation results in Section 3.1, and validate them empirically in Section 3.2.

#### 3.1 Parameter Trade-offs in Information Extraction

We study the information extraction task, formulated as learning linear combinations of input tokens. Many practical Transformer behaviors can be traced to such information extraction mechanisms. For example, sparse memory ([82]) extends n-gram-style pattern extraction, while induction heads implement copy-like operations underlying in-context learning and certain reasoning behaviors ([65]).

Our analysis also extends to nonlinear sequence modeling. From an operator-learning perspective, Transformers approximate causal mappings over sequences, where information extraction corresponds to the linear regime and more complex dependencies arise from higher-order interactions. Such continuous and time-homogeneous nonlinear systems can be modeled by Volterra series, which have recently re-emerged in time-series modeling ([36]), signal processing ([74]), and quantitative finance ([25]).

We first present results for the information extraction task, formulated as follows.

**Linear sequence modeling.** Assume that  $\mathbf{x}_t$  are random vectors defined on a common probability space. They are not required to be independent, identically distributed, isotropic, or sub-Gaussian. Assume only that their  $\ell_2$ -norms are uniformly upper bounded by  $B$ . Consider a sequence-to-sequence modeling framework widely-adopted in prior theoretical studies ([48, 47, 83, 39, 84]):<sup>1</sup>  $\mathbf{y}_t = \text{H}_t(\mathbf{X}) = \sum_{i=0}^L \boldsymbol{\rho}_i \mathbf{x}_{t-i}$  (with a zero padding on  $\mathbf{x}_t$ ), i.e., the output sequences are convolutions

<sup>1</sup>This (linear) sequence modeling framework has been shown to yield important and profound insights into the interplay between model approximation property and data memory structure for multiple sequence models, such as SSMs ([48, 47, 83, 84]), temporal CNNs ([39]), and Transformers ([38]).

of input sequences with the target “kernel”  $\rho_i \in \mathbb{R}^{d \times d}$ . For information extraction tasks, the target kernel  $\rho_i$  can be viewed as the extractor.

Under this setting, the hypothesis space consists of single-layer Transformers with a fixed total parameter budget (i.e. the fixed embedding/model dimension)  $D$ . Let  $\mathcal{T}^D$  denote all single-layer Transformers with a fixed embedding dimension  $D$ , then the approximation problem is defined as

$$\mathcal{E}_D(\mathbf{X}) = \inf_{T \in \mathcal{T}^D} \|\mathbf{H}_t(\mathbf{X}) - T(\mathbf{X})\|_{2,\infty}.$$

In this formulation, with a fixed budget  $D$ , parameters (i.e. the number of heads and head dimensions) must be allocated judiciously to learn the target in the form of linear combinations of the most informative tokens. We establish the following approximation error estimate.

**Theorem 3.1.** *Suppose there are  $H_m$  heads each with the dimension  $d_m$  ( $m \in [M]$ ), and the model dimension  $D = \sum_{m=1}^M H_m \cdot d_m$  is fixed. To approximate the linear target  $\mathbf{H}_t(\mathbf{X}) = \sum_{i=0}^L \rho_i \mathbf{x}_{t-i}$ , for any  $\delta \in (0, 1)$ , with the probability at least  $1 - \delta$ , we have*

$$\begin{aligned} \mathcal{E}_D(X) \leq & B \sum_{m=1}^M \|\rho_m\|_2 \left[ \mathbb{I}_{\{d_m < d\}} \left( \sqrt{1 - \frac{d_m}{d}} \right) + \frac{1.3e^{0.02m}}{H_m} \right] + B \sum_{k=M+1}^L \|\rho_k\|_2 \\ & + B \sum_{m=1}^M \|\rho_m\|_2 \cdot C \sqrt{\frac{\log(2M/\delta)}{d}}. \end{aligned} \quad (1)$$

We provide the detailed proof of Theorem 3.1 in Appendix D, and briefly summarize the main construction ideas as follows:

1. Since the embedding dimension is fixed, we first compress the information contained in each token, which yields the first term in the estimate (Eq. (1)).
2. We then employ  $M$  groups of heads, where each group is responsible for extracting a single token ahead. The approximation error decreases as more heads are employed, corresponding to the second term in the estimate.
3. Finally, since only the  $M$  most informative tokens are extracted, which truncates the target at the time step  $M$ , we derive the third term in the estimate.
4. Without loss of generality,  $\rho_i$  can be reordered so that  $\|\rho_i\|_2$  decreases with  $i$ , minimizing the third term.

According to Theorem 3.1, on one hand, setting a large head dimension is beneficial for preserving token information, reducing the first error term in equation (1); on the other hand, increasing the number of heads is also favorable for approximating the extraction function, reducing the second error term. However, since the total budget/number of parameters is fixed, one cannot increase the number of heads and head dimensions simultaneously beyond certain constraints, which inevitably leads to a trade-off, see Appendix F.1 for illustration.

Here, we formulate the information extraction task as learning to construct linear combinations of input tokens, typically represented as a convolutional form. For a single Transformer layer, [40] also identifies a trade-off phenomenon in this type of tasks, as illustrated in their experiments (see Figure 14 in Appendix G): The trade-off between the number of attention heads and learning errors emerges in different sequence modeling tasks, with its manifestation depending on the type and strength of underlying memory structures captured by the convolution kernel  $\rho$ . Their results demonstrate that these parameter-performance trade-offs consistently arise across a variety of memory structures (e.g., exponentially/polynomically decaying kernels, delta/Airy function kernels). However, in certain cases such as exponentially and polynomically decaying kernels, these trade-offs may vanish when the memory strength varies. Theorem 3.1 also provides a theoretical explanation for this phenomenon.

Based on Theorem 3.1, we consequently have the following parameter allocation strategy.

**Corollary 3.2** (Principled parameter allocation). *Under the same conditions of Theorem 3.1, the optimal allocation of parameters can be achieved by solving the following optimization problem*

$$\begin{aligned} \min_{M, H_m, d_m} \quad & \sum_{m=1}^M \|\rho_m\|_2 \left( \mathbb{I}_{\{d_m \leq d\}} \sqrt{1 - \frac{d_m}{d}} + \frac{1.3 e^{0.02m}}{H_m} \right) + \sum_{k=M+1}^L \|\rho_k\|_2, \\ \text{s.t.} \quad & \sum_{m=1}^M H_m \cdot d_m = D. \end{aligned} \quad (2)$$

Corollary 3.2 is a direct conclusion of Theorem 3.1. Next, we provide some specific tasks for illustrations of Corollary 3.2.

**Example I: N-gram.** The  $n$ -gram task in language modeling involves predicting the next token in a sequence using the previous  $n$  tokens as contexts ([42, 43, 13, 11, 82]). This task aims to capture local statistical dependencies in texts and serves as a basic benchmark for language modeling. We take a simple 4-gram task as an example, where

$$\mathbf{y}_t = \mathbf{x}_{t-1} + \mathbf{x}_{t-2} + \mathbf{x}_{t-3} + \mathbf{x}_{t-4}. \quad (3)$$

In this case,  $\rho_i = \mathbb{I}_{\{1 \leq i \leq 4\}} \mathbf{I}_{d \times d}$ . Let  $d = 8$ ,  $D = 256$  and  $B = 1$ , then Eq. (2) becomes

$$\begin{aligned} \min_{M, H_m, d_m} \quad & \sum_{m=1}^M \mathbb{I}_{\{d_m \leq 8\}} \sqrt{1 - \frac{d_m}{8}} + \sum_{m=1}^M \frac{1.3 e^{0.02m}}{H_m} + (4 - M), \\ \text{s.t.} \quad & \sum_{m=1}^M H_m \cdot d_m = 256, \quad M \leq 4. \end{aligned}$$

We find numerically by search that the optimal solution is attained at  $M = 4$ ,  $d_m = H_m = 8$  ( $m = 1, 2, 3, 4$ ). See further experimental validations in Figure 1a.

**Example II: Induction head.** The induction head (IH) is a mechanism in Transformers that learns to copy patterns by attending from a repeated token to its earlier occurrence, enabling sequence continuation beyond training contexts ([26]). Implementing an IH requires two Transformer layers ([69, 70, 68, 81]). In this construction, the first layer plays a role similar to an  $n$ -gram task, extracting linear combinations of tokens. This allows us to derive an allocation strategy, with details provided in Appendix F.3.

**Nonlinear sequence modeling.** Theorem 3.1 can be further extended to nonlinear settings. Specifically, we consider a continuous, time-homogeneous system that admits the decomposition in [83, 21]. The corresponding results and proofs are deferred to Appendix F.4.

### 3.2 Empirical Results

In this section, we present experimental results that validate Theorem 3.1 or Corollary 3.2. It is numerically shown that the trade-off between the number of attention heads and head dimensions exists to enhance model performance under a fixed total parameter budget.

**Verification of Corollary 3.2.** We empirically validate the parameter allocation strategy on the above 4-gram task (Eq. (3)). A single-layer Transformer encoder with sinusoidal positional encoding ([78]) is trained on a synthetic dataset (white noises) to learn the target Eq. (3). With a fixed embedding dimension  $D = 256$ , we conduct experiments under multiple learning rates and random seeds, and report the results corresponding to the best-performed configuration as an estimate of the approximation error. As shown in Figure 1a, the optimal allocation occurs at  $M = 4$  with  $d_m = H_m = 8$  ( $m = 1, 2, 3, 4$ ), which is consistent with the above theoretical prediction and confirms the proposed strategy (Eq. (2)).

<sup>1</sup>For other values of  $D$  in Figure 1(b) and (c), please refer to Appendix F.2 (see Figure 8, 9, 10 and 11).

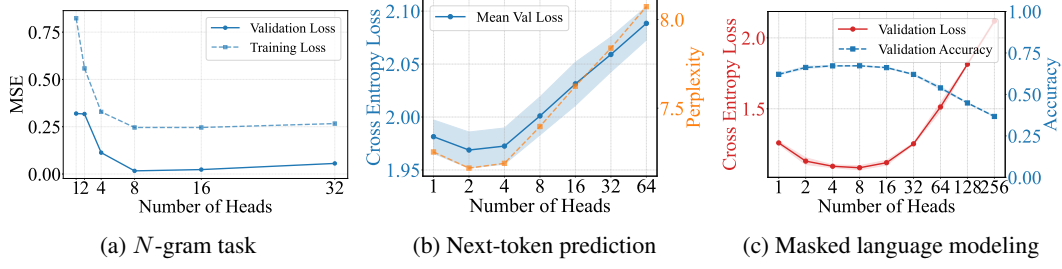


Figure 1: Trade-off to enhance model performance with a fixed total parameter budget.<sup>1</sup>

**More trade-off examples.** In Figure 1b and 1c, we also investigate the trade-off phenomenon under a fixed total parameter budget. In Figure 1b, a three-layer causal Transformer decoder is trained on the Tiny Shakespeare corpus for the next-token prediction task. Fix the bottleneck dimension as  $D = 128$ , we vary the number of attention heads and evaluate model performance using both cross entropy loss and perplexity (PPL), with results averaged over multiple random seeds. In Figure 1c, we consider a masked language modeling (MLM) task by training a three-layer character-level Transformer encoder (with full self-attention) also on the Tiny Shakespeare corpus. Following the standard MLM protocol [22], 15% of input tokens are randomly masked, and the model is trained to reconstruct the masked tokens while ignoring unmasked positions during optimization. Fix the model dimension as  $D = 256$ , we report both cross-entropy loss and masked token prediction accuracy, with all results averaged over three independent runs. Across both experiments, we consistently observe a clear trade-off phenomenon, demonstrating that appropriate architectural choices can improve model performance under a fixed total parameter budget. We additionally include code, math, and logic-intensive tasks on a 330M model to broaden the scope in Appendix F.2 Figure 12.

**Practical guideline.** Based on these observations, we suggest that practitioners should start with relatively smaller  $H$  (thus larger  $d_H$ ), and then perform a local sweep over a small range of  $H$  to identify the optimal tradeoff point. In particular, overly large  $H$  should be avoided, as excessively small  $d_H$  can hurt performance even in simple tasks.

## 4 Reducing Parameters with Maintained Performance

In this section, we turn to problem (ii), aiming to reduce the required total parameters by decreasing head dimensions without significantly degrading model performance. This is provably guaranteed by the saturation pattern of softmax-based attention, meaning that continuously increasing head dimensions can lead to diminishing returns in learning errors, particularly for long sequences. This saturation pattern is first mathematically proved, implying potential model efficiency guidance with reduced parameters (head dimensions) while maintaining “lossless” model capacity.

### 4.1 Softmax Saturation and Parameter Reduction

The theory in this section is motivated by understanding modern architectures such as multi-head latent attention (MLA; [51]), as well as the phenomenon of rank collapse in attention. While several prior works have pointed out the low-rank nature of attention, the theoretical understanding remains incomplete. Our results contribute to this line of research by quantifying a redundancy-related estimate particularly when the sequence length is sufficiently large.

We first introduce the notation. Let

$$\text{Logits}_{d_H}(\mathbf{W}_Q, \mathbf{W}_K, \mathbf{X}) := \langle \mathbf{W}_Q \mathbf{X}, \mathbf{W}_K \mathbf{X} \rangle + \mathbf{R} \in \mathbb{R}^{L \times L},$$

with  $\mathbf{X} \in \mathbb{R}^{D \times L}$ ,  $\mathbf{W}_K, \mathbf{W}_Q \in \mathbb{R}^{d_H \times D}$ , and  $\mathbf{R} \in \mathbb{R}^{L \times L}$ . For  $l \in [L]$ , let  $\mathbf{e}_l$  be the vector with the  $l$ -th element as 1 and all other elements as 0. For each query position  $l \in [L]$ , let

$$\mathcal{S}_l := \{j \in [L] : R_{lj} > -\infty\}, \quad s_l := |\mathcal{S}_l|,$$

be the set and number of unmasked key positions. Define the  $l$ -th query logit vector and attention vector, restricted to their active coordinates, by

$$\mathbf{z}_l := (\mathbf{e}_l^\top \text{Logits}_{d_H})_{\mathcal{S}_l}, \quad \mathbf{a}_l := \text{softmax}(\mathbf{z}_l),$$

and let

$$J_l := \frac{\partial a_l}{\partial z_l} = \text{diag}(a_l) - a_l a_l^\top.$$

Thus  $s_l = L$  in the unmasked case, whereas  $s_l = l$  under the standard causal mask. Then, we have the following estimate on the column-wise Jacobian of softmax-based attention.

**Theorem 4.1** (Softmax sensitivity under bounded active logits). *Assume  $\|W_Q\|_2 \leq M_Q$ ,  $\|W_K\|_2 \leq M_K$ ,  $\|x_t\|_2 \leq B$  ( $t \in [L]$ ), and that the finite positional-bias entries satisfy  $|R_{lj}| \leq R_0$  ( $j \in \mathcal{S}_l$ ,  $l \in [L]$ ). Let  $\Delta := M_Q M_K B^2 + R_0$ . Then, for every  $l \in [L]$ ,*

$$\|J_l\|_2 \leq \frac{2e^{2\Delta}}{s_l}.$$

Consequently,

$$\|J_l\|_2 = \mathcal{O}(L^{-1})$$

for unmasked attention, while under the standard causal mask,  $\|J_l\|_2 = \mathcal{O}(l^{-1})$ . The hidden constants depend only on  $M_Q, M_K, B$ , and  $R_0$ .

According to Theorem 4.1, the ‘‘slope’’ of softmax activations is quite flat, particularly for long sequences. This further implies the following corollary, which characterizes the approximation error when ‘‘compressing’’ attention score matrices with reduced head dimensions.

**Corollary 4.2** (Head-dimension reduction). *Under the assumptions of Theorem 4.1, let  $M := W_Q^\top W_K$ , and let  $M_{d_h}$  be its rank- $d_h$  truncated SVD. Choose  $\hat{W}_Q, \hat{W}_K \in \mathbb{R}^{d_h \times D}$  such that  $\hat{W}_Q^\top \hat{W}_K = M_{d_h}$ . Let  $\hat{A}$  denote the attention-score matrix obtained by replacing  $W_Q, W_K$  with  $\hat{W}_Q, \hat{W}_K$  while keeping the same positional bias and mask. Then for every query  $l \in [L]$ ,*

$$\left\| e_l^\top (\text{AttnScore}_{d_H} - \hat{A}) \right\|_2 \leq \frac{2e^{2\Delta}}{s_l} \|X\|_2^2 \sigma_{d_h+1}(M).$$

Hence the per-query approximation error is

$$\mathcal{O}\left(\frac{\|X\|_2^2 \sigma_{d_h+1}(M)}{L}\right)$$

in the unmasked case and

$$\mathcal{O}\left(\frac{\|X\|_2^2 \sigma_{d_h+1}(M)}{l}\right)$$

under the standard causal mask.

In particular, for any fixed  $\alpha \in (0, 1)$ , the late-query block  $\mathcal{I}_\alpha := \{l : l \geq \lceil \alpha L \rceil\}$  satisfies

$$\left( \sum_{l \in \mathcal{I}_\alpha} \left\| e_l^\top (\text{AttnScore}_{d_H} - \hat{A}) \right\|_2^2 \right)^{1/2} \leq \frac{2e^{2\Delta}}{\alpha \sqrt{L}} \|X\|_2^2 \sigma_{d_h+1}(M). \quad (4)$$

**Remark 4.3** (Causal masking and early queries). For causal attention, the global Frobenius error over all queries does not in general inherit an  $L^{-1/2}$  rate from Theorem 4.1, because

$$\sum_{l=1}^L l^{-2} = \mathcal{O}(1).$$

The vanishing  $L^{-1/2}$  Frobenius rate remains valid for any fixed late-query fraction  $l \geq \alpha L$ . Early queries have much smaller active support and, moreover,  $\text{rank}(J_l) \leq l - 1$ .

**Extension to rotary positional embeddings.** The preceding analysis uses an additive relative positional bias. The same head-dimension reduction principle can also be extended to rotary positional embeddings (RoPE), although the corresponding low-rank approximation takes a slightly different form. Let  $\mathcal{R}_t \in \mathbb{R}^{d_H \times d_H}$  denote the RoPE rotation at position  $t$ . Since  $\mathcal{R}_t$  is orthogonal, the relative rotation

$$\mathcal{R}_{l,j} := \mathcal{R}_l^\top \mathcal{R}_j$$

---

**Algorithm 1** Head Dimension Reduction Strategy

---

**Input:** Teacher parameters  $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d_H \times D}$ , student dimension  $d_h$   
**Output:** Approximated student parameters  $\hat{\mathbf{W}}_Q, \hat{\mathbf{W}}_K \in \mathbb{R}^{d_h \times D}$   
**1. SVD Initialization:** Initialize  $\hat{\mathbf{W}}_Q, \hat{\mathbf{W}}_K$  via truncated SVD of  $\mathbf{W}_Q^\top \mathbf{W}_K$   
**2. Refinement :** Further train  $\hat{\mathbf{W}}_Q, \hat{\mathbf{W}}_K$  with other parameters fixed

---

also satisfies

$$\mathcal{R}_{l,j}^\top \mathcal{R}_{l,j} = I, \quad \|\mathcal{R}_{l,j}\|_2 = 1.$$

Accordingly, the  $(l, j)$ -th attention logit can be written as

$$z_{lj} = x_l^\top \mathbf{W}_Q^\top \mathcal{R}_{l,j} \mathbf{W}_K x_j.$$

Let  $\mathbf{W}_{Q,r}$  and  $\mathbf{W}_{K,r}$  denote the rank- $r$  truncated-SVD approximations of  $\mathbf{W}_Q$  and  $\mathbf{W}_K$ , respectively. Then, uniformly over all relative positions,

$$\|\mathbf{W}_Q^\top \mathcal{R}_{l,j} \mathbf{W}_K - \mathbf{W}_{Q,r}^\top \mathcal{R}_{l,j} \mathbf{W}_{K,r}\|_2 \leq \|\mathbf{W}_Q\|_2 \sigma_{r+1}(\mathbf{W}_K) + \sigma_{r+1}(\mathbf{W}_Q) \|\mathbf{W}_K\|_2.$$

Thus, the approximation error caused by reducing the Q/K head dimension remains controlled uniformly over the RoPE rotations. Combining this estimate with the softmax sensitivity bound in Theorem 4.1 yields the same inverse-support-size saturation behavior. In particular, under causal masking, the sensitivity scales as  $\mathcal{O}(1/l)$  for the  $l$ -th query. A formal statement and proof are provided in Appendix E.2.

The detailed proofs of Theorem 4.1 and Corollary 4.2 are provided in Appendix E.1. Specifically, Corollary 4.2 is derived by applying an SVD-based low-rank approximation to original logits matrices, together with Theorem 4.1. According to Corollary 4.2, for long sequences ( $L \gg 1$ ), compressing original attention score matrices in large head dimensions with those in smaller head dimensions can lead to minor errors, potentially implying a principled strategy for additional parameter reduction with approximately maintained model capacity.

**Principled strategy for parameter reduction.** Corollary 4.2 shows that a student head with reduced dimensions can approximate a teacher head with larger dimensions by incurring only small errors, particularly for long sequences. Therefore, a principled and practical strategy for additional parameter reduction can be motivated as Algorithm 1. This procedure enables further model compression with enhanced parameter efficiency and lossless learning performance.

*Remark 4.4.* The dimension-reduction strategy based on Corollary 4.2 is complementary to existing low-rank techniques rather than a direct replacement. Building upon existing low-rank methods such as GQA (mainly used in LLAMA in Figure 4 and Figure 5), our approach further reduces the parameter budget without significantly degrading model performance. Rather than designing new architectures or training compression-aware models from scratch, we consider the scenario where a large pretrained model is already available, and the goal is to compress it into a smaller model with reduced computation costs and lossless performance. This setting is particularly relevant for deployment on resource-constrained devices or for users with limited computation resources.

**Complexity analysis.** We consider computation and memory costs arising from self-attention. By reducing  $\mathbf{W}_Q, \mathbf{W}_K$  with  $\hat{\mathbf{W}}_Q, \hat{\mathbf{W}}_K$ , the computation complexity for linear projections  $\hat{\mathbf{Q}} = \hat{\mathbf{W}}_Q \mathbf{X}$  and  $\hat{\mathbf{K}} = \hat{\mathbf{W}}_K \mathbf{X}$  drops to  $\mathcal{O}((d_h/d_H)^2 L D^2)$ , and  $\mathcal{O}((d_h/d_H) L^2 D)$  for the attention score  $\langle \hat{\mathbf{Q}}, \hat{\mathbf{K}} \rangle$ , definitely saving the parameters budget for storing query/key matrices ([78, 85]).

## 4.2 Empirical Results

In this section, we first numerically verify Theorem 4.1 and Corollary 4.2. We then implement Algorithm 1 on LLM-scale experiments (LLAMA-3-8B and LLAMA-3.2-1B), to empirically demonstrate that compressing head dimensions does not significantly degrade model performance for both pre-training and down-stream tasks. Finally, we validate the efficiency gains of the proposed algorithm.

**Verification of Theorem 4.1.** In Figure 2, we empirically validate Theorem 4.1 by analyzing how the spectral norm of the attention Jacobian varies with sequence lengths. Specifically, we study the TinyLlama-1.1B model and focus on Layer 0 and Head 0, using sequences from the WikiText-103 dataset. For sequence lengths  $L \in \{4, 8, 16, 32, 64, 128, 256, 512, 1024\}$ , we sample 10 different sequences for each length.<sup>2</sup> For each sequence, we compute the attention Jacobian with respect to logits, calculate its spectral norm, and take average over the ten sequences.

**Verification of Corollary 4.2.** In Figure 3, we numerically validate Corollary 4.2 by analyzing the error to fit attention heads of pre-trained Transformer language models with those of reduced head dimensions. Using the LiteLlama-460M model as the teacher, we isolate a single teacher head (Layer 5, Head 0) and train a student head with reduced query-key dimensions  $d_h \in \{4, 8, 16, 24, 32, 48, 64\}$  (input and rotary positional embeddings are retained). For  $d_h < d_H = 64$ , the student query and key matrices are initialized via truncated SVD of the teacher’s. The training objective is to fit the teacher attention weights on a mixed dataset of synthetic and WikiText-103 sequences with the length 1024.<sup>1</sup> We use the Adam optimizer with the learning rate  $10^{-3}$  for 10,000 epochs, and the performance is evaluated by mean squared error (Eq. (4)) over masked causal attention weights.

The above experiments reflect that a single attention head can be effectively compressed via head dimension reduction for real-world language models and datasets. We further show that head dimension reduction applied to multi-head attention can largely maintain LLM’s performance in terms of both pre-training perplexity and downstream accuracy.

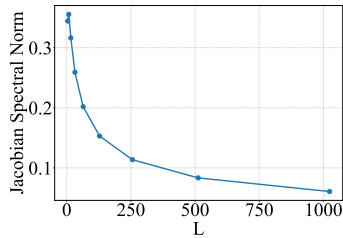


Figure 2: Numerical verification of Theorem 4.1: Softmax-based attention becomes flat (w.r.t. logits) for long sequences.

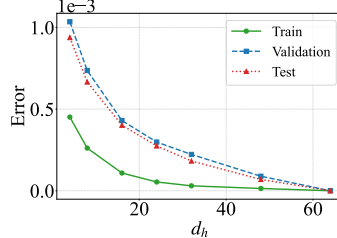


Figure 3: Numerical verification of Corollary 4.2: Students with reduced dimensions ( $d_h$ ) can approximate teacher heads ( $d_H = 64 > d_h$ ).

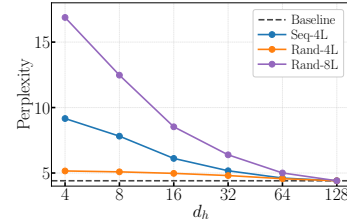


Figure 4: Reduced head dimensions in LLAMA-3-8B lead to saturated perplexity on WikiText-2.

**LLAMA-3-8B: Saturated perplexity.** To further study the model reduction effect of pre-trained large Transformer models, we apply head-wise dimension reduction to the multi-head self-attention modules of LLAMA-3-8B (with an original head dimension  $d_H = 128$ ). Specifically, for selected layers, we parameterize reduced models with query/key weight matrices as low-rank decompositions ( $\text{rank-}d_h < d_H$ ) of those in original LLAMA-3-8B, and retain all the other configurations. These low-rank weights are initialized via truncated SVD of those in the original model, and we freeze all the other original parameters during training, to minimize the mean squared error (MSE) between the reconstructed low-rank weights and their original (full-rank) counterparts. The training is performed using the AdamW optimizer (with a learning rate  $2 \times 10^{-4}$ ) for 150 steps. We evaluate performance using perplexity (PPL) on WikiText-2 via a sliding-window protocol. Experiments cover multiple head dimensions  $d_h \in \{4, 8, 32, 64, 128\}$  across different layers: *Seq-4L* (four sequential layers), *Rand-4L* (four random layers), and *Rand-8L* (eight random layers). As shown in Figure 4, increasing head dimensions yields diminishing returns in perplexity, indicating that head dimensions can be reduced with approximately maintained model performance.

**LLAMA-3.2-1B: Saturated downstream accuracy.** We further evaluate head-dimension reduction on downstream tasks using LLAMA-3.2-1B, whose original head dimension is  $d_H = 64$ . For selected layers (early 1–8, middle 5–12, late 9–16, odd, and all 1–16), each query/key projection head

<sup>2</sup>Each sequence is generated by randomly selecting contiguous spans from WikiText-103, which is tokenized to a fixed length  $L$  (padded if necessary). For large  $L$ , padding ensures a consistent input size while maintaining semantically meaningful contents.

is permanently reparameterized as a rank- $d_h$  factorization,

$$W_h = L_h R_h, \quad \text{rank}(W_h) \leq d_h,$$

with  $d_h \in \{2, 4, 8, 16, 32, 64\}$ , while all other model components are kept unchanged. The low-rank factors are initialized using the truncated SVD of the corresponding pretrained Q/K projections. Training consists of two stages. In Stage I, the compressed model is refined against the uncompressed language-model teacher on WikiText-2 for 400 steps by minimizing a combination of hidden-state MSE and output-logit KL divergence, while only the low-rank Q/K factors are optimized. In Stage II, a full-rank SST-2 classifier is used as the teacher to distill the compressed model for three epochs, with an objective combining label cross-entropy, output KL divergence, and hidden-state matching; only the low-rank Q/K factors and the classification head are trainable. As shown in Figure 5, the compressed models retain strong downstream performance even under substantial reductions of the Q/K head dimension, while the degree of robustness depends noticeably on which layers are compressed, indicating non-uniform redundancy across depth.

**Ablation study.** To further isolate the contribution of each component in this pipeline, we conduct controlled ablations by fixing the compressed layers to 5–12 and  $d_h = 16$ , with complete results reported in Appendix G.6. Replacing the SVD initialization with Kaiming initialization increases the post-refinement WikiText-2 loss from 3.215 to 3.582 and decreases SST-2 accuracy from 93.81% to 91.74%. Removing Stage I refinement decreases the downstream accuracy from 93.81% to 91.86%, whereas removing Stage II downstream distillation results in only 47.48% accuracy. These comparisons demonstrate complementary roles of the three components: truncated-SVD initialization provides a stronger low-rank starting point, intermediate refinement better preserves the pretrained representation after compression, and downstream distillation is essential for recovering task-specific performance.

**Computation efficiency analysis.** To assess the improvement of model computation efficiency caused by head dimension reduction, we test for a customized Transformer model, which is a 6-layer encoder with the embedding dimension  $D = 256$  and  $H = 8$  attention heads. We vary the head dimension  $d_h \in \{4, 8, 16, 32, 64, 128\}$  to calculate the computation overhead scaling. Here, we use synthetic sequences with the sequence length  $L = 256$  and the batch size  $B = 256$  to ensure a compute-bound regime. To measure realistic training overhead, for the classification task, we run 400 optimization steps using the AdamW optimizer with a learning rate  $2 \times 10^{-4}$  to minimize a cross-entropy loss. Benchmarks are evaluated on an NVIDIA A100 GPU (80GB). The FLOPs are calculated via `thop` and averaged over 100 iterations. As indicated in Figure 6, both GFLOPs and empirical forward/backward time scale non-linearly with  $d_h$ . The results confirm that low head dimensions can help to significantly reduce forward/backward time and model calculations, offering a path to more efficient inference and training.

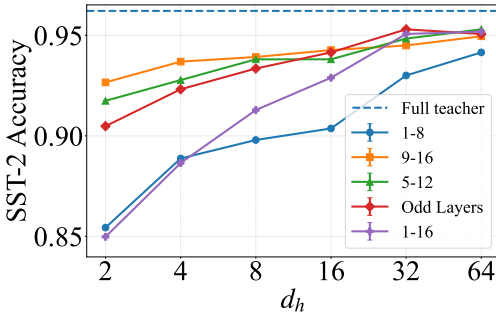


Figure 5: Reduced head dimensions in LLAMA-3.2-1B yield saturated SST-2 accuracy after distillation and fine-tuning.<sup>3</sup>

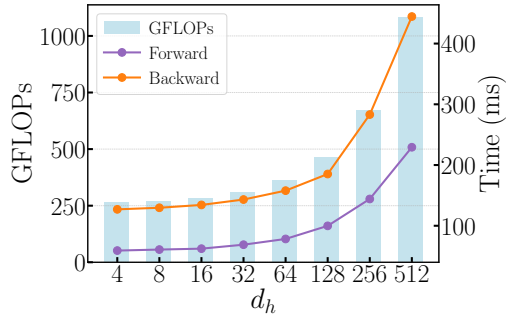


Figure 6: Computation efficiency evaluations of Transformers regarding head dimensions.

<sup>3</sup>The performance gap stems from the varied tuning capacity, as the student model is partially fine-tuned only for the selected layer given reduced head dimensions, while the teacher model is fully fine-tuned.

## 5 Conclusion and Limitations

In this work, we study performance–efficiency trade-offs in Transformers from an approximation theory perspective, focusing on the allocation of attention heads and head dimensions. Under a fixed parameter budget, we characterize an intrinsic trade-off between the number of heads and their dimensions, and additionally show that softmax-based attention exhibits saturation, where increasing head dimensions yields diminishing returns, especially for long sequences. Experiments across synthetic and language modeling tasks support these findings and provide guidance for the parameter-efficient architecture design. For limitations, our theoretical analysis mainly focuses on idealized settings, and thus may not capture all architectural details. Extending the analysis to more realistic architectures and intricate optimization dynamics remains promising as an important direction for future work.

## References

- [1] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.
- [2] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.
- [3] Noah Amsel, Gilad Yehudai, and Joan Bruna. Quality over quantity in attention layers: When adding more heads hurts. In *The Thirteenth International Conference on Learning Representations*, 2024.
- [4] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection, 2023. URL <https://arxiv.org/abs/2306.04637>.
- [5] Brian R Bartoldson, Bhavya Kailkhura, and Davis Blalock. Compute-efficient deep learning: Algorithmic trends and opportunities. *Journal of Machine Learning Research*, 24(122):1–77, 2023.
- [6] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.
- [7] Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of transformers to recognize formal languages. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- [8] Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *International conference on machine learning*, pages 864–873. PMLR, 2020.
- [9] A. M. Bran et al. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 2024. URL <https://www.nature.com/articles/s42256-024-00832-8>.
- [10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [11] Christian Buck, Kenneth Heafield, and Bas van Ooyen. N-gram counts and language models from the common crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 3579–3584, Reykjavik, Iceland, May 2014. URL [http://www.lrec-conf.org/proceedings/lrec2014/pdf/1097\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/1097_Paper.pdf).
- [12] F. Busch. Current applications and challenges in large language models for patient care: a systematic review. *Nature Medicine*, 2025. URL <https://www.nature.com/articles/s43856-024-00717-2>.

- [13] Ciprian Chelba, Mohammad Norouzi, and Samy Bengio. N-gram language modeling using recurrent neural network estimation. *CoRR*, abs/1703.10724, 2017. URL <https://arxiv.org/abs/1703.10724>.
- [14] Xingwu Chen and Difan Zou. What can transformer learn with varying depth? case studies on sequence learning tasks, 2024. URL <https://arxiv.org/abs/2404.01601>.
- [15] Minsik Choi, Hyegang Son, Changhoon Kim, and Young Geun Kim. Entropy meets importance: A unified head importance-entropy score for stable and efficient transformer pruning. *arXiv preprint arXiv:2510.13832*, 2025.
- [16] Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarloš, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers, 2020. arXiv:2009.14794.
- [17] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Hailong Ren, Xiaolin Wei, Huaxin Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- [18] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989. doi: 10.1007/BF02551274.
- [19] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019. URL <https://arxiv.org/abs/1901.02860>.
- [20] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems*, volume 35, 2022.
- [21] R.J.P. de Figueiredo. The volterra and wiener theories of nonlinear systems. *Proceedings of the IEEE*, 70(3):316–317, 1982. doi: 10.1109/PROC.1982.12307.
- [22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 2019.
- [23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [24] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022.
- [25] Omar El Euch and Mathieu Rosenbaum. The characteristic function of rough heston models. *Mathematical Finance*, 29(1):3–38, 2019.
- [26] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.
- [27] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.
- [28] Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. *International Conference on Machine Learning*, 2023.
- [29] Yuan Gong, Cheng-I Jeff Lai, Yu-An Chung, and James Glass. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10699–10709, 2022.

- [30] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [31] Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.
- [32] Ali Hassani, Steven Walton, Nikhil Shah, Jiachen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6185–6194, 2023.
- [33] Ali Hatamizadeh, Hongxu Yin, Jan Kautz, and Pavlo Molchanov. Fastervit: Fast vision transformers with hierarchical attention. *arXiv preprint arXiv:2306.06189*, 2023.
- [34] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. doi: 10.1016/0893-6080(89)90020-8.
- [35] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer. *arXiv preprint arXiv:1809.04281*, 2018.
- [36] Yixue Huang, Hao Li, Jiaming Fan, Jiadi Li, Ding Wanli, Ying Jiang, Meikang Qiu, and Hao Jiang. Rethinking nonlinear dynamics in deep time series models.
- [37] Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024.
- [38] Haotian Jiang and Qianxiao Li. Approximation rate of the transformer architecture for sequence modeling. *Advances in Neural Information Processing Systems*, 37:68926–68955, 2024.
- [39] Haotian Jiang, Zhong Li, and Qianxiao Li. Approximation theory of convolutional architectures for time series modelling. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4961–4970, 2021. URL <https://proceedings.mlr.press/v139/jiang21d.html>.
- [40] Haotian Jiang, Zeyu Bao, Shida Wang, and Qianxiao Li. Numerical investigation of sequence modeling theory using controllable memory functions, 2025. URL <https://arxiv.org/abs/2506.05678>.
- [41] Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? *arXiv preprint arXiv:2307.14023*, 2023.
- [42] S M Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3): 400–401, 1987.
- [43] R Kneser and H Ney. Improved backing-off for m-gram language modeling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 181–184. IEEE, 1995.
- [44] Siddique Latif, Aun Zaidi, Heriberto Cuayahuitl, Fahad Shamshad, Moazzam Shoukat, and Junaid Qadir. Transformers in speech processing: A survey. *arXiv preprint arXiv:2303.11607*, 2023.
- [45] Eunho Lee and Youngbae Hwang. Automatic channel pruning for multi-head attention. *arXiv preprint arXiv:2405.20867*, 2024.
- [46] Qianxiao Li, Ting Lin, and Zuowei Shen. Deep learning via dynamical systems: An approximation perspective. *Journal of the European Mathematical Society*, 25(5):1671–1709, 2022.

- [47] Zhong Li, Jiequn Han, Qianxiao Li, and Weinan E. On the curse of memory in recurrent neural networks: Approximation and optimization analysis. *International Conference on Learning Representations*, 2021.
- [48] Zhong Li, Jiequn Han, Weinan E, and Qianxiao Li. Approximation and optimization theory for linear continuous-time recurrent neural networks. *Journal of Machine Learning Research*, 23(42):1–85, 2022. URL <http://jmlr.org/papers/v23/21-0368.html>.
- [49] Valerii Likhoshesterov, Krzysztof Choromanski, and Adrian Weller. On the expressive flexibility of self-attention matrices. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8773–8781, Jun. 2023. doi: 10.1609/aaai.v37i7.26055. URL <https://ojs.aaai.org/index.php/AAAI/article/view/26055>.
- [50] Hongzhou Lin and Stefanie Jegelka. Resnet with one-neuron hidden layers is a universal approximator. *Advances in neural information processing systems*, 31, 2018.
- [51] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [52] Y. Liu, Q. Sun, and D. R. Kapadia. Integrating large language models into robotic autonomy: A review of motion, voice, and training pipelines. *AI*, 6(7):158, 2025. URL <https://www.mdpi.com/2673-2688/6/7/158>.
- [53] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.
- [54] Wei-Tsung Lu, Ju-Chiang Wang, Minz Won, Keunwoo Choi, and Xuchen Song. Spectnt: a time-frequency transformer for music audio. *arXiv preprint arXiv:2110.09127*, 2021.
- [55] Chao Ma, Lei Wu, et al. A priori estimates of the population risk for two-layer neural networks. *arXiv preprint arXiv:1810.06397*, 2018.
- [56] Chao Ma, Stephan Wojtowytsch, Lei Wu, and Weinan E. Towards a mathematical understanding of neural network-based machine learning: what we know and what we don’t. *arXiv preprint arXiv:2009.10713*, 2020.
- [57] Weicheng Ma, Xinyu Chen, Lijun Wu, Ran He, and Yefeng Wang. Contributions of transformer attention heads in multi- and cross-lingual tasks. *arXiv preprint arXiv:2108.08375*, 2021.
- [58] Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*, 2023.
- [59] Xiangbin Meng, Xiangyu Yan, Kuo Zhang, Da Liu, Xiaojuan Cui, Yaodong Yang, Muhan Zhang, Chunxia Cao, Jingjia Wang, Xuliang Wang, et al. The application of large language models in medicine: A scoping review. *Iscience*, 27(5), 2024.
- [60] William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. *arXiv preprint arXiv:2310.07923*, 2023.
- [61] William Merrill, Ashish Sabharwal, and Noah A Smith. Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics*, 10: 843–856, 2022.
- [62] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.
- [63] Timur Mudarisov, Mikhail Burtsev, Tatiana Petrova, and Radu State. Limitations of normalization in attention mechanism, 2025. URL <https://arxiv.org/abs/2508.17821>.
- [64] Ken M. Nakanishi. Scalable-softmax is superior for attention. *arXiv preprint arXiv:2501.19399*, 2025.

- [65] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.
- [66] Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation, 2022. URL <https://arxiv.org/abs/2108.12409>.
- [67] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- [68] Nived Rajaraman, Marco Bondaschi, Kannan Ramchandran, Michael Gastpar, and Ashok V Makkuva. Transformers on markov data: Constant depth suffices. *Advances in Neural Information Processing Systems*, 37:137521–137556, 2024.
- [69] Clayton Sanford, Daniel Hsu, and Matus Telgarsky. One-layer transformers fail to solve the induction heads task, 2024. URL <https://arxiv.org/abs/2408.14332>.
- [70] Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth, 2024. URL <https://arxiv.org/abs/2402.09268>.
- [71] Hemant Saratchandran, Jianqiao Zheng, Yiping Ji, Wenbo Zhang, and Simon Lucey. Rethinking softmax: Self-attention with polynomial activations. *arXiv preprint arXiv:2410.18613*, 2024.
- [72] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.
- [73] Lingfeng Shen, Aayush Mishra, and Daniel Khashabi. Do pre-trained transformers really learn in-context by gradient descent? 2023.
- [74] Sidney B Shiki, Samuel Da Silva, and Michael D Todd. On the application of discrete-time volterra series for the damage detection problem in initially nonlinear systems. *Structural Health Monitoring*, 16(1):62–78, 2017.
- [75] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.
- [76] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.
- [77] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL <https://arxiv.org/abs/2104.09864>.
- [78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
- [79] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, 2019.
- [80] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.
- [81] Mingze Wang, Ruoxi Yu, Weinan E, and Lei Wu. How transformers get rich: Approximation and dynamics analysis, 2025. URL <https://arxiv.org/abs/2410.11474>.

- [82] Mingze Wang et al. Understanding the expressive power and mechanisms of transformer for sequence modeling. *Advances in Neural Information Processing Systems*, 37:25781–25856, 2024.
- [83] Shida Wang and Qianxiao Li. Stableness: Alleviating the curse of memory in state-space models through stable reparameterization, 2024. URL <https://arxiv.org/abs/2311.14495>.
- [84] Shida Wang, Zhong Li, and Qianxiao Li. Inverse approximation theory for nonlinear recurrent neural networks. *arXiv preprint arXiv:2305.19190*, 2023.
- [85] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.
- [86] E Weinan, Chao Ma, and Lei Wu. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, 2022.
- [87] Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In *International Conference on Machine Learning*, pages 11080–11090. PMLR, 2021.
- [88] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021.
- [89] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [90] Fanqi Yan, Huy Nguyen, Pedram Akbarian, Nhat Ho, and Alessandro Rinaldo. Sigmoid self-attention is better than softmax self-attention: A mixture-of-experts perspective. *arXiv preprint arXiv:2502.00281*, 2025.
- [91] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS (2019) / arXiv preprint arXiv:1906.08237*, 2019. URL <https://arxiv.org/abs/1906.08237>.
- [92] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL <https://arxiv.org/abs/2210.03629>.
- [93] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019.
- [94] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- [95] F. Zeng, W. Gan, Y. Wang, N. Liu, and P. S. Yu. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*, 2023. URL <https://arxiv.org/abs/2311.07226>.
- [96] Zining Zhang, Bingsheng He, and Zhenjie Zhang. Transmask: A compact and fast speech separation model based on transformer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5764–5768. IEEE, 2021.
- [97] Chuanyang Zheng, Yihang Gao, Guoxuan Chen, Han Shi, Jing Xiong, Xiaozhe Ren, Chao Huang, Zhenguo Li, and Yu Li. Self-adjust softmax. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7827–7847, 2025.
- [98] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 11106–11115, 2021.

---

# Appendix

---

## Contents

|          |                                                                     |           |
|----------|---------------------------------------------------------------------|-----------|
| <b>A</b> | <b>Use of Large Language Models</b>                                 | <b>17</b> |
| <b>B</b> | <b>Related Works</b>                                                | <b>17</b> |
| <b>C</b> | <b>Lemmas used in Proofs</b>                                        | <b>18</b> |
| <b>D</b> | <b>Proof of Theorem 3.1</b>                                         | <b>22</b> |
| <b>E</b> | <b>Proofs for Section 4</b>                                         | <b>24</b> |
| E.1      | Proof of Theorem 4.1 and Corollary 4.2 . . . . .                    | 24        |
| E.2      | Extension to Rotary Positional Embeddings . . . . .                 | 26        |
| <b>F</b> | <b>Approximation and tradeoff results for nonlinear target</b>      | <b>27</b> |
| F.1      | Tradeoff in Theorem 3.1 . . . . .                                   | 27        |
| F.2      | Further Experiments for Section 3.1 . . . . .                       | 28        |
| F.3      | Induction Head . . . . .                                            | 29        |
| F.4      | Continuous Time-homogeneous System . . . . .                        | 29        |
| <b>G</b> | <b>Additional Experiments</b>                                       | <b>34</b> |
| G.1      | Inspiration of Parameter Allocation . . . . .                       | 34        |
| G.2      | Verification of Theorem 4.1 and Corollary 4.2 . . . . .             | 35        |
| G.3      | Role of Layers in Figure 4 . . . . .                                | 35        |
| G.4      | Reducing Budget on Number of Heads . . . . .                        | 36        |
| G.5      | Analysis of Attention Logits Norms . . . . .                        | 37        |
| G.6      | Controlled Ablations for the Two-stage Reduction Pipeline . . . . . | 37        |

## A Use of Large Language Models

In the preparation of this manuscript, we employed a large language model (ChatGPT, developed by OpenAI) as a writing assistant. The model was used exclusively to:

- Improve the clarity, coherence, and conciseness of the text.
- Rephrase sentences to enhance grammatical correctness and readability.
- Ensure consistent terminology and smooth transitions across sections.

All outputs generated by the LLM were carefully reviewed, edited, and validated by the authors to guarantee technical accuracy, logical consistency, and fidelity to the research content. No text was included without human oversight. Importantly, the LLM was not employed for data generation, model training, or experimental design.

## B Related Works

**Approximation ability of Transformers.** The work [93] is the first to establish the universal approximation property of Transformers. Subsequent studies analyzed the efficiency of Transformers

in representing specific functions or tasks, including sparse functions ([24]), targets with nonlinear temporal kernels ([38]), practical computer programs ([28]), long but sparse memories ([82]), induction heads ([69, 70, 68]), and memorization and reasoning ([14]). Moreover, several studies suggest that Transformers achieve in-context learning by effectively approximating gradient-based iterations across layers ([27, 2, 80, 58, 4, 73]). Other work has examined limitations of Transformers' expressivity, particularly in modeling formal languages or simulating circuits ([31, 87, 7, 61, 60]).

**Attention structure.** Prior work has extensively studied the representational power and structural properties of self-attention in Transformers. [49] showed that a fixed self-attention layer can approximate arbitrary sparse patterns with hidden size scaling logarithmically in sequence length. Several works proposed replacing standard attention with linearized variants for improved efficiency ([85, 16]). From a rank perspective, [8] addressed the low-rank bottleneck by setting head dimension proportional to sequence length, while [3] demonstrated that a single full-rank head can implement nearest-neighbor search, whereas exponentially many low-rank heads are required otherwise. Finally, [41] established memorization capacity for one-layer, single-head Transformers, whereas [31] showed that modeling certain formal languages requires the number of layers or heads to grow with input length.

**Property of softmax.** Recent studies have questioned the suitability of softmax attention and analyzed its theoretical limitations. [63] showed that as sequence length increases, softmax attention collapses toward uniformity, token separability saturates, and gradients become unstable, consistent with empirical observations in GPT-2. From a different perspective, [71] argued that the effectiveness of softmax stems mainly from implicit Frobenius norm regularization rather than its probabilistic interpretation, and demonstrated that polynomial activations can achieve comparable performance. Several alternatives have been proposed to address these issues, including SA-Softmax to mitigate gradient vanishing ([97]), sigmoid attention with improved sample efficiency ([90]), and Scalable-Softmax, which preserves sharper attention distributions for long contexts ([64]).

**Model efficiency.** Multi-head attention (MHA) is fundamental to Transformers, yet extensive empirical evidence reveals significant redundancy across attention heads. A large fraction of heads can be pruned with negligible performance loss, and only a small subset consistently captures meaningful patterns across tasks and languages ([62, 79, 57]). Building on this observation, subsequent work proposes principled and automated pruning criteria that substantially improve parameter efficiency while preserving accuracy ([15, 45]). In parallel, efficiency has been further enhanced through architectural and system-level modifications, including sparse and structured attention for long contexts ([6, 94]), shared key-value projections ([1]), and optimized attention kernels ([20, 33]). Despite these advances, a theoretical understanding of when and why multiple attention heads are necessary remains largely open.

## C Lemmas used in Proofs

The first lemma provides a convergence rate for approximating the delta function using exponential sums.

**Lemma C.1.** *For any  $T \in \mathbb{N}_+$ ,  $q, m \in \mathbb{N}_+$ , there exists a  $\phi_m^{\text{exp}}(t) = \sum_{k=1}^m \alpha_k e^{-\beta_k t}$  such that*

$$\|\mathbb{I}(\cdot = T) - \phi_m^{\text{exp}}(\cdot)\|_{\ell_1(\mathbb{N})} \leq \frac{1.3 e^{0.02T}}{m}$$

*where  $\beta_k > 0$  holds for any  $k \in [m]$ , and  $A, C > 0$  are absolute constants.*

*Proof.* This lemma is a corollary of Lemma F.1 in [82]. By Lemma F.1 in [82] and its proof: for any  $T \in \mathbb{N}_+$ ,  $q, m \in \mathbb{N}_+$ , there exists a  $C(q) > 0$  and a  $\phi_m^{\text{exp}}(t) = \sum_{k=1}^m \alpha_k e^{-\beta_k t}$  such that

$$\|\mathbb{I}(\cdot = T) - \phi_m^{\text{exp}}(\cdot)\|_{\ell_1(\mathbb{N})} \leq \frac{C(q) e^{0.01(q+1)T}}{m^q},$$

where  $\beta_k > 0$  holds for any  $k \in [m]$ . Moreover,

$$C(q) = \frac{M(q)}{(1 - 1/e)^q}, \quad M(q) = \max_{0 \leq k \leq q} \sup_{x \in [-1, 1]} \left| \Psi^{(k)}(x) \right|,$$

where  $\Psi(x) = \begin{cases} \exp\left(-\frac{1}{1-x^2}\right), & x \in (-1, 1) \\ 0, & \text{otherwise} \end{cases}$  is the standard bump function on  $[-1, 1]$ . We mainly focus on small  $q$  since  $C(q)$  grows rapidly and through directly computation, we have  $C(1) \approx 1.3$ . Hence,

$$\|\mathbb{I}(\cdot = T) - \phi_m^{\exp}(\cdot)\|_{\ell_1(\mathbb{N})} \leq \frac{1.3 e^{0.02T}}{m}$$

□

The following lemma is for technical usage.

**Lemma C.2** (Tensor-product perturbation). *Let  $u_1, \dots, u_n$  and  $v_1, \dots, v_n$  be vectors satisfying*

$$\|u_j\|_2 \leq B, \quad \|u_j - v_j\|_2 \leq \varepsilon, \quad j \in [n].$$

*Then*

$$\|u_1 \otimes \dots \otimes u_n - v_1 \otimes \dots \otimes v_n\|_2 \leq (B + \varepsilon)^n - B^n.$$

*Proof.* Since  $\|v_j\|_2 \leq B + \varepsilon$ , the telescoping identity gives

$$\begin{aligned} & u_1 \otimes \dots \otimes u_n - v_1 \otimes \dots \otimes v_n \\ &= \sum_{j=1}^n v_1 \otimes \dots \otimes v_{j-1} \otimes (u_j - v_j) \otimes u_{j+1} \otimes \dots \otimes u_n. \end{aligned}$$

Therefore,

$$\begin{aligned} \|u_1 \otimes \dots \otimes u_n - v_1 \otimes \dots \otimes v_n\|_2 &\leq \sum_{j=1}^n (B + \varepsilon)^{j-1} \varepsilon B^{n-j} \\ &= (B + \varepsilon)^n - B^n. \end{aligned}$$

□

The following lemma guarantees that, with high probability, a bounded random vector can be projected onto a  $d$ -dimensional subspace such that the resulting error, after applying any linear transformation  $A$ , is tightly controlled by  $\|A\|_2$ , the projection dimension, and the vector's norm bound.

**Lemma C.3** (Uniform high-probability projection for bounded random vectors). *Let*

$$X_1, \dots, X_N \in \mathbb{R}^D$$

*be random vectors defined on a common probability space. They are not required to be independent, identically distributed, isotropic, or sub-Gaussian. Assume only that*

$$\|X_\ell\|_2 \leq B \quad \text{almost surely for all } \ell \in [N].$$

*Let*

$$E \sim \text{Unif}(G_{D,r})$$

*be a uniformly random  $r$ -dimensional subspace of  $\mathbb{R}^D$ , independent of  $(X_1, \dots, X_N)$ , and let  $\Pi_E$  denote the orthogonal projection onto  $E$ . Equivalently, let  $P \in \mathbb{R}^{r \times D}$  have orthonormal rows spanning  $E$ , so that*

$$PP^\top = I_r, \quad P^\top P = \Pi_E.$$

*Then there exists an absolute constant  $c > 0$  such that, for every  $\delta \in (0, 1)$ , with probability at least  $1 - \delta$  over the joint randomness of  $(X_1, \dots, X_N, E)$ , simultaneously for all  $\ell \in [N]$ ,*

$$\|(I - \Pi_E)X_\ell\|_2 \leq B \left( \sqrt{1 - \frac{r}{D}} + \sqrt{\frac{\log(2N/\delta)}{cD}} \right).$$

More strongly, conditional on  $(X_1, \dots, X_N)$ , the above event has probability at least  $1 - \delta$  with respect to the randomness of  $E$  almost surely.

Moreover, for every  $\ell \in [N]$ ,

$$\mathbb{E}_E [\|(I - \Pi_E)X_\ell\|_2^2 | X_\ell] = \left(1 - \frac{r}{D}\right) \|X_\ell\|_2^2 \quad \text{almost surely,}$$

and consequently

$$\mathbb{E}_{X_\ell, E} \|(I - \Pi_E)X_\ell\|_2^2 = \left(1 - \frac{r}{D}\right) \mathbb{E} \|X_\ell\|_2^2.$$

When  $r = D$ , the projection error is identically zero.

*Proof.* Let

$$Q_E := I - \Pi_E,$$

which is the orthogonal projection onto the  $(D - r)$ -dimensional subspace  $E^\perp$ .

We first establish the concentration inequality for a fixed deterministic vector  $x \in \mathbb{R}^D$ . If  $x = 0$ , the result is trivial. Hence suppose  $x \neq 0$  and write

$$u = \frac{x}{\|x\|_2} \in \mathbb{S}^{D-1}.$$

By rotational invariance of the uniform measure on the Grassmannian, the random variable

$$\|Q_E u\|_2$$

has the same distribution as

$$\|Q_0 U\|_2,$$

where  $U$  is uniformly distributed on the unit sphere  $\mathbb{S}^{D-1}$  and  $Q_0$  is any fixed orthogonal projection of rank  $D - r$ .

Consider the function

$$f(u) := \|Q_0 u\|_2, \quad u \in \mathbb{S}^{D-1}.$$

Since  $Q_0$  is an orthogonal projection,

$$|f(u) - f(v)| \leq \|Q_0(u - v)\|_2 \leq \|u - v\|_2,$$

so  $f$  is 1-Lipschitz.

By the standard concentration inequality for Lipschitz functions on the sphere, there exists an absolute constant  $c > 0$  such that, for every  $\varepsilon > 0$ ,

$$\Pr(f(U) \geq \mathbb{E}f(U) + \varepsilon) \leq 2 \exp(-cD\varepsilon^2).$$

On the other hand, Jensen's inequality gives

$$\mathbb{E}f(U) \leq \sqrt{\mathbb{E}f(U)^2}.$$

Using

$$\mathbb{E}[UU^\top] = \frac{1}{D} I_D,$$

we obtain

$$\mathbb{E}f(U)^2 = \mathbb{E}\|Q_0 U\|_2^2 \leq \text{Tr}(Q_0 \mathbb{E}[UU^\top]) = \frac{\text{Tr}(Q_0)}{D} = 1 - \frac{r}{D}.$$

Therefore,

$$\mathbb{E}f(U) \leq \sqrt{1 - \frac{r}{D}}.$$

It follows that

$$\Pr_E \left[ \|Q_E u\|_2 \geq \sqrt{1 - \frac{r}{D}} + \varepsilon \right] \leq 2 \exp(-cD\varepsilon^2).$$

Multiplying by  $\|x\|_2$  yields

$$\Pr_E \left[ \|Q_E x\|_2 \geq \|x\|_2 \left( \sqrt{1 - \frac{r}{D}} + \varepsilon \right) \right] \leq 2 \exp(-cD\varepsilon^2). \quad (1)$$

We now return to the random vectors  $X_1, \dots, X_N$ . Let

$$\mathcal{F}_X := \sigma(X_1, \dots, X_N).$$

Since  $E$  is independent of  $\mathcal{F}_X$ , conditional on  $\mathcal{F}_X$  the subspace  $E$  is still uniformly distributed over  $G_{D,r}$ .

Fix an arbitrary realization

$$X_1 = x_1, \dots, X_N = x_N$$

satisfying

$$\|x_\ell\|_2 \leq B, \quad \ell \in [N].$$

Applying (1) to each  $x_\ell$  gives

$$\Pr_E \left[ \|Q_E x_\ell\|_2 \geq B \left( \sqrt{1 - \frac{r}{D}} + \varepsilon \right) \right] \leq 2 \exp(-cD\varepsilon^2),$$

where we used  $\|x_\ell\|_2 \leq B$ .

Hence, by a union bound over  $\ell \in [N]$ ,

$$\begin{aligned} \Pr_E \left[ \max_{\ell \in [N]} \|Q_E x_\ell\|_2 \geq B \left( \sqrt{1 - \frac{r}{D}} + \varepsilon \right) \middle| X_1 = x_1, \dots, X_N = x_N \right] \\ \leq 2N \exp(-cD\varepsilon^2). \end{aligned}$$

Since this bound holds for every realization satisfying the almost sure boundedness assumption, we have almost surely

$$\Pr_E \left[ \max_{\ell \in [N]} \|Q_E X_\ell\|_2 \geq B \left( \sqrt{1 - \frac{r}{D}} + \varepsilon \right) \middle| \mathcal{F}_X \right] \leq 2N \exp(-cD\varepsilon^2).$$

Taking expectation with respect to  $\mathcal{F}_X$  gives

$$\Pr_{X,E} \left[ \max_{\ell \in [N]} \|Q_E X_\ell\|_2 \geq B \left( \sqrt{1 - \frac{r}{D}} + \varepsilon \right) \right] \leq 2N \exp(-cD\varepsilon^2).$$

Choosing

$$\varepsilon = \sqrt{\frac{\log(2N/\delta)}{cD}}$$

proves the uniform high-probability bound.

For the expectation identity, rotational invariance of the uniform distribution on the Grassmannian implies

$$\mathbb{E}_E[\Pi_E] = \frac{r}{D} I_D.$$

Consequently,

$$\mathbb{E}_E[Q_E] = \left(1 - \frac{r}{D}\right) I_D.$$

Conditioning on  $X_\ell$ , we obtain

$$\begin{aligned} \mathbb{E}_E \left[ \|Q_E X_\ell\|_2^2 \middle| X_\ell \right] &= X_\ell^\top \mathbb{E}_E[Q_E^\top Q_E] X_\ell \\ &= X_\ell^\top \mathbb{E}_E[Q_E] X_\ell \\ &= \left(1 - \frac{r}{D}\right) \|X_\ell\|_2^2, \end{aligned}$$

where we used  $Q_E^2 = Q_E$ . Taking expectation over  $X_\ell$  yields

$$\mathbb{E}_{X_\ell, E} \|Q_E X_\ell\|_2^2 = \left(1 - \frac{r}{D}\right) \mathbb{E} \|X_\ell\|_2^2.$$

Finally, when  $r = D$ , we have  $E = \mathbb{R}^D$  and hence  $Q_E = 0$ , so the projection error is identically zero.  $\square$

The following lemma is Corollary 1 in [63].

**Lemma C.4.** Assume  $\|\text{Logits}_{d_H}(\mathbf{W}_Q, \mathbf{W}_K, \mathbf{X})\|_\infty$  is bounded, the attention weights satisfy:

$$\frac{1}{L} \exp(-2\Delta) \leq \|\text{AttnScore}_{d_H}(\mathbf{W}_Q, \mathbf{W}_K, \mathbf{X})\|_\infty \leq \frac{1}{L} \exp(2\Delta),$$

where  $\Delta = \|\mathbf{W}_Q\|_2 \|\mathbf{W}_K\|_2 B^2$ .

The Gershgorin circle theorem localizes all eigenvalues of a matrix within discs determined by its diagonal entries and row sums of off-diagonal magnitudes.

**Lemma C.5** (Gershgorin circle theorem). Let  $A = (a_{ij}) \in \mathbb{C}^{n \times n}$ . For each  $i$  define the Gershgorin radius

$$R_i := \sum_{j \neq i} |a_{ij}|.$$

Then every eigenvalue  $\lambda$  of  $A$  lies in the union of the Gershgorin discs

$$\lambda \in \bigcup_{i=1}^n \{z \in \mathbb{C} : |z - a_{ii}| \leq R_i\}.$$

In particular, if the discs  $\{z : |z - a_{ii}| \leq R_i\}$  are contained in a region  $S \subset \mathbb{C}$ , then all eigenvalues of  $A$  lie in  $S$ .

The mean value theorem expresses the difference  $F(\mathbf{y}) - F(\mathbf{x})$  as the Jacobian at an intermediate point applied to  $\mathbf{y} - \mathbf{x}$ .

**Lemma C.6** (Mean Value Inequality for vector-valued functions). Let  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  be continuously differentiable on an open convex set  $D \subset \mathbb{R}^n$ . Then for any  $\mathbf{x}, \mathbf{y} \in D$ , there exists  $\theta \in (0, 1)$  such that

$$\|F(\mathbf{y}) - F(\mathbf{x})\| \leq \|JF(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x}))\| \cdot \|\mathbf{y} - \mathbf{x}\|,$$

where  $JF(\cdot)$  denotes the Jacobian matrix of  $F$  and  $\|\cdot\|$  is an appropriate vector/matrix norm.

**Lemma C.7** (Eckart–Young–Mirsky Theorem). Let  $\mathbf{A} \in \mathbb{R}^{m \times n}$  have singular value decomposition  $\mathbf{A} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ , where  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ . For any  $k < r$ , the rank- $k$  truncated SVD

$$\mathbf{A}_k := \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

is an optimal rank- $k$  approximation of  $\mathbf{A}$  in both the spectral norm and the Frobenius norm, i.e.,

$$\mathbf{A}_k \in \arg \min_{\text{rank}(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_2 \quad \text{and} \quad \mathbf{A}_k \in \arg \min_{\text{rank}(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_F.$$

Moreover, the optimal errors satisfy

$$\|\mathbf{A} - \mathbf{A}_k\|_2 = \sigma_{k+1}, \quad \|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i=k+1}^r \sigma_i^2.$$

## D Proof of Theorem 3.1

In this section, we give the detailed proof of Theorem 3.1. We first restate the theorem:

**Theorem D.1** (Restate of Theorem 3.1.). To approximate the linear target  $\mathbf{H}_t(\mathbf{X}) = \sum_{i=0}^L \boldsymbol{\rho}_i \mathbf{x}_{t-i}$ , suppose we have  $H_m$  heads each of the dimension  $d_m$  ( $m=1, \dots, M$ ), and the model dimension  $D = \sum_{m=1}^M H_m \cdot d_m$  is fixed. Then, with probability at least  $1 - \delta$ , we have

$$\begin{aligned} \mathcal{E}_D(X) \leq & B \sum_{m=1}^M \|\rho_m\|_2 \left[ \mathbb{I}_{\{d_m < d\}} \left( \sqrt{1 - \frac{d_m}{d}} \right) + \frac{1.3e^{0.02m}}{H_m} \right] + B \sum_{k=M+1}^L \|\rho_k\|_2 \\ & + B \sum_{m=1}^M \|\rho_m\|_2 \cdot C \sqrt{\frac{\log(2M/\delta)}{d}}. \end{aligned} \tag{5}$$

*Proof.* We only consider the case  $d_m \leq d$  in our proof, since otherwise there is no information loss and the first term in the approximation error vanishes. Since we only use a single layer of the Transformer, we omit the layer index.

**Embedding layer.** We embed each token by

$$\mathbf{W}_E = \begin{pmatrix} \mathbf{I}_d \\ \mathbf{0}_{(D-d) \times d} \end{pmatrix} \in \mathbb{R}^{D \times d}, \quad \mathbf{b}_E = \mathbf{0} \in \mathbb{R}^D,$$

for  $\mathbf{x}_t$ , then each token  $\mathbf{x}_t^{(0)}$  after embedding is

$$\mathbf{x}_t^{(0)} = \mathbf{W}_E \mathbf{x}_t + \mathbf{b}_E = \begin{pmatrix} \mathbf{x}_t \\ \mathbf{0} \end{pmatrix} \in \mathbb{R}^D.$$

**Multi-head attention layer with residual connection.** We implement multi-head attention using  $M$  groups of heads. Each group  $m$  comprises  $H_m$  heads with dimension  $d_m$  for  $(m = 1, 2, \dots, M)$ , aimed at extracting the corresponding representation  $\tilde{\mathbf{x}}_{t-m}$  which is  $\mathbf{x}_{t-m}$  after projection to  $d_m$ -dimension space. This step can be presented as follows for each token:

$$\begin{pmatrix} \mathbf{x}_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} \tilde{\mathbf{x}}_t \\ \tilde{\mathbf{x}}_{t-1} \\ \vdots \\ \tilde{\mathbf{x}}_{t-M} \\ \mathbf{0}_{D-(\sum_{m=1}^M d_m)} \end{pmatrix} := \mathbf{x}_t^{(1/2)}.$$

where  $\tilde{\mathbf{x}}_t$  is obtained by residual connection.

We now focus on the extraction of an arbitrary  $\tilde{\mathbf{x}}_{t-m}$  and give the details of construction.

By lemma C.1, for any rate  $q \in \mathbb{N}^+$ , there exists a function

$$\phi_m^{\text{exp}}(t) = \sum_{h=1}^{H_m} \alpha_{h,m} e^{-\beta_{h,m}(t-1)}$$

such that  $\beta_h > 0$  and

$$\|\mathbb{I}(\cdot = m) - \phi_m^{\text{exp}}(\cdot)\|_{\ell_1(\mathbb{N})} = \sum_{s=0}^{+\infty} |\mathbb{I}(s = m) - \phi^{\text{exp}}(s)| \leq \frac{1.3 e^{0.02m}}{H_m}.$$

By Lemma C.3, there exist an projection  $\mathbf{P}_m \in \mathbb{R}^{d_m \times d}$  such that

$$\|(I - \Pi_E)x_\ell\|_2 \leq B \sqrt{1 - \frac{d_m}{D}} + B \cdot \sqrt{\frac{\log(2N/\delta)}{cD}}$$

for any  $l = 1, \dots, m$  with probability at least  $1 - \delta$ . Let  $\tilde{\mathbf{x}}_{t-m} := \mathbf{P}_m \mathbf{x}_{t-m}$ . For  $h = \sum_{i=1}^{m-1} H_i, \sum_{j=1}^{m-1} H_j + 1, \dots, \sum_{k=1}^m H_k$ , we choose parameters as follows

$$p^{(h)} = \beta_{h,m}, \quad \mathbf{W}_V^{(h)} = \alpha_{h,m} \left( \sum_{i=0}^{H_m} \exp(-\beta_{h,m}(i-1)) \right) \mathbf{P}_m, \quad \mathbf{W}_K^{(h)} = \mathbf{W}_Q^{(h)} = 0.$$

The output of  $H$  heads are concatenated together, and we take

$$\mathbf{W}_O = \sum_{m=1}^M \sum_{h=\sum_{i=1}^{m-1} H_i + 1}^{\sum_{j=1}^m H_j} \mathbf{S}_{m,h},$$

where  $\mathbf{S}_{m,h} \in \mathbb{R}^{D \times D}$  moves the output of the  $m$ -th set of heads to the same position

$$\mathbf{S}_{m,h} = \begin{pmatrix} \mathbf{0}_{(\sum_{i=1}^{m-1} d_i) \times (\sum_{i=1}^{m-1} d_i + (h-1)d_m)} & \mathbf{0}_{(\sum_{i=1}^{m-1} d_i) \times d_m} & \mathbf{0}_{(\sum_{i=1}^{m-1} d_i) \times (D - \sum_{i=1}^{m-1} d_i - h d_m)} \\ \mathbf{0}_{d_m \times (\sum_{i=1}^{m-1} d_i + (h-1)d_m)} & \mathbf{I}_{d_m \times d_m} & \mathbf{0}_{d_m \times (D - \sum_{i=1}^{m-1} d_i - h d_m)} \\ \mathbf{0}_{(D - \sum_{i=1}^m d_m) \times (\sum_{i=1}^{m-1} d_i + (h-1)d_m)} & \mathbf{0}_{(D - \sum_{i=1}^m d_m) \times d_m} & \mathbf{0}_{(D - \sum_{i=1}^m d_m) \times (D - \sum_{i=1}^{m-1} d_i - h d_m)} \end{pmatrix}$$

We define

$$\tilde{\mathbf{P}}_m := (0_{d_H \times (m-1)d_H}, I_{d_m \times d_m}, 0_{d_H \times (D-md_H)}) \in \mathbb{R}^{d_m \times D}, \quad m \in [M],$$

then

$$\tilde{\mathbf{P}}_m \cdot \text{Concat} \left( \text{Attn}^{(h)}(\mathbf{X}^{(0)}) \right)_{h=1}^H = \sum_{h=\sum_{i=1}^{m-1} H_i}^{\sum_{j=1}^m H_j} \alpha_{h,m} \sum_{s=1}^{\infty} e^{-\beta_{h,m}(s-1)} \tilde{\mathbf{x}}_{t-s} := \hat{\mathbf{x}}_{t-m}.$$

And the approximation error of this step is

$$\begin{aligned} \varepsilon_{\text{Attn}} &\leq \sup_t \sum_{m=1}^M \|\tilde{\mathbf{x}}_{t-m} - \hat{\mathbf{x}}_{t-m}\|_{\infty} \\ &\leq \sup_t \sum_{m=1}^M \|\mathbb{I}(\cdot = m) - \phi_m^{\text{exp}}(\cdot)\|_{\ell_1(\mathbb{N})} \cdot B \\ &\leq B \cdot \sum_{m=1}^M \frac{1.3 e^{0.02m}}{H_m}, \end{aligned}$$

**Feed forward network.** FFN is implemented component-wise to realize convolution. The final output for the  $t$ -th token is

$$\hat{H}_t(\mathbf{X}) = \sum_{m=0}^M \tilde{\rho}_m \hat{\mathbf{x}}_{t-m},$$

where  $\tilde{\rho}_m = \rho_m \mathbf{P}_m^{\top}$ . In fact we only need a linear map to achieve this, and FFN won't cause any approximation error.

**Approximation error.** By Lemma C.3, with probability at least  $1 - \delta$ , we have

$$\begin{aligned} \mathcal{E}_D(\mathbf{X}) &= \left\| \sum_{m=0}^{\infty} \rho_m \mathbf{x}_{t-m} - \sum_{m=0}^M \tilde{\rho}_m \hat{\mathbf{x}}_{t-m} \right\|_2 \\ &\leq \sum_{m=1}^M \|\rho_m \mathbf{x}_{t-m} - \tilde{\rho}_m \hat{\mathbf{x}}_{t-m}\|_2 + B \sum_{k=M+1}^{\infty} \|\rho_k\|_2 \\ &\leq \sum_{m=1}^M \|\rho_m \mathbf{x}_{t-m} - \tilde{\rho}_m \tilde{\mathbf{x}}_{t-m}\|_2 + \sum_{m=1}^M \|\tilde{\rho}_m \tilde{\mathbf{x}}_{t-m} - \tilde{\rho}_m \hat{\mathbf{x}}_{t-m}\|_2 + B \sum_{k=M+1}^{\infty} \|\rho_k\|_2 \\ (\text{By Lemma C.3}) &\leq B \sum_{m=1}^M \|\rho_m\|_2 \left[ \mathbb{I}_{\{d_m < d\}} \left( \sqrt{1 - \frac{d_m}{d}} \right) + \frac{1.3e^{0.02m}}{H_m} \right] + B \sum_{k=M+1}^L \|\rho_k\|_2 \\ &\quad + B \sum_{m=1}^M \|\rho_m\|_2 \cdot C \sqrt{\frac{\log(2M/\delta)}{d}}. \end{aligned}$$

□

## E Proofs for Section 4

### E.1 Proof of Theorem 4.1 and Corollary 4.2

In this section, we prove Theorem 4.1 and Corollary 4.2, we first restate the results:

**Theorem E.1** (Restate of Theorem 4.1). Assume  $\|W_Q\|_2 \leq M_Q$ ,  $\|W_K\|_2 \leq M_K$ ,  $\|x_t\|_2 \leq B$  ( $t \in [L]$ ), and that the finite positional-bias entries satisfy  $|R_{lj}| \leq R_0$  ( $j \in \mathcal{S}_l$ ,  $l \in [L]$ ). Let  $\Delta := M_Q M_K B^2 + R_0$ . Then, for every  $l \in [L]$ ,

$$\|J_l\|_2 \leq \frac{2e^{2\Delta}}{s_l}.$$

Consequently,

$$\|J_l\|_2 = \mathcal{O}(L^{-1})$$

for unmasked attention, while under the standard causal mask,  $\|J_l\|_2 = \mathcal{O}(l^{-1})$ . The hidden constants depend only on  $M_Q, M_K, B$ , and  $R_0$ .

**Corollary E.2** (Restate of Theorem 4.2). Under the assumptions of Theorem 4.1, let  $M := W_Q^\top W_K$ , and let  $M_{d_h}$  be its rank- $d_h$  truncated SVD. Choose  $\hat{W}_Q, \hat{W}_K \in \mathbb{R}^{d_h \times D}$  such that  $\hat{W}_Q^\top \hat{W}_K = M_{d_h}$ . Let  $\hat{A}$  denote the attention-score matrix obtained by replacing  $W_Q, W_K$  with  $\hat{W}_Q, \hat{W}_K$  while keeping the same positional bias and mask. Then for every query  $l \in [L]$ ,

$$\left\| e_l^\top (\text{AttnScore}_{d_h} - \hat{A}) \right\|_2 \leq \frac{2e^{2\Delta}}{s_l} \|X\|_2^2 \sigma_{d_h+1}(M).$$

Hence the per-query approximation error is

$$\mathcal{O}\left(\frac{\|X\|_2^2 \sigma_{d_h+1}(M)}{L}\right)$$

in the unmasked case and

$$\mathcal{O}\left(\frac{\|X\|_2^2 \sigma_{d_h+1}(M)}{l}\right)$$

under the standard causal mask.

In particular, for any fixed  $\alpha \in (0, 1)$ , the late-query block  $\mathcal{I}_\alpha := \{l : l \geq \lceil \alpha L \rceil\}$  satisfies

$$\left( \sum_{l \in \mathcal{I}_\alpha} \left\| e_l^\top (\text{AttnScore}_{d_h} - \hat{A}) \right\|_2^2 \right)^{1/2} \leq \frac{2e^{2\Delta}}{\alpha \sqrt{L}} \|X\|_2^2 \sigma_{d_h+1}(M).$$

*Proof of Theorem 4.1.* Fix  $l \in [L]$  and write  $p = a_l \in \mathbb{R}^{s_l}$ . For every active coordinate  $j \in \mathcal{S}_l$ ,

$$|z_{lj}| \leq \|W_Q\|_2 \|W_K\|_2 \|x_l\|_2 \|x_j\|_2 + |R_{lj}| \leq \Delta.$$

Therefore

$$p_j = \frac{e^{z_{lj}}}{\sum_{k \in \mathcal{S}_l} e^{z_{lk}}} \leq \frac{e^\Delta}{s_l e^{-\Delta}} = \frac{e^{2\Delta}}{s_l}.$$

The softmax Jacobian is

$$J_l = \text{diag}(p) - pp^\top,$$

which is positive semidefinite. By the Gershgorin circle theorem (Lemma C.5), every eigenvalue  $\lambda$  of  $J_l$  satisfies

$$0 \leq \lambda \leq 2 \max_j p_j (1 - p_j) \leq 2 \max_j p_j \leq \frac{2e^{2\Delta}}{s_l}.$$

Since  $J_l$  is symmetric positive semidefinite,

$$\|J_l\|_2 = \lambda_{\max}(J_l) \leq \frac{2e^{2\Delta}}{s_l}.$$

Substituting  $s_l = L$  for unmasked attention and  $s_l = l$  for the standard causal mask proves the result.  $\square$

*Proof of Corollary 4.2.* Let  $M_{d_h}$  be the rank- $d_h$  truncated SVD of  $M = W_Q^\top W_K$ . By the Eckart–Young–Mirsky theorem (Lemma C.7),

$$\|M - M_{d_h}\|_2 = \sigma_{d_h+1}(M), \quad \|M_{d_h}\|_2 \leq \|M\|_2.$$

Writing  $M_{d_h} = U\Sigma V^\top$ , one may take

$$\widehat{W}_Q = \Sigma^{1/2}U^\top, \quad \widehat{W}_K = \Sigma^{1/2}V^\top,$$

so that  $\widehat{W}_Q^\top \widehat{W}_K = M_{d_h}$ .

For query  $l$ , let  $z_l$  and  $\hat{z}_l$  denote the original and compressed active logit vectors. Their difference obeys

$$\begin{aligned} \|z_l - \hat{z}_l\|_2 &\leq \|X^\top(M - M_{d_h})X\|_2 \\ &\leq \|X\|_2^2 \sigma_{d_h+1}(M). \end{aligned}$$

Moreover, both endpoints and every point on the line segment between  $z_l$  and  $\hat{z}_l$  have active coordinates bounded in absolute value by  $\Delta$ , because  $\|M_{d_h}\|_2 \leq \|M\|_2$ . Hence Theorem 4.1 applies uniformly along this segment. Using the integral form of the mean-value inequality,

$$\begin{aligned} \|\text{softmax}(z_l) - \text{softmax}(\hat{z}_l)\|_2 &\leq \sup_{t \in [0,1]} \|J_l((1-t)z_l + t\hat{z}_l)\|_2 \|z_l - \hat{z}_l\|_2 \\ &\leq \frac{2e^{2\Delta}}{s_l} \|X\|_2^2 \sigma_{d_h+1}(M). \end{aligned}$$

This proves the per-query bound. For  $\mathcal{I}_\alpha = \{l : l \geq \lceil \alpha L \rceil\}$  under causal masking,  $s_l = l \geq \alpha L$ , and therefore

$$\begin{aligned} &\left( \sum_{l \in \mathcal{I}_\alpha} \|e_l^\top (\text{AttnScore}_{d_H} - \hat{A})\|_2^2 \right)^{1/2} \\ &\leq \sqrt{L} \frac{2e^{2\Delta}}{\alpha L} \|X\|_2^2 \sigma_{d_h+1}(M), \end{aligned}$$

which gives the stated  $L^{-1/2}$  late-query bound.  $\square$

## E.2 Extension to Rotary Positional Embeddings

We next show that the head-dimension reduction argument extends to rotary positional embeddings (RoPE). Unlike the additive positional bias considered in Theorem 4.1, RoPE rotates the query and key representations before their inner product.

For position  $t \in [L]$ , let

$$\mathcal{R}_t \in \mathbb{R}^{d_H \times d_H}$$

be the corresponding RoPE rotation matrix. By construction,

$$\mathcal{R}_t^\top \mathcal{R}_t = I_{d_H}.$$

For a query position  $l$  and a key position  $j$ , define the relative rotation

$$\mathcal{R}_{l,j} := \mathcal{R}_l^\top \mathcal{R}_j.$$

Therefore,

$$\mathcal{R}_{l,j}^\top \mathcal{R}_{l,j} = I_{d_H}, \quad \|\mathcal{R}_{l,j}\|_2 = 1.$$

The RoPE attention logit between positions  $l$  and  $j$  is

$$\begin{aligned} z_{lj} &= (\mathcal{R}_l W_Q x_l)^\top (\mathcal{R}_j W_K x_j) \\ &= x_l^\top W_Q^\top \mathcal{R}_l^\top \mathcal{R}_j W_K x_j \\ &= x_l^\top W_Q^\top \mathcal{R}_{l,j} W_K x_j. \end{aligned}$$

Thus, for each relative position, the effective interaction matrix is

$$B_{l,j} := W_Q^\top \mathcal{R}_{l,j} W_K \in \mathbb{R}^{D \times D}.$$

**Proposition E.3** (Head-dimension reduction under RoPE). *Let  $W_Q, W_K \in \mathbb{R}^{d_H \times D}$ , and let  $W_{Q,r}$  and  $W_{K,r}$  be their rank- $r$  truncated-SVD approximations, where  $r < d_H$ . Define*

$$\tilde{B}_{l,j} := W_{Q,r}^\top \mathcal{R}_{l,j} W_{K,r}.$$

*Then  $\text{rank}(\tilde{B}_{l,j}) \leq r$ , and uniformly over all  $l, j \in [L]$ ,*

$$\|B_{l,j} - \tilde{B}_{l,j}\|_2 \leq \|W_Q\|_2 \sigma_{r+1}(W_K) + \sigma_{r+1}(W_Q) \|W_K\|_2.$$

*Consequently, if  $\|x_t\|_2 \leq B$  for all  $t \in [L]$ , then*

$$|z_{lj} - \tilde{z}_{lj}| \leq B^2 [\|W_Q\|_2 \sigma_{r+1}(W_K) + \sigma_{r+1}(W_Q) \|W_K\|_2].$$

*Proof.* Write

$$W_Q = W_{Q,r} + E_Q, \quad W_K = W_{K,r} + E_K.$$

By the Eckart–Young–Mirsky theorem,

$$\|E_Q\|_2 = \sigma_{r+1}(W_Q), \quad \|E_K\|_2 = \sigma_{r+1}(W_K),$$

and

$$\|W_{Q,r}\|_2 = \|W_Q\|_2, \quad \|W_{K,r}\|_2 = \|W_K\|_2.$$

For any pair  $(l, j)$ , we have

$$\begin{aligned} B_{l,j} - \tilde{B}_{l,j} &= W_Q^\top \mathcal{R}_{l,j} W_K - W_{Q,r}^\top \mathcal{R}_{l,j} W_{K,r} \\ &= W_{Q,r}^\top \mathcal{R}_{l,j} E_K + E_Q^\top \mathcal{R}_{l,j} W_K. \end{aligned}$$

Hence,

$$\|B_{l,j} - \tilde{B}_{l,j}\|_2 \leq \|W_{Q,r}\|_2 \|\mathcal{R}_{l,j}\|_2 \|E_K\|_2 + \|E_Q\|_2 \|\mathcal{R}_{l,j}\|_2 \|W_K\|_2.$$

Since  $\mathcal{R}_{l,j}$  is orthogonal,

$$\|\mathcal{R}_{l,j}\|_2 = 1.$$

Therefore,

$$\|B_{l,j} - \tilde{B}_{l,j}\|_2 \leq \|W_Q\|_2 \sigma_{r+1}(W_K) + \sigma_{r+1}(W_Q) \|W_K\|_2.$$

Moreover,

$$|z_{lj} - \tilde{z}_{lj}| = \left| x_l^\top (B_{l,j} - \tilde{B}_{l,j}) x_j \right| \leq \|x_l\|_2 \|B_{l,j} - \tilde{B}_{l,j}\|_2 \|x_j\|_2,$$

which, together with  $\|x_t\|_2 \leq B$ , gives the stated logit approximation bound.  $\square$

## F Approximation and tradeoff results for nonlinear target

### F.1 Tradeoff in Theorem 3.1

We numerically validate the parameter trade-off in the approximation error estimate (i.e. the first equation of (2)). Assume that all  $d_m$ 's are equal and there is no truncation error, we take  $\rho = 1$ ,  $d = 16$  and  $D = 128$ . Then

$$\mathcal{E}_{128} = \sqrt{1 - \frac{8}{H}} + \frac{1.3 e^{0.02}}{H}$$

without concentration error. We plot  $\mathcal{E}(128)$  in Figure 7, which clearly shows a parameter trade-off trend. The trend of the continuous curve closely resembles that of the discrete polyline in Figure 1.

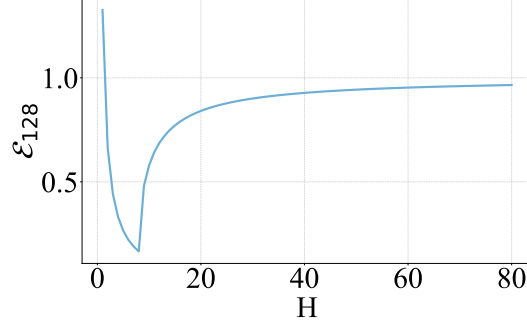


Figure 7: Induction Head Tradeoff

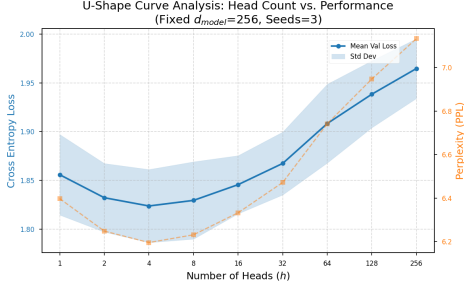


Figure 8: Next token prediction  $D = 256$ .

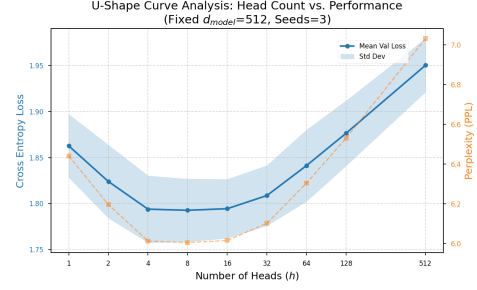


Figure 9: Next token prediction  $D = 512$ .

## F.2 Further Experiments for Section 3.1

We first give out the empirical results for other values of  $D$  for Figure 1 in Figure 8, 9, 10 and 11.

To investigate the architectural scaling properties of multi-head attention (MHA), we conducted a controlled experiment in Figure 12 to determine the optimal balance between the number of attention heads ( $H$ ) and the head dimension ( $d_h$ ) for a fixed model capacity. We configured a series of Transformer-based models with approximately 330 million parameters, maintaining a constant hidden dimension ( $d_{model} = 1024$ ) and depth ( $L = 12$ ). We varied the head count  $h \in \{4, 8, 16, 32, 64, 128\}$ , which inversely corresponds to head dimensions  $d_h \in \{256, 128, 64, 32, 16, 8\}$ . The models were trained on a high-quality subset of the Stanford Alpaca dataset consisting of 15,000 samples, specifically filtered to emphasize logic and code-related reasoning. Each configuration was trained for 1,000 steps using a cosine learning rate scheduler (peak learning rate  $2.5 \times 10^{-4}$ ) and a global batch size of 16. To ensure statistical robustness, we averaged the results across two distinct random seeds ( $S \in \{42, 2026\}$ ). Generalization performance was measured via cross-entropy loss on a held-out validation set. This experimental setup allows us to identify the "Goldilocks" zone for attention granularity, where the model maximizes its representational capacity without suffering from the noise associated with excessively small head dimensions.

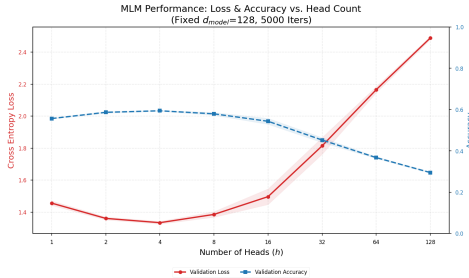


Figure 10: Masked language modeling  $D = 128$ .

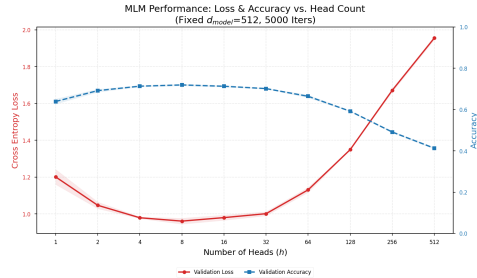


Figure 11: Masked language modeling  $D = 512$ .

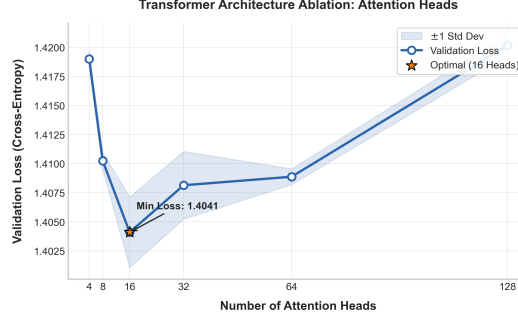


Figure 12: Induction Head Tradeoff

### F.3 Induction Head

The former work [81] defined a generalized induction head:

$$\text{IH}_{n:t}(\mathbf{X}) = \sum_{s=n}^{t-1} \pi_s \mathbf{x}_s,$$

where

$$\pi_s = \text{softmax} \left( \left( \mathbf{X}_{t-n+2:t} \mathbf{W}^* \mathbf{X}_{v-n+1:v-1} \right)_{v=n}^{L-1} \right)_{v=s},$$

and empirically pointed out that the first Transformer layer is used to extract information. In fact, the parameter allocation problem for the first layer reduces to the  $n$ -gram case.

We systematically investigated the effect of Transformer architectural variants on the Induction Head task in Figure 13, which tests a model’s ability to recognize and extend repeated patterns—a fundamental reasoning mechanism for in-context learning. The task was constructed using synthetic sequences of length 128 with 10 unique tokens and embedded 5-token repeating patterns, yielding a dataset of 10,000 sequences (80%/20% train/validation).

The model architecture was a two-layer Transformer encoder with  $d_{\text{model}} = 256$ , pre-norm configuration, GELU feed-forward networks, learned positional encodings, and a two-layer MLP output head. We varied the number of attention heads in 1, 2, 4, 8, 16, 32 while keeping  $d_{\text{model}}$  fixed. Models were trained for 500 epochs using AdamW (learning rate  $\in 10^{-3}, 5 \times 10^{-4}, 10^{-4}$ , weight decay  $10^{-4}$ ), cosine scheduling with ReduceLROnPlateau, gradient clipping (1.0), and mixed precision. Each configuration was repeated with three random seeds, and the best-performing model was selected by validation mean squared error (MSE).

Evaluation included validation MSE, heatmaps of performance across architectural variants, and qualitative prediction analysis. The results reveal how the number of heads influences the ability of the model to capture the Induction Head mechanism and demonstrate the robustness of certain configurations across seeds. This provides practical guidance for architectural design in tasks requiring in-context learning.

### F.4 Continuous Time-homogeneous System

In this subsection, we analyze the approximation error of a continuous target function and derive the resulting trade-off.

We consider a sequence-to-sequence model in which both the input  $\mathbf{X} = (\mathbf{x}_t)_{t \in \mathbb{Z}} \subset \mathcal{X} \subset \mathbb{R}^{d \times \mathbb{Z}}$  and output  $\mathbf{Y} = (\mathbf{y}_t)_{t \in \mathbb{Z}} \subset \mathcal{Y} \subset \mathbb{R}^{d \times \mathbb{Z}}$  are sequences of infinite length and input tokens  $\mathbf{x}_t$  are i.i.d. (we need a stronger assumption due to the target) random vectors with zero mean and identity covariance matrix and whose norm is bounded by  $B$ .  $H$  denotes the mapping from  $\mathcal{X}$  to  $\mathcal{Y}$  with  $\mathbf{y}_t = H_t(\mathbf{X})$ .

To give the mapping  $H$  practical relevance, we must impose certain properties on the target function space. We now present the definitions of these properties.

**Definition F.1.**  $H = \{H_t : \mathcal{X} \mapsto \mathbb{R}^d; t \in \mathbb{R}\}$  be a sequence of functionals.

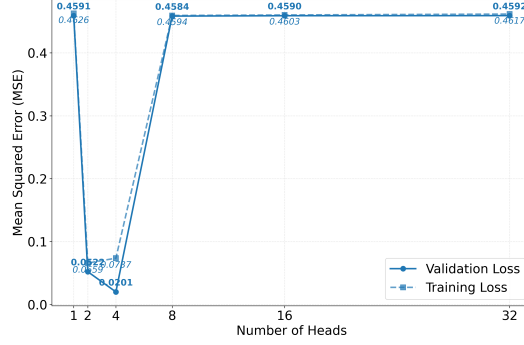


Figure 13: Induction Head Tradeoff

1. **(Linear)**  $H_t$  is linear if for any  $\lambda, \lambda' \in \mathbb{R}$  and  $\mathbf{X}, \mathbf{X}' \in \mathcal{X}$ ,  $H_t(\lambda\mathbf{X} + \lambda'\mathbf{X}') = \lambda H_t(\mathbf{X}) + \lambda' H_t(\mathbf{X}')$ .
2. **(Continuous)**  $H_t$  is continuous if for any  $\mathbf{X}, \mathbf{X}' \in \mathcal{X}$ ,  $\lim_{\mathbf{X}' \rightarrow \mathbf{X}} |H_t(\mathbf{X}') - H_t(\mathbf{X})| = 0$ .
3. **(Bounded)**  $H_t$  is bounded if  $\sup_{\mathbf{X} \in \mathcal{X}, \mathbf{X} \neq \mathbf{0}} \frac{|H_t(\mathbf{X})|}{\|\mathbf{X}\|_\infty} \leq \infty$ .
4. **(Time-homogeneous)**  $H = H_t : t \in \mathbb{R}$  is time-homogeneous (or shift-equivariant) if the input-output relationship commutes with time shift: let  $[S_\tau(\mathbf{X})]_t = \mathbf{x}_{t-\tau}$  be a shift operator, then  $H(S_\tau \mathbf{X}) = S_\tau H(\mathbf{X})$ .
5. **(Causal)**  $H_t$  is causal if it does not depend on future values of the input. That is, if  $\mathbf{X}, \mathbf{X}'$  satisfy  $\mathbf{x}_t = \mathbf{x}'_t$  for any  $t \leq t_0$ , then  $H_t(\mathbf{X}) = H_t(\mathbf{X}')$  for any  $t \leq t_0$ .
6. **(Regular)**  $H_t$  is regular if for any sequence  $\mathbf{X}^{(n)} \in \mathcal{X} : n \in \mathbb{Z}$  such that  $\mathbf{x}_t^{(n)} \rightarrow \mathbf{0}$ , then

$$\lim_{n \rightarrow \infty} H_t(\mathbf{X}^{(n)}) = \mathbf{0}.$$

According to the following lemma, the target we studied in Section 3.1 is a continuous, linear, regular, causal and time-homogeneous functional.

**Lemma F.2.** *Let  $H$  be a family of continuous, linear, regular, causal and time-homogeneous functionals on  $\mathcal{X}$ . Then, there exist a sequence  $\rho : \mathbb{N} \rightarrow \mathbb{R}^d$  that is summable, i.e.*

$$\|\rho\|_{l^1} := \sum_{i=1}^d \sum_{j=0}^{\infty} |(\rho_i)_j| < \infty$$

and

$$H_t(\mathbf{X}) = \sum_{i=0}^{\infty} \rho_i \mathbf{x}_{t-i}, \quad t \in \mathbb{Z}.$$

In particular,  $H$  is uniformly bounded with  $\sup_t \|H_t\| = \|\rho\|_{l^1}$ .

*Proof.* We prove the existence of a summable sequence  $\rho : \mathbb{N} \rightarrow \mathbb{R}^d$  satisfying the stated properties. Let  $\mathcal{X} = \ell^\infty(\mathbb{Z}; \mathbb{R}^d)$  with the sup-norm  $\|\mathbf{X}\| = \sup_{k \in \mathbb{Z}} \max_{1 \leq i \leq d} |X_{k,i}|$ .

### Step 1: Riesz representation and causality.

Since each  $H_t : \mathcal{X} \rightarrow \mathbb{R}$  is continuous, linear, and regular, by the Riesz representation theorem for  $\ell^\infty$  spaces with regular functionals, there exists a unique *finitely additive* signed vector measure  $\mu_t$  on  $\mathbb{Z}$  such that

$$H_t(\mathbf{X}) = \int_{\mathbb{Z}} \mathbf{X}^T d\mu_t = \sum_{k \in \mathbb{Z}} \mathbf{X}_k^T \mu_t(\{k\}),$$

where  $\mu_t(\{k\}) \in \mathbb{R}^d$  is the mass at integer  $k$ , and  $\|H_t\| = \|\mu_t\| = \sum_{k \in \mathbb{Z}} \sum_{i=1}^d |\mu_{t,i}(\{k\})|$ .

Causality implies that  $H_t(\mathbf{X})$  depends only on  $\{\mathbf{X}_k : k \leq t\}$ . Hence,  $\mu_t(\{k\}) = \mathbf{0}$  for all  $k > t$ , and the representation simplifies to:

$$H_t(\mathbf{X}) = \sum_{k=-\infty}^t \mathbf{X}_k^T \mu_t(\{k\}). \quad (1)$$

**Step 2: Time homogeneity and shift invariance.**

For any  $\tau \in \mathbb{Z}$ , time homogeneity implies  $H_t(\mathbf{X}) = H_{t+\tau}(\mathbf{X}^{(\tau)})$  where  $\mathbf{X}_k^{(\tau)} = \mathbf{X}_{k-\tau}$ . Applying (1) to the right-hand side:

$$H_{t+\tau}(\mathbf{X}^{(\tau)}) = \sum_{m=-\infty}^{t+\tau} \mathbf{X}_{m-\tau}^T \boldsymbol{\mu}_{t+\tau}(\{m\}).$$

Substituting  $j = m - \tau$ :

$$H_{t+\tau}(\mathbf{X}^{(\tau)}) = \sum_{j=-\infty}^t \mathbf{X}_j^T \boldsymbol{\mu}_{t+\tau}(\{j + \tau\}).$$

Equating with (1) and comparing coefficients:

$$\boldsymbol{\mu}_t(\{j\}) = \boldsymbol{\mu}_{t+\tau}(\{j + \tau\}) \quad \forall j \leq t. \quad (2)$$

**Step 3: Construction of  $\rho$ .**

Fix  $t = 0$  in (2) and set  $\tau = -t$ :

$$\boldsymbol{\mu}_t(\{j\}) = \boldsymbol{\mu}_0(\{j + t\}) \quad \forall j \leq t.$$

Define  $\boldsymbol{\mu}(\{k\}) = \boldsymbol{\mu}_0(\{k\})$ . By causality,  $\boldsymbol{\mu}(\{k\}) = \mathbf{0}$  for  $k > 0$ . Using time homogeneity:

$$H_t(\mathbf{X}) = H_0(\mathbf{X}^{(-t)}) = \sum_{k=-\infty}^0 \mathbf{X}_{k+t}^T \boldsymbol{\mu}(\{k\}).$$

Substituting  $i = -k$ :

$$H_t(\mathbf{X}) = \sum_{i=0}^{\infty} \mathbf{X}_{t-i}^T \boldsymbol{\mu}(\{-i\}). \quad (3)$$

Define  $\rho_i = \boldsymbol{\mu}(\{-i\})$  for  $i \geq 0$ . Then (3) becomes:

$$H_t(\mathbf{X}) = \sum_{i=0}^{\infty} \mathbf{x}_{t-i}^T \rho_i.$$

**Step 4: Summability and norm equality.**

From the Riesz representation and causality:

$$\|\rho\|_{\ell^1} = \sum_{i=0}^{\infty} \sum_{j=1}^d |(\rho_i)_j| = \sum_{k=-\infty}^0 \sum_{j=1}^d |\mu_j(\{k\})| = \|\boldsymbol{\mu}\| = \|H_0\| < \infty.$$

For any  $t \in \mathbb{Z}$ , using (2) and shift invariance:

$$\|H_t\| = \sum_{j=-\infty}^t \sum_{i=1}^d |\mu_{t,i}(\{j\})| = \sum_{j=-\infty}^t \sum_{i=1}^d |\mu_i(\{j + t\})| = \|\boldsymbol{\mu}\| = \|\rho\|_{\ell^1}.$$

Thus  $\sup_{t \in \mathbb{Z}} \|H_t\| = \|\rho\|_{\ell^1}$ . □

We now consider a continuous time-homogeneous system. According to Theorem C.6 in [83] and [21], we obtain the following decomposition result for such systems:

**Lemma F.3** (Volterra Series Decomposition). *H is a continuous time-homogeneous system with input  $\mathbf{X}$  and output  $\mathbf{Y}$ , then H can be expanded in the Volterra series as follows*

$$\mathbf{y}_t = H_t(\mathbf{X}) = h_0 + \sum_{n=1}^{\infty} \sum_{\tau_1=0}^{\infty} \cdots \sum_{\tau_n=0}^{\infty} H^{(n)}(\tau_1, \dots, \tau_n) (\mathbf{x}_{t-\tau_1} \otimes \cdots \otimes \mathbf{x}_{t-\tau_n}), \quad (6)$$

where  $\otimes$  denotes the Kronecker product. In particular, we call the expansion order  $n$  to be the series' order.

We focus on the term  $H^{(n)}$  and derive its approximation error when using a single-layer Transformer to approximate it. We begin by stating the lemmas required for our analysis:

**Lemma F.4.** *Let  $\{x_i\}_{i=1}^n$  be i.i.d. random variables with  $\mathbb{E}[x_i] = 0$  and  $\mathbb{E}[x_i^2] = 1$ . Then  $\bigotimes_{i=1}^n x_i$  is a random vector with mean zero and identity covariance matrix.*

*Proof.* It suffices to consider the scalar case  $d = 1$ . Define  $y := \prod_{i=1}^n x_i$ . Then

$$\mathbb{E}[y] = \mathbb{E}\left[\prod_{i=1}^n x_i\right] = \prod_{i=1}^n \mathbb{E}[x_i] = 0,$$

and

$$\mathbb{E}[y^2] = \mathbb{E}\left[\left(\prod_{i=1}^n x_i\right)^2\right] = \prod_{i=1}^n \mathbb{E}[x_i^2] = 1.$$

Thus  $\text{Var}(y) = 1$ , which proves the claim.  $\square$

**Lemma F.5.**  *$\mathbf{x}_i$  are random vectors with  $\mathbb{E}\|\mathbf{x}_i\|_2 = B_i$ , and  $\|\mathbf{x}_i - \bar{\mathbf{x}}_i\|_2 \leq \varepsilon_i \leq \varepsilon$ . Then  $\mathbb{E}\|\bigotimes_{i=1}^n \tilde{\mathbf{x}}_i - \bigotimes_{i=1}^n \hat{\mathbf{x}}_i\|_2 \leq \prod_{i=1}^n (B_i + \varepsilon_i) - \prod_{i=1}^n B_i$*

*Proof.* From the property of projection operation,

$$\|\tilde{\mathbf{x}}_i - \hat{\mathbf{x}}_i\|_2 \leq \|P\|_2 \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2 \leq \varepsilon_i \leq \varepsilon.$$

By direct calculation,

$$\begin{aligned} \mathbb{E}\|\bigotimes_{i=1}^n \tilde{\mathbf{x}}_i - \bigotimes_{i=1}^n \hat{\mathbf{x}}_i\|_2 &\leq \sum_{\phi \neq S \subset \{1, \dots, n\}} \left(\prod_{i \in S} \varepsilon_i\right) \mathbb{E}\|\bigotimes_{j \notin S} \tilde{\mathbf{x}}_j\|_2 \\ &= \sum_{\phi \neq S \subset \{1, \dots, n\}} \left(\prod_{i \in S} \varepsilon_i\right) \prod_{j \notin S} B_j \\ &= \prod_{i=1}^n (B_i + \varepsilon_i) - \prod_{i=1}^n B_i. \end{aligned}$$

$\square$

We use the FFN to implement the Kronecker product in our proof; therefore, the universal approximation property of FFNs is also required.

**Definition F.6** (Barron space ([55, 86, 56])). Consider functions  $f : X \rightarrow \mathbb{R}$  that admit the following representation:  $f(\mathbf{x}) = \int_{\Omega} a \sigma(\mathbf{b}^\top \mathbf{x} + c) \rho(\mathrm{d}a, \mathrm{d}\mathbf{b}, \mathrm{d}c)$ ,  $\mathbf{x} \in X$ . For any  $p \in [1, +\infty]$ , we define the Barron norm as  $\|f\|_{\mathcal{B}_p} := \inf_{\rho} \left( \mathbb{E}_{\rho} [|a|^p (\|\mathbf{b}\|_1 + |c|)^p] \right)^{1/p}$ . Then the Barron space are defined as:  $\mathcal{B}_p := \{f \in \mathcal{C} : \|f\|_{\mathcal{B}_p} < +\infty\}$ .

**Lemma F.7** ([56]). *For any  $f \in \mathcal{B}$ , there exists a two-layer ReLU neural network  $\text{FFN}(\mathbf{x}) = \sum_{k=1}^K a_k \sigma(\mathbf{b}_k^\top \mathbf{x} + c_k)$  with  $M$  neurons such that*

$$\|f - \text{FFN}\|_{L^\infty[-2, 2]} \leq \mathcal{O}\left(\frac{\|f\|_{\mathcal{B}} \sqrt{\log K}}{\sqrt{K}}\right).$$

We now present the approximation and tradeoff result:

**Theorem F.8** (Tradeoff for Nonlinear Target). *Consider the  $n$ -th order Volterra term*

$$H_t(\mathbf{X}) = \sum_{\tau_1=0}^{\infty} \cdots \sum_{\tau_n=0}^{\infty} H^{(n)}(\tau_1, \dots, \tau_n)(\mathbf{x}_{t-\tau_1} \otimes \cdots \otimes \mathbf{x}_{t-\tau_n}).$$

Suppose that we have  $H_m$  heads of dimension  $d_m \leq d$  for  $m = 1, \dots, M$ , and that the model dimension

$$D = \sum_{m=1}^M H_m d_m$$

is fixed.

For  $m \in [M]$ , define

$$\gamma_m := \mathbb{I}_{\{d_m < d\}} \left( \sqrt{1 - \frac{d_m}{d}} + C \sqrt{\frac{\log(2M/\delta)}{d}} \right),$$

and

$$\eta_m := B \frac{1.3e^{0.02m}}{H_m}.$$

For the current token, which is preserved by the residual connection, set

$$\gamma_0 = \eta_0 = 0.$$

Then, with probability at least  $1 - \delta$ ,

$$\mathcal{E}_D(\mathbf{X}) \leq \sum_{\tau_1=0}^M \cdots \sum_{\tau_n=0}^M \left\| \mathbf{H}^{(n)}(\tau_1, \dots, \tau_n) \right\|_2 \left[ B^n \sum_{j=1}^n \gamma_{\tau_j} + \prod_{j=1}^n (B + \eta_{\tau_j}) - B^n \right] + \varepsilon_H + \varepsilon_{\text{FFN}}. \quad (7)$$

Here,

$$\begin{aligned} \varepsilon_H := \sup_t \left\| \sum_{\tau_1=0}^{\infty} \cdots \sum_{\tau_n=0}^{\infty} \mathbf{H}^{(n)}(\tau_1, \dots, \tau_n) (\mathbf{x}_{t-\tau_1} \otimes \cdots \otimes \mathbf{x}_{t-\tau_n}) \right. \\ \left. - \sum_{\tau_1=0}^M \cdots \sum_{\tau_n=0}^M \mathbf{H}^{(n)}(\tau_1, \dots, \tau_n) (\mathbf{x}_{t-\tau_1} \otimes \cdots \otimes \mathbf{x}_{t-\tau_n}) \right\|_2, \end{aligned} \quad (8)$$

and  $\varepsilon_{\text{FFN}}$  denotes the approximation error incurred when the FFN implements the required Kronecker-product operations.

**Corollary F.9** (Parameter allocation via trade-offs). *Under the same conditions as Theorem F.8, a principled parameter allocation can be obtained by solving*

$$\begin{aligned} \min_{M, \{H_m, d_m\}_{m=1}^M} \quad & \sum_{\tau_1=0}^M \cdots \sum_{\tau_n=0}^M \left\| \mathbf{H}^{(n)}(\tau_1, \dots, \tau_n) \right\|_2 \left[ B^n \sum_{j=1}^n \gamma_{\tau_j} \right. \\ & \left. + \prod_{j=1}^n (B + \eta_{\tau_j}) - B^n \right] + \varepsilon_H(M) \end{aligned} \quad (9)$$

$$\begin{aligned} \text{s.t.} \quad & \sum_{m=1}^M H_m d_m = D, \\ & M, H_m, d_m \in \mathbb{N}_+, \quad d_m \leq d, \end{aligned}$$

where

$$\gamma_m = \mathbb{I}_{\{d_m < d\}} \left( \sqrt{1 - \frac{d_m}{d}} + C \sqrt{\frac{\log(2M/\delta)}{d}} \right), \quad \eta_m = B \frac{1.3e^{0.02m}}{H_m},$$

with  $\gamma_0 = \eta_0 = 0$ . If  $\varepsilon_{\text{FFN}}$  is fixed independently of the allocation, it can be omitted from the optimization objective.

*Proof.* The construction of the embedding layer and the multi-head attention layer follows the same procedure as in the proof of Theorem 3.1. After these two steps, for each token we obtain

$$\begin{pmatrix} \tilde{\mathbf{x}}_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} \tilde{\mathbf{x}}_t \\ \tilde{\mathbf{x}}_{t-1} \\ \vdots \\ \tilde{\mathbf{x}}_{t-M} \\ 0 \end{pmatrix} := \mathbf{x}_t^{(1/2)}.$$

For each  $m \in [M]$ , let  $\mathbf{P}_m \in \mathbb{R}^{d_m \times d}$  be the random projection used in the proof of Theorem 3.1, and define  $\mathbf{\Pi}_m = \mathbf{P}_m^\top \mathbf{P}_m$ . By Lemma C.3 and a union bound over the  $M$  projection groups, with probability at least  $1 - \delta$ ,

$$\|(\mathbf{I} - \mathbf{\Pi}_m)\mathbf{x}\|_2 \leq B \left[ \mathbb{I}_{\{d_m < d\}} \sqrt{1 - \frac{d_m}{d}} + C \sqrt{\frac{\log(2M/\delta)}{d}} \right], \quad m \in [M].$$

Define

$$\gamma_\delta := \mathbb{I}_{\{d_{\min} < d\}} \sqrt{1 - \frac{d_{\min}}{d}} + C \sqrt{\frac{\log(2M/\delta)}{d}}, \quad d_{\min} := \min_{m \in [M]} d_m.$$

Then, for every  $m \in [M]$ ,

$$\|(\mathbf{I} - \mathbf{\Pi}_m)\mathbf{x}\|_2 \leq B\gamma_\delta.$$

For any  $(\tau_1, \dots, \tau_n) \in \{0, \dots, M\}^n$ , the difference between the original and projected tensor representations admits the telescoping decomposition

$$\begin{aligned} & \mathbf{x}_{t-\tau_1} \otimes \dots \otimes \mathbf{x}_{t-\tau_n} - \tilde{\mathbf{x}}_{t-\tau_1} \otimes \dots \otimes \tilde{\mathbf{x}}_{t-\tau_n} \\ &= \sum_{j=1}^n \tilde{\mathbf{x}}_{t-\tau_1} \otimes \dots \otimes \tilde{\mathbf{x}}_{t-\tau_{j-1}} \otimes (\mathbf{x}_{t-\tau_j} - \tilde{\mathbf{x}}_{t-\tau_j}) \otimes \mathbf{x}_{t-\tau_{j+1}} \otimes \dots \otimes \mathbf{x}_{t-\tau_n}. \end{aligned}$$

Since  $\|\mathbf{x}_t\|_2 \leq B$  and orthogonal projection is non-expansive,  $\|\tilde{\mathbf{x}}_t\|_2 \leq B$ . Therefore,

$$\|\mathbf{x}_{t-\tau_1} \otimes \dots \otimes \mathbf{x}_{t-\tau_n} - \tilde{\mathbf{x}}_{t-\tau_1} \otimes \dots \otimes \tilde{\mathbf{x}}_{t-\tau_n}\|_2 \leq nB^n\gamma_\delta.$$

Furthermore, by Lemma C.2,

$$\|\tilde{\mathbf{x}}_{t-\tau_1} \otimes \dots \otimes \tilde{\mathbf{x}}_{t-\tau_n} - \hat{\mathbf{x}}_{t-\tau_1} \otimes \dots \otimes \hat{\mathbf{x}}_{t-\tau_n}\|_2 \leq (B + \varepsilon_{\text{Attn}})^n - B^n,$$

where

$$\varepsilon_{\text{Attn}} = \max_{m \in [M]} \frac{1.3e^{0.02m}}{H_m}.$$

Combining the projection error, attention-extraction error, kernel truncation error  $\varepsilon_H$ , and FFN approximation error  $\varepsilon_{\text{FFN}}$ , with probability at least  $1 - \delta$ ,

$$\begin{aligned} \mathcal{E}_D(\mathbf{X}) &\leq \sum_{\tau_1=0}^M \dots \sum_{\tau_n=0}^M \left\| \mathbf{H}^{(n)}(\tau_1, \dots, \tau_n) \right\|_2 [nB^n\gamma_\delta + (B + \varepsilon_{\text{Attn}})^n - B^n] \\ &\quad + \varepsilon_H + \varepsilon_{\text{FFN}}. \end{aligned}$$

□

## G Additional Experiments

### G.1 Inspiration of Parameter Allocation

The former work [40] empirically investigated how the head trade-off is influenced by the memory structure and the difficulty of the target relationship in Figure 14. As illustrated in Figure 4 of their work, the trade-off depends on both the underlying memory structure and the level of task difficulty. Each subfigure corresponds to a specific memory structure, with the plots showing loss as a function of the number of heads. The color of the curves encodes the parameter  $\alpha$ , which reflects task difficulty. The results indicate that, for certain tasks, the trade-off remains consistent across different levels of difficulty (Figure (c)). In tasks (a) and (b), when the task is relatively easy, no clear trade-off is observed. However, in task (c), a trade-off appears consistently, regardless of task difficulty.

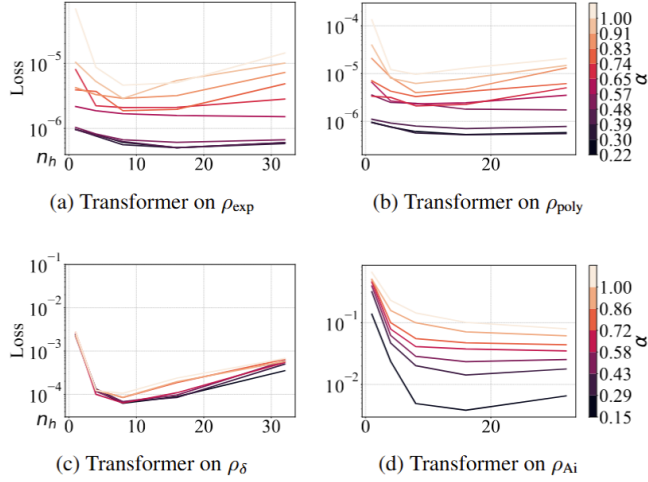


Figure 4: Loss versus  $n_h$  for Transformers with fixed  $m = 128$  and varying  $\alpha$ .

Figure 14: Linear Convolution Tradeoff

Table 1: Jacobian spectral norm at different query positions.

| Query position $l \in [L]$ | Mean Jacobian spectral norm $\ J_l\ _2$ | $l\ J_l\ _2$ |
|----------------------------|-----------------------------------------|--------------|
| 2                          | $4.512 \times 10^{-1}$                  | 0.9024       |
| 4                          | $2.792 \times 10^{-1}$                  | 1.1168       |
| 8                          | $1.506 \times 10^{-1}$                  | 1.2048       |
| 16                         | $7.413 \times 10^{-2}$                  | 1.1861       |
| 32                         | $3.582 \times 10^{-2}$                  | 1.1462       |
| 64                         | $1.753 \times 10^{-2}$                  | 1.1219       |
| 128                        | $1.122 \times 10^{-2}$                  | 1.4362       |
| 256                        | $5.718 \times 10^{-3}$                  | 1.4638       |
| 512                        | $3.312 \times 10^{-3}$                  | 1.6957       |
| 1024                       | $1.650 \times 10^{-3}$                  | 1.6896       |

## G.2 Verification of Theorem 4.1 and Corollary 4.2

To verify Theorem 4.1: We rerun Figure 2 for a complementary evaluation protocol. In the original manuscript, we varied the sequence length  $L$  and computed the Jacobian (spectral) norm associated with the last query. Here, under the same model and data setting, we instead fix  $L = 1024$  and compute the Jacobian (spectral) norm associated with the  $l$ -th query (averaged over 10 sampled sequences) as follows. It is clearly to have  $l\|J_l\|_2 \in [0.9, 1.7]$  across reported positions, verifying the  $\mathcal{O}(1/l)$ -bound.

Consistency between Corollary 4.2 and Figure 3: An implication of Corollary 4.2 is, for most of queries at larger positions (say  $l \geq 0.1L$ ), the distillation error per query scales like  $\mathcal{O}(1/L) \rightarrow 0$  ( $L \gg 1$ ); while for a small portion of queries at smaller positions (say  $l \leq 0.1L$ ), the corresponding Jacobian mapping is highly degenerated ( $\text{rank}(J_l) \leq l \leq 0.1L$ ) and hence controls the distillation error incurred by truncated SVD. This makes the SVD initialization a warm start, but also need further refinements to reduce the distillation error corresponding to earlier queries. Both steps together form Algorithm 1, which is implemented and verified in Figure 3.

## G.3 Role of Layers in Figure 4

We further evaluate layer-wise head-dimension reduction on LLAMA-3-8B, measured by perplexity (PPL), by separately compressing the deepest layers (21–24) and several early layers (2, 3, and 4). The results are reported in Table 2.

Table 2: Layer-wise head-dimension reduction on LLAMA-3-8B, evaluated by perplexity (PPL).

| Reduced rank $d_h$ | Deepest layers 21–24 | Layer 2 | Layer 3 | Layer 4 |
|--------------------|----------------------|---------|---------|---------|
| 1                  | 4.72                 | 6.02    | 4.69    | 4.75    |
| 4                  | 4.72                 | 5.61    | 4.61    | 4.76    |
| 8                  | 4.71                 | 5.41    | 4.59    | 4.67    |
| 16                 | 4.68                 | 5.22    | 4.57    | 4.58    |
| 32                 | 4.61                 | 4.88    | 4.52    | 4.52    |
| 64                 | 4.51                 | 4.57    | 4.45    | 4.47    |
| 128                | 4.41                 | 4.41    | 4.41    | 4.41    |

These results confirm a clear layer dependence of model reduction. Compressing the deepest layers causes almost no perplexity degradation even at very small ranks. Layers 3 and 4 are similarly robust, whereas layer 2 is moderately more sensitive. We will add these new experiments in the revision to emphasize that even the saturation behavior is layer-dependent, it is particularly pronounced in deep layers and hence leads to lossless later-layer model reduction.

#### G.4 Reducing Budget on Number of Heads

We conducted a systematic investigation of Transformer architectural hyperparameters on the WikiText-103 language modeling task. The goal was to understand how attention head dimension and number of heads affect model performance in next-token prediction.

The models were 6-layer GPT-style Transformer decoders with learned token and positional embeddings (sequence length 256), RMSNorm, SwiGLU feed-forward networks, residual connections, and weight-tied output projections. Two hyperparameter sweeps were performed: (1) varying head dimension with 8 attention heads fixed, and (2) varying the number of heads (1–32) with head dimension fixed at 64. The model dimension  $d_{\text{model}}$  was set as the product of number of heads and head dimension.

Training was performed on 1% of WikiText-103 ( 1M tokens) for 2 epochs using AdamW (learning rate  $3 \times 10^{-4}$ , weight decay 0.1) with cosine scheduling, 200-step warmup, gradient clipping, dropout 0.1, and mixed precision. Models were evaluated on validation and test sets using cross-entropy loss and perplexity. Each configuration recorded training curves, validation losses, and test performance for analysis.

Comparative analysis reveals that cutting head dimension is a more effective way to reduce model parameters than reducing the number of heads, which incurs higher loss. The study highlights how architectural choices influence performance and parameter efficiency in language modeling, offering insights for balancing model capacity and computational budget.

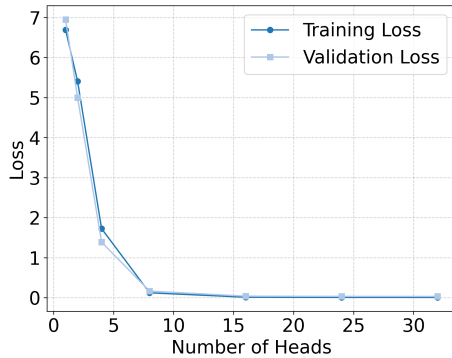


Figure 15: Loss vs Number of Heads

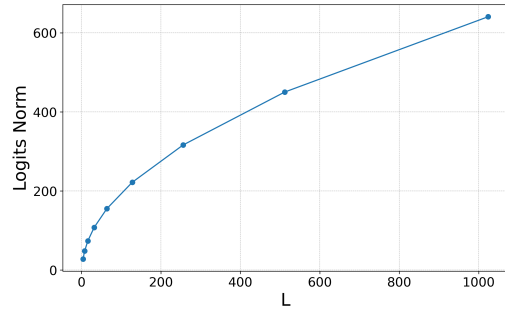


Figure 16: Logits Norm vs L

### G.5 Analysis of Attention Logits Norms

In Figure 16, we investigate how the  $\ell_2$  norm of attention logits (pre-softmax) varies with sequence length. For each sequence length  $L$ , we randomly sampled 10 text sequences from the Wikitext-103 dataset and performed forward passes using the TinyLlama-1.1B model. We focused on the logits of the first attention head in the first layer (layer 0, head 0) and computed the  $\ell_2$  norm at the middle token of each sequence (index =  $L/2$ ).

This analysis has several theoretical motivations:

- **Gradient stability:** The magnitude of logits directly affects the softmax gradients. Excessively large norms may cause gradient explosion, while very small norms lead to overly uniform attention distributions.
- **Numerical stability:** Extreme logits values can result in numerical overflow, impacting training stability.
- **Attention concentration:** Differences in logits magnitudes indicate the degree of attention focus, with larger variations corresponding to more concentrated attention.

By analyzing the logits norm at the middle token, we avoid boundary effects and ensure comparability across sequences of different lengths. We hypothesize that the logits norm exhibits a specific scaling behavior with sequence length and should remain relatively stable to support effective gradient propagation. These observations provide insights into the internal mechanisms of Transformers when handling long sequences and can guide design choices for long-sequence modeling.

### G.6 Controlled Ablations for the Two-stage Reduction Pipeline

We conduct controlled ablations on the corrected permanent low-rank implementation. The reported sweep uses seed 42. The main comparison fixes layers 5–12 and  $d_h = 16$ , while all other settings remain unchanged.

Table 3: Controlled ablation at layers 5–12 and  $d_h = 16$ . WikiText-2 losses are evaluated immediately after initialization and after Stage I; SST-2 accuracy is evaluated after Stage II (or without Stage II when it is disabled).

| Initialization | Stage I | Stage II | Init. loss | Stage-I loss | SST-2 Acc. (%) |
|----------------|---------|----------|------------|--------------|----------------|
| SVD            | ✓       | ✓        | 5.072      | 3.215        | 93.81          |
| Kaiming        | ✓       | ✓        | 7.180      | 3.582        | 91.74          |
| SVD            | ×       | ✓        | 5.072      | 5.072        | 91.86          |
| SVD            | ✓       | ×        | 5.072      | 3.215        | 47.48          |

Table 4: SST-2 validation accuracy (%) for Kaiming initialization + Stage I + Stage II.

| Layers \ $d_h$ | 2     | 4     | 8     | 16    | 32    | 64    |
|----------------|-------|-------|-------|-------|-------|-------|
| 1–8            | 77.87 | 79.93 | 83.03 | 86.81 | 87.96 | 88.30 |
| 5–12           | 88.65 | 91.06 | 91.63 | 91.74 | 92.66 | 92.78 |
| 9–16           | 93.35 | 93.23 | 93.23 | 93.58 | 93.00 | 94.61 |
| Odd layers     | 81.77 | 84.63 | 88.19 | 89.91 | 90.37 | 90.83 |
| 1–16           | 78.10 | 80.16 | 82.45 | 84.40 | 87.16 | 88.07 |

Table 5: SST-2 validation accuracy (%) for SVD initialization + Stage II, without Stage I.

| Layers \ $d_h$ | 2     | 4     | 8     | 16    | 32    | 64    |
|----------------|-------|-------|-------|-------|-------|-------|
| 1–8            | 71.79 | 79.93 | 81.65 | 86.70 | 89.91 | 94.50 |
| 5–12           | 89.68 | 91.28 | 91.63 | 91.86 | 94.04 | 94.84 |
| 9–16           | 91.86 | 93.69 | 94.38 | 94.38 | 94.95 | 95.07 |
| Odd layers     | 82.22 | 88.88 | 90.14 | 91.63 | 93.12 | 94.84 |
| 1–16           | 76.03 | 81.77 | 84.06 | 87.73 | 90.60 | 95.53 |

Table 6: SST-2 validation accuracy (%) for SVD initialization + Stage I, without Stage II.

| Layers \ $d_h$ | 2     | 4     | 8     | 16    | 32    | 64    |
|----------------|-------|-------|-------|-------|-------|-------|
| 1–8            | 48.28 | 49.89 | 47.94 | 47.59 | 47.71 | 48.51 |
| 5–12           | 48.39 | 47.82 | 48.85 | 47.48 | 47.71 | 48.74 |
| 9–16           | 49.54 | 48.28 | 50.11 | 48.51 | 48.51 | 49.08 |
| Odd layers     | 46.22 | 47.36 | 46.79 | 47.94 | 48.28 | 49.08 |
| 1–16           | 47.25 | 48.74 | 47.13 | 47.94 | 47.13 | 48.28 |