ARTIFICIAL ENTANGLEMENT IN THE FINE-TUNING OF LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) can be adapted to new tasks using parameter-efficient fine-tuning (PEFT) methods that modify only a small number of trainable parameters, often through low-rank updates. In this work, we adopt a quantum-information-inspired perspective to understand their effectiveness. From this perspective, low-rank parameterizations naturally correspond to low-dimensional Matrix Product States (MPS) representations, which enable entanglement-based characterizations of parameter structure. Thereby, we term and measure *"Artificial Entanglement"*, defined as the entanglement entropy of the parameters in artificial neural networks (in particular the LLMs). We first study the representative low-rank adaptation (LoRA) PEFT method, alongside full fine-tuning (FFT), using LLaMA models at the 1B and 8B scales trained on the Tulu3 and OpenThoughts3 datasets, and uncover: (i) *Internal artificial entanglement* in the updates of query and value projection matrices ($\Delta W_Q, \Delta W_V$) in LoRA follows a *volume law* with a central suppression (termed as the *"Entanglement Valley"*), which is sensitive to hyper-parameters and is distinct from that in FFT; (ii) *External artificial entanglement* in attention matrices, corresponding to token–token correlations in representation space, follows an *area law* with logarithmic corrections and remains robust to LoRA hyper-parameters and training steps. Drawing a parallel to the No-Hair Theorem in black hole physics, we propose that although LoRA and FFT induce distinct internal entanglement signatures, such differences do not manifest in the attention outputs, suggesting a "no-hair" property that results in the effectiveness of low rank updates. We further provide theoretical support based on random matrix theory, and extend our analysis to an *MPS Adaptation* PEFT method, which exhibits qualitatively similar behaviors.

1 INTRODUCTION

Recently, large language models (LLMs) have demonstrated desirable abilities by scaling model size and pretraining on large corpora of data (Radford et al., 2018; 2019; Brown et al., 2020; Devlin et al., 2019; Raffel et al., 2020; Vaswani et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022; Henighan et al., 2020; Chowdhery et al., 2023; Achiam et al., 2023). Post-training, usually via fine-tuning, is adopted to adapt pretrained LLMs for downstream tasks. However, it is often considered unnecessary and computationally expensive to update the full set of parameters. Therefore, many parameter-efficient fine-tuning (PEFT) methods are proposed, including low-rank adaptation (LoRA) (Hu et al., 2022) and its variants (Dettmers et al., 2023; Zhang et al., 2023b; Liu et al., 2024; Kopiczko et al., 2024; Hayou et al., 2024; Valipour et al., 2023; Zhang et al., 2023a; 2024; Zi et al., 2023; Lialin et al., 2023). Previous studies (Hu et al., 2022; Li et al., 2018; Shuttleworth et al., 2024; Hayou et al., 2024; Kalajdjieski, 2024; Pan et al., 2024) also contribute significantly to explaining their effectiveness. However, a fundamental question remains unsolved: *how complex are the internal parameter structures induced by distinct fine-tuning methods, and to what extent are such complexities reflected in attention-level representations?*

We investigate this question by adopting a quantum information perspective, and draw a parallel with a well-known phenomenon in theoretical physics: the tension between internal quantum correlations and external simplicity. This tension has been most prominently explored within black hole physics (Bombelli et al., 1986; Srednicki, 1993; Eisert et al., 2008). A fundamental conflict (Israel, 1967; Carter, 1971; Bhattacharya & Lahiri, 2007; Mavromatos, 1996) emerges between the quantum

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

mechanical description of black holes (via entanglement entropy) and their classical geometric description. While the scaling behavior of entanglement entropy implies that black holes are complex systems characterized by a vast number of degrees of freedom, classical gravitational theories predict that black holes are surprisingly simple objects with no distinct features. This classical perspective is formalized by uniqueness theorems (Israel, 1968; Carter, 1971; Robinson, 1975). The physical essence of these theorems is encapsulated in the No-Hair Theorem (Israel, 1967; 1968; Carter, 1971; Robinson, 1975), which states that all stationary black hole solutions of the Einstein–Maxwell equations can be completely characterized by only three externally observable classical parameters: mass, electric charge, and angular momentum. All other semi-classical information about the matter that formed the black hole (the "hair") disappears behind the event horizon and is inaccessible to an external observer (Harlow, 2016).

Motivated by this dilemma, we define and probe the *"artificial entanglement"*, where "artificial" refers to the application of entanglement entropy measures to artificial neural networks (particularly the LLMs). Notably, "artificial" emphasizes that large language models are entirely classical systems and do not possess any physical quantum degrees of freedom or genuine quantum entanglement. Therefore, this adopted perspective serves purely as a mathematical tool. Specifically, we represent the updates of the query and value projection matrices ($\Delta W_Q, \Delta W_V$) as Matrix Product States (MPS) (Eisert et al., 2008; Vidal & Werner, 2002; Vidal, 2003; Vidal et al., 2003; Vidal, 2004; 2007; Orús, 2019; Berezutskii et al., 2025; Milsted et al., 2022) and compute the von Neumann entanglement entropy across bond partitions, yielding an *artificial entanglement profile* that captures the internal parameter structure. We refer to the corresponding entanglement as the *"internal artificial entanglement"*. Meanwhile, we apply the same routine to the attention matrices and outputs produced during the forward computation rather than to the model parameters, and refer to it as the *"external artificial entanglement"*, which corresponds to token–token correlations in representation space. Our experiments are conducted on the Llama 3 series models (Dubey et al., 2024) at the 1B and 8B scales, and employ the Tulu3 (Lambert et al., 2024) and OpenThoughts3 (Guha et al., 2025) datasets, which target instruction following and reasoning tasks respectively. Our key findings are summarized in FIG. 1. **Our contributions are as follows:**

(i) We show that the internal artificial entanglement under LoRA follows a *volume law* (Eisert et al., 2008; Amico et al., 2008) with a pronounced central suppression, termed as the *"Entanglement Valley"*, which is highly sensitive to hyperparameters such as the rank r and the scaling factor α and is qualitatively distinct from full fine-tuning (FFT). In contrast, the external artificial entanglement follows an approximate *area law with logarithmic corrections* and remains robust across hyper-parameters and training stages. Drawing a parallel to the No-Hair Theorem, we show that the attention mechanism exhibits a "no-hair" property, and propose that LoRA learns intrinsically different internal parameter structures from FFT but remains effective because the attention mechanism coarse-grains these differences.

(ii) Based on random matrix theory, we establish an *"Attention Cardy Formula"*, which shows that the entanglement entropy of attention matrices scales logarithmically with sequence length under certain initialization conditions, analogous to the Cardy formula (Calabrese & Cardy, 2004; 2009) for critical quantum systems. Moreover, we relate low external artificial entanglement to *stable rank collapse* (Saada et al., 2024), see *Appendix E*.

(iii) We extend our analysis beyond LoRA to an *MPS Adaptation* PEFT strategy, in which weight updates are parameterized directly via MPS, and demonstrate that it exhibits qualitatively similar internal and external artificial entanglement behaviors, validating the generality of our findings.

2 Related Works

Some empirical evidence provided by Thinking Machines Lab (Schulman & Lab, 2025) offers intuitive visualizations of the performance in LoRA and FFT. We echo some findings in our study and advance these observations. An important feature of ours can be concluded as identifying a "no-hair" behavior that justifies the effectiveness of LoRA. (Shuttleworth et al., 2024) discusses different spectral properties of the resulting final weights obtained by LoRA and FFT, compared with pretrained weights. We focus on (i) the direct fine-tuning solutions ΔW, the attention matrices and outputs; (ii) the entanglement structure; (iii) test performance w.r.t hyper-parameters, while they focus on generalization and forgetting. (Saada et al., 2024) shows a phenomenon termed as *rank*

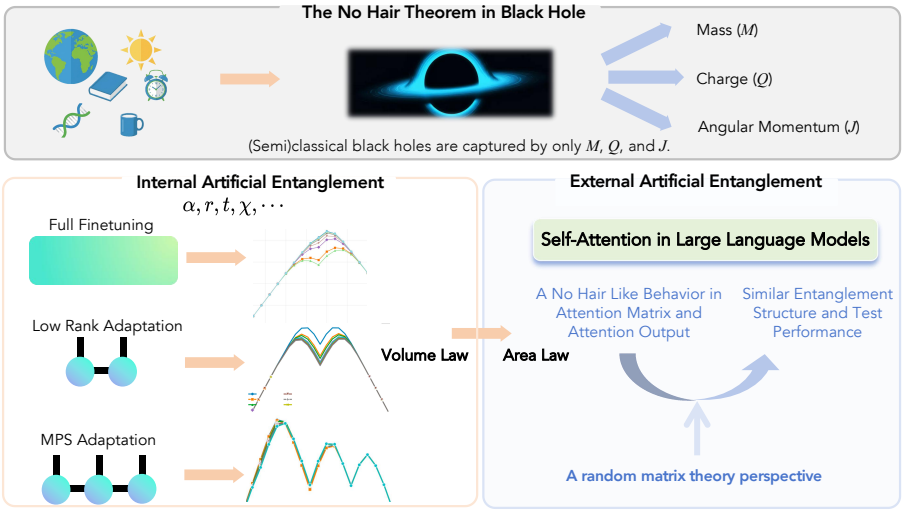

Figure 1: Key Findings. **(i)** Projection matrices (ΔW_Q and ΔW_V) exhibit volume-law internal entanglement profiles with distinctive entanglement valleys that differ between FFT, LoRA and MPS adaptation. **(ii)** Attention matrices show area-law scaling with logarithmic corrections. Random matrix theory explains this through the Attention Cardy Formula. **(iii)** Despite internal differences, external attention outputs remain invariant, revealing a no-hair–like effect where the attention mechanism acts as a coarse-graining operator.

collapse in width, both in attention matrices and attention outputs. We theoretically link our findings to this perspective (See Appendix E). Notably, TT-LoRA (Anjum et al., 2024) utilizes an equivalent Tensor Train decomposition. While it focuses on storage efficiency, we utilize the tensor network formalism to probe entanglement signatures. We do not focus on other variants of LoRA but the original LoRA itself.

3 Preliminaries and Backgrounds

FFT and LoRA. In this section, we first review FFT and LoRA (Hu et al., 2022) methods. FFT adapts a pre-trained model by updating all model parameters. Given a pre-trained weight projection matrix W_0, the model learns a full matrix update ΔW_{Full}, resulting in $W = W_0 + \Delta W_{\text{Full}}$. In contrast, LoRA injects trainable low-rank matrices to approximate the weight updates. Specifically, LoRA decomposes the update into two smaller matrices, denoted as $A \in \mathbb{C}^{r \times d_{\text{in}}}$ and $B \in \mathbb{C}^{d_{\text{out}} \times r}$,[1] where the rank $r \ll \min(d_{\text{in}}, d_{\text{out}})$. In addition, LoRA uses a scaling hyperparameter α to control the magnitude of the low-rank update. Formally, LoRA update is formulated as

$$\Delta W_{\text{LoRA}} := \frac{\alpha}{r} BA \in \mathbb{C}^{d_{\text{out}} \times d_{\text{in}}}, \tag{1}$$

which is then added to the frozen pre-trained weight W_0.

Matrix Product States. Among various tensor network architectures, the *Matrix Product State* (MPS) is the most widely used one-dimensional tensor network. An MPS expresses a high-order tensor as a chain of Order-3 tensors connected by virtual bonds. The MPS representation approximates this tensor by factorizing it into a sequence of Order-3 tensors:

$$c_{i_1,\ldots,i_M} = \sum_{\alpha_1=1}^{\chi_1} \cdots \sum_{\alpha_{M-1}=1}^{\chi_{M-1}} \mathcal{T}^{[1]\,i_1}_{1,\alpha_1} \mathcal{T}^{[2]\,i_2}_{\alpha_1,\alpha_2} \cdots \mathcal{T}^{[M]\,i_M}_{\alpha_{M-1},1}, \tag{2}$$

[1]For notational generality, we allow the matrices to be complex-valued. All experiments in this work are conducted with real-valued parameters as in standard LoRA implementations.

where each $\mathcal{T}^{[k]}$ is an Order-3 tensor with one physical index i_k and two virtual indices α_{k-1}, α_k. The virtual indices are contracted, yielding the chain-like network structure. The parameter χ_k denotes the bond dimension at bond k, which may vary along the chain. The quantity $\chi := \max_k \chi_k$ is often referred to as the (maximum) *bond dimension*, which determines the expressive power of the MPS and controls the entanglement it can capture: larger χ allows more accurate approximations of c, while small χ corresponds to a compressed representation.

4 METHODS

To analyze artificial entanglement in fine-tuning, we employ a framework that includes MPS decomposition with entanglement entropy computation. Our approach consists of four key steps: (i) reshaping matrices (including the updates of weight projection matrices and attention-related matrices) into higher-order tensors, (ii) factorizing them into MPS representations, (iii) interpreting the MPS as a mathematical formalism of many-body states (formally analogous to quantum many-body states, though the underlying systems are entirely classical), and (iv) computing the von Neumann entanglement entropy across different bonds[2] by performing SVD and obtaining singular values (Schmidt indices), yielding an *artificial entanglement profile*. This profile quantifies how correlations vary across different positions, revealing the fruitful internal structure. Below we take the update of a projection matrix ΔW (ΔW_Q and ΔW_V) as an example and detail the essential steps (See preliminaries on tensor basis in *Appendix B*):

Step 1: Although $\Delta W \in \mathbb{C}^{d_{\text{out}} \times d_{\text{in}}}$ is originally an Order-2 tensor, we first reshape it into a higher-order tensor by factorizing its input and output dimensions into their prime components. This produces the finest-grained decomposition

$$d_{\text{out}} = \prod_{k=1}^{n} f_k, \qquad d_{\text{in}} = \prod_{\ell=1}^{m} g_\ell, \tag{3}$$

where n and m denote the number of prime factors in d_{out} and d_{in}, respectively. This allows us to view ΔW as a tensor living on $(n+m)$ sites. This provides the "lattice" on which we build the MPS.

Step 2: We then perform a sequence of singular value decompositions (SVDs) along the reshaped tensor dimensions to factorize ΔW into a chain of local Order-3 tensors. The resulting representation

$$\Delta W \longrightarrow \{\mathcal{T}^{[1]}, \mathcal{T}^{[2]}, \ldots, \mathcal{T}^{[n+m]}\} \tag{4}$$

is an MPS whose virtual bond indices capture how different tensor sites are correlated. For LoRA, the factorization naturally reflects its low-rank structure: the update BA induces a restricted bond dimension. For FFT, the MPS is obtained directly from ΔW_{Full} without such constraints.

Step 3: Once factorized, the MPS can be interpreted as a mathematical formalism of many-body states. Specifically, the MPS factorization of ΔW is formally analogous to the MPS representation of the amplitude tensor of a quantum many-body state $|\Psi\rangle$ (Liu et al., 2023; Chen et al., 2023) (though the underlying system is entirely classical). Each "site" corresponds to one of the factorized input/output degrees of freedom, and the virtual bonds represent the latent coupling strength between subsystems.

Step 4: When performing an SVD at each bond during the MPS decomposition, we obtain the singular values $\{\sigma_{\alpha_k}\}$, where k denotes the bond position (the k-th bond in the MPS chain) and α_k indexes the Schmidt modes at bond k. We normalize the singular values according to

$$\lambda_{\alpha_k} = \frac{\sigma_{\alpha_k}}{\sqrt{\sum_{\alpha_k} \sigma_{\alpha_k}^2}}, \tag{5}$$

such that $\sum_{\alpha_k} \lambda_{\alpha_k}^2 = 1$. These normalized Schmidt coefficients define the bipartite entanglement entropy,

$$S_k = -\sum_{\alpha_k} \lambda_{\alpha_k}^2 \log \lambda_{\alpha_k}^2, \tag{6}$$

[2]With a slight abuse of terminology, we sometimes use the terms *bond*, *cut*, *cut position*, and *bi-partition* interchangeably, as they are in one-to-one correspondence in the one-dimensional settings considered here. Specifically, a cut at a given position induces a bi-partition of the Hilbert space, while in an MPS representation such a cut corresponds to cutting a virtual bond.

where S_k quantifies the entanglement entropy across bond k. Scanning across all bonds yields an *"artificial entanglement profile"*, allowing us to probe correlation patterns. The full details are provided in Sec. C.

Remark (On the use of the term "artificial entanglement"). The entanglement entropy considered in this work is purely a mathematical quantity derived from the Schmidt decomposition of high-dimensional tensors, and should not be interpreted as physical quantum entanglement. Although large language models are entirely classical systems, MPS representations and their associated entanglement measures provide a principled, quantum-inspired, basis-independent way to characterize how correlations are distributed under different bipartitions. In this sense, the resulting "artificial entanglement profiles" are used as descriptive diagnostics of structural correlations, rather than as real physical entanglement in an underlying physical Hilbert space.

5 Experiments

Default Experimental Setups. In the main text, we present the results in fine-tuning the Llama-3.2-1B-Instruct, an instruction-tuned model with 1 billion parameters from Llama-3.2 family (Touvron et al., 2023; Dubey et al., 2024), on a subset of the tulu-3-sft-mixture instruction-tuning (Tulu3) corpus (Lambert et al., 2024). For more experiments in other model and dataset setups, see *Appendix I*. Unless otherwise stated, we adopt and keep the following settings: we uniformly subsample 5,000 examples from the full dataset and reserve 5% as a test split. For LoRA we use rank $r = 256$ and a scaling hyperparameter $\alpha = 16$, applied only to the self-attention query and value projections (W_Q and W_V) (Hu et al., 2022). This is to follow the default settings (Schulman & Lab, 2025), serving as representative cases to study the resulting behaviors. The learning rate is set to 3×10^{-4} for LoRA and 3×10^{-5} for FFT, since under this setup the testing loss is approximately equal (Schulman & Lab, 2025). The test loss is defined as the average token-level cross-entropy, *i.e.*, negative log-likelihood (NLL). We quantify artificial entanglement using the von Neumann entropy (Eisert et al., 2008; Orús, 2019; Berezutskii et al., 2025) and denote it as S, the standard measure of entanglement in quantum information theory.

5.1 Internal Artificial Entanglement

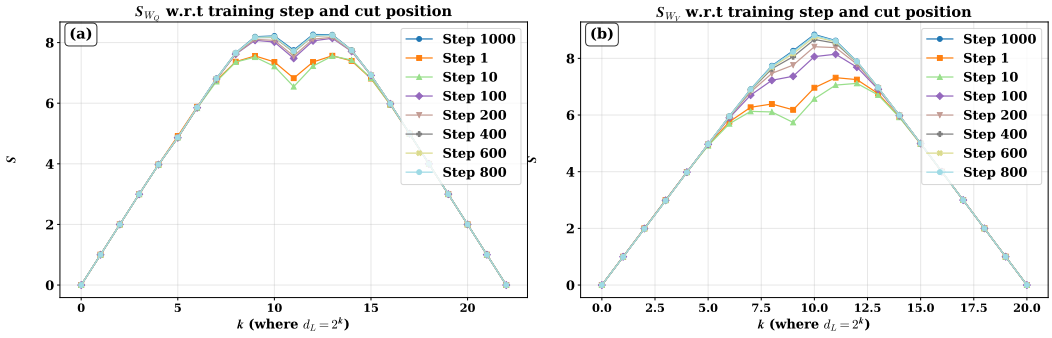

Figure 2: Artificial entanglement profiling of ΔW_Q (a) and ΔW_V (b) across different bi-partition positions k during FFT. Each curve corresponds to a training step, which shows how S evolves and gradually converges as fine-tuning progresses.

Artificial Entanglement Profiling $S_{\Delta W_Q}$ and $S_{\Delta W_V}$ [3] when α is small[4]. We compare the artificial entanglement profiling of FFT and LoRA across bonds, see FIG. 2 and FIG. 3. Both methods exhibit overall **volume-law** scaling of entanglement with a *central dip*. We term it as the *entanglement valley* (A detailed analysis is provided in *Appendix I.2*). However, their entanglement evolution w.r.t time differs substantially: In LoRA with a small scaling parameter $\alpha = 16$, the valley deepens as

[3] For notational convenience, we interchangeably use S_{W_Q} and S_{W_V} to denote $S_{\Delta W_Q}$ and $S_{\Delta W_V}$ respectively.

[4] Here "small" refers to relatively small values of the scaling parameter α (*e.g.*, $\alpha = 16$) compared to larger values used in experiments (e.g., $\alpha = 256$ or higher).

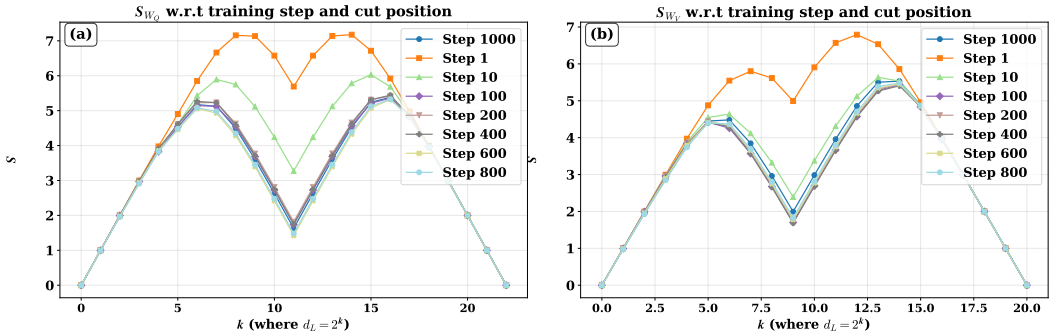

Figure 3: Artificial entanglement profiling of ΔW_Q (a) and ΔW_V (b) during LoRA fine-tuning, similar to FIG. 2.

training progresses, making the valley more pronounced, whereas FFT gradually lifts the valley. This contrast is also type-dependent: in W_V, the valley almost disappears in later FFT steps, while in W_Q it remains clearly visible even at late time.

Remark. From the perspective of quantum information theory, a volume-law entanglement profile signals high correlation across many effective degrees of freedom. The rank of a LoRA update bounds their number, yet the measured valley is far deeper than that bound requires. Therefore, we propose that ***LoRA converges to updates whose spectral weight concentrates on a few modes****, resulting in distinct signatures.*

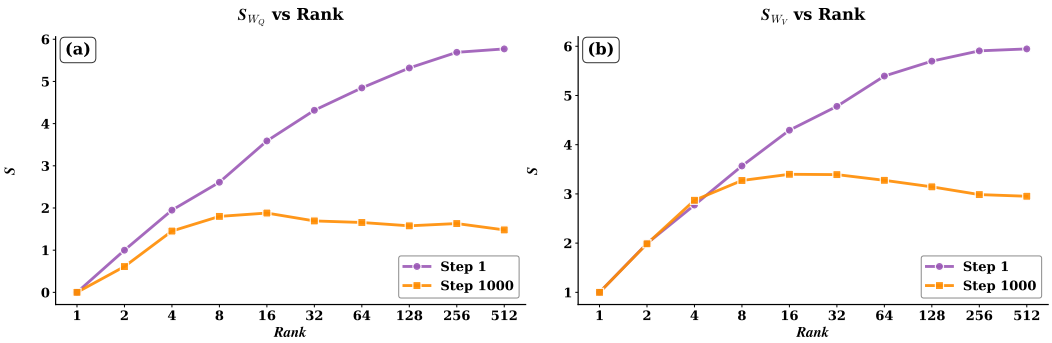

Figure 4: Artificial entanglement profiling of ΔW_Q and ΔW_V as a function of LoRA rank r in different steps. (a) $S_{\Delta W_Q}$ at the early time of fine-tuning (Step 1) and the late time (Step 1000). (b) $S_{\Delta W_V}$ at the early time of fine-tuning (Step 1) and the late time (Step 1000). These reveal how differing r and training time jointly shape the entanglement structure of the learned updates.

To study the relationship between rank settings in LoRA and the deepening of the valley, we examine how the LoRA rank r influences the artificial entanglement profiling of ΔW_Q and ΔW_V. FIG. 4 reports $S_{\Delta W_Q}$ and $S_{\Delta W_V}$ as a function of r at the early time in step 1 and the late time in step 1000. At early time, S increases monotonically with r. This is consistent with viewing the LoRA rank r as an effective bond dimension χ of the MPS factorization: a larger χ enlarges the admissible subspace of tensor network states and raises the maximal entanglement entropy that can be represented across each bond, see FIG. 12. Consequently, as r (and hence the effective χ) grows, the learned updates can exploit higher entanglement, and the measured S rises accordingly. In contrast, at late time S no longer grows with r. It instead quickly approaches a plateau and remains nearly unchanged beyond a certain rank. Heuristically, this suggests that in LoRA, optimization converges to a relatively low-entanglement solution whose intrinsic correlation is already captured at moderate bond dimension, so that additional entanglement capacity from larger r remains largely unused and does not translate into further growth of entanglement entropy. In contrast, FFT does not exhibit such saturation, as it distributes spectral weight over many more modes and thereby reaches higher entanglement. This contrast indicates that **LoRA consistently converges to a concentrated spectrum across different LoRA ranks**.

The concentrated spectrum raises the question of whether the entanglement valley is an artifact of the comparison itself, since FFT updates the full model while LoRA adapts only W_Q and W_V, and since the profiling depends on the chosen tensorization. To rule out these possibilities, we conduct below controlled experiments:

(i) We train Dense and LoRA updates on exactly the same W_Q and W_V projections under one shared learning rate, and profile them over three paired training seeds, see *Appendix J*.

(ii) We swap the singular values and singular vectors between the trained Dense and LoRA updates without retraining, so that each hybrid carries the spectrum of one method and the singular vectors of the other, see *Appendix K*.

(iii) We reprofile the same updates under 21 alternative factor orderings and eight basis transforms, see *Appendix L*.

We first draw some observations: When both methods update the same projections, LoRA retains a pronounced entanglement valley in every module and layer whereas Dense exhibits no two-sided dip at all, with a paired depth difference of 3.5–4.2 bits at step 1000 that stays positive at every saved step. Moreover, the valley in LoRA deepens as training progresses while the Dense depth stays near or below zero, reproducing under matched scope the time-dependent contrast between FIG. 3 and FIG. 2. Transplanting the LoRA spectrum onto the Dense singular vectors closes 97.7% of the profile gap and 98.0% of the depth gap, while the reverse transplant closes essentially none of either. Notably, all four matrices in the swap are truncated to the same nominal rank of 256, so the rank limit is held fixed and the difference between the two transplants isolates the distribution of singular values alone. Under every tested ordering and basis, LoRA keeps a two-sided valley and Dense stays below zero, so the sign of the contrast never reverses. Combining these results we can see that **the entanglement valley is carried by the concentrated singular values of the LoRA update rather than by its singular vectors, its adaptation scope, or the tensorization we adopt**.

We can further observe the empirical relation between how r impacts the test loss (See FIG. 13) and the perspective from S (FIG. 4). In FIG. 13 we visualize the test loss with respect to different r and training steps, and observe that increasing r does not necessarily indicate a large improvement in the test performance, along with the saturation of S (FIG. 4). Furthermore, investigating the best learning rate for FFT and LoRA (See FIG. 14), we observe a similar phenomenon as (Biderman et al., 2024; Schulman & Lab, 2025). From FIG. 14, LoRA requires a larger learning rate to achieve the best test loss than FFT. These observations motivate us to further investigate how hyper-parameters, where we specifically investigate the scaling parameter α, affects the entanglement structure and optimization behavior in LoRA.

Artificial Entanglement Profiling of $S_{\Delta W_Q}$ and $S_{\Delta W_V}$ when α is large. We now investigate the case when α is increased. We first tune α from 16 to 256. From FIG. 15 we see that in this setting LoRA will learn a solution that mitigates the deepening of the entanglement valley in ΔW_Q and even lifts the valley in ΔW_V during late time, suggesting a closer yet still distinct solution to the one learned by FFT. To better characterize this phenomenon in terms of both S, learning rates and final test performance, we conduct below experiments:

(i) We visualize the entanglement entropy S in the middle bi-partition position with respect to the training steps (Specifically evaluating at the early time in step 1 and the late time in step 1000 respectively) and α, see FIG. 16;

(ii) We visualize how α changes the final test loss while keeping all other hyper-parameters fixed, see FIG. 17.

(iii) We sweep the learning rates by setting $\alpha = \{16, 256, 512, 1024, 4096, 8192\}$, where the α is set higher or equal to the one in FIG. 14, see FIG. 18 (a).

We first draw some observations: As α increases from 1, test performance is initially improved. Then, the test performance reaches the best when α is around $\alpha = 16$. However, the test performance is worse as α increases. Meanwhile, from $\alpha = 1$ to $\alpha = 256$ we can observe an increasing fluctuation along the path. This indicates that the optimal learning rate in LoRA has been shifted, see FIG. 18 (a): the learning rates corresponding to the best performance in LoRA and FFT are within less gap, suggesting more aligned fine-tuning solutions. And the optimal learning rates are not impacted by a relatively small rank, see FIG. 18 (b). Combining these results we can see that the artificial

entanglement signatures are actually sensitive to hyper-parameter settings, including α (FIG. 16), time (FIG. 3) and rank (FIG. 4). Besides, **entanglement signatures might serve as an indicator of the combined effect of multiple hyper-parameter choices** (See detailed discussion in Sec. 6).

While the above analysis indicates that LoRA and FFT exhibit different artificial entanglement profiles, exhibiting different fine-tuning solutions (*e.g.*, the updates of projection matrices and the optimal learning rate), a puzzle still remains: *Why are the resulting test losses still approximate even when the artificial entanglement profiles are different?* When only focusing on LoRA, we can formulate this problem as: *Why different settings of hyper-parameters that result in different artificial entanglement profiles in LoRA can still result in similar test loss (e.g., see FIG. 17, the final test loss is approximately the same when $\alpha = 1$ and $\alpha = 64$, ignoring the fluctuations)?*

5.2 External Artificial Entanglement

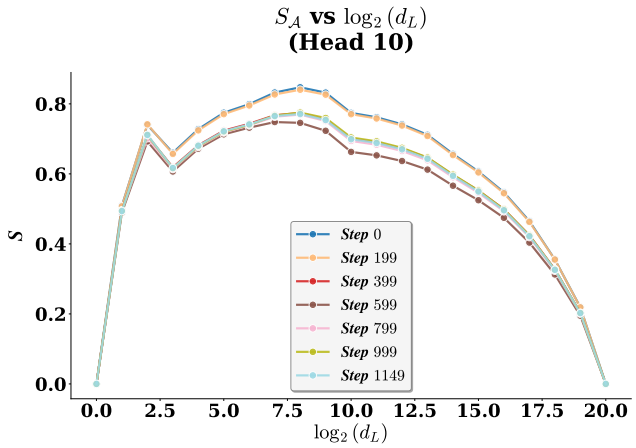

Figure 5: Artificial Entanglement Profiling of the attention matrix (Head 10) S_A across training steps. S_A remains uniformly low throughout training and exhibits a characteristic of an approximate area-law scaling, where the entropy grows only mildly that roughly following a shallow logarithmic rise. This indicates that the attention matrix operates in a strongly low-entanglement regime.

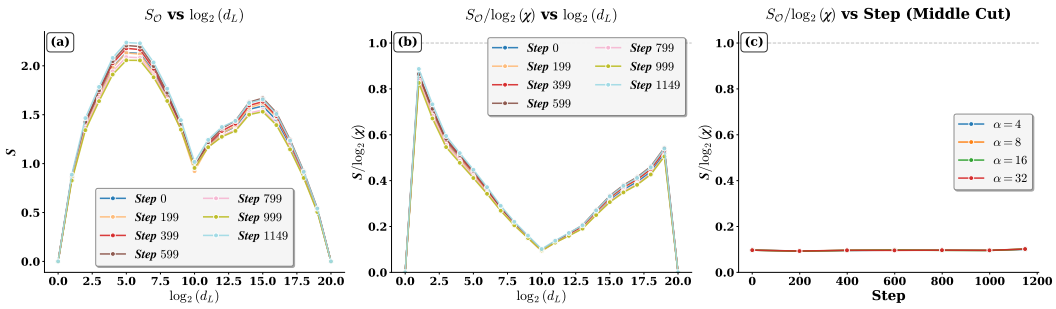

Figure 6: Entanglement entropy $S_\mathcal{O}$ of the output operator $\mathcal{O} = XX^\top$ across bi-partition positions and training steps. Panels (a) and (b) show that $S_\mathcal{O}$ remains significantly below the theoretical maximum. Panel (c) illustrates that this behavior is invariant across the variation of scaling coefficient α and training steps, indicating a no-hair like behavior as well.

To explore this problem, we extend our artificial entanglement analysis framework from only the updates of W_Q and W_V (*i.e.*, ΔW_Q and ΔW_V) to attention matrix $A(X_0)$ and attention outputs X, specifically the XX^T. This choice is natural because $XX^\top$ captures the pairwise correlations between token representations after attention mixing, analogous to how correlation is probed in quantum many-body systems via reduced density matrices (*i.e.*, $XX^\top$ serves as the analogue of an output-state density operator).

Artificial Entanglement Profiling of S_A and S_{XX^T}. Generally, for an input $\mathcal{X}_0 \in \mathbb{R}^{B\times T\times d_{\text{model}}}$, we have the attention tensor $\mathcal{A}(\mathcal{X}_0) \in \mathbb{R}^{B\times H\times T\times T}$ where B is the batch size, T is the number of tokens, d_{model} is the token embedding dimension and H is the number of heads. Here we select samples where the sequence length T can be directly factorized by 2 (*e.g.*, $T = 1024$). Then, we focus on each attention matrix in separate attention head. Therefore, for each head, we obtain a matrix $A \in \mathbb{R}^{T\times T}$ that can be formalized to an MPS to profile the artificial entanglement. Here we show an entanglement profiling specifically in Head 10 (More heads are shown in normalized form in FIG. 19), see FIG. 5. Across all training steps, the entanglement entropy S_A remains uniformly low. Such a profile is consistent with an approximate **area law scaling**: S exhibits a shallow, nearly logarithmic rise as the bi-partitioning position moving (in this case from right to left across the MPS, followed by a saturation around the central partition). Though not a strict area law scaling, the profiling deviates much from volume law and can often be treated as area law with logarithmic correction (Eisert et al., 2008). This behavior indicates that the correlations encoded by this attention head are highly localized and remains far from the high-entanglement (volume-law) regime. FIG. 19 shows that the phenomenon is prevalent across multiple heads through normalizing the S via dividing it by the theoretical allowed entanglement $\log(\chi)$, where $\chi = \min(d_L, d_R)$. Notably, χ changes across different cuts, so as the theoretical allowed entanglement $\log(\chi)$. Specifically, χ reaches its maximum at the central partition (where $d_L = d_R$), making the theoretical allowed entanglement largest at the middle cuts. This variation in the normalization factor makes the dip in the middle of the normalized entropy curves more pronounced. Besides, a **"no hair" property exhibits:** the entanglement structures in attention are largely insensitive with respect to training steps and α, see FIG. 20. This property persists when the profiling is extended from a single head to all 32 heads on 100 nontraining inputs kept at their actual length without padding, where the attention entropies of checkpoints trained under different parameterizations agree to within hundredths of a bit at matched loss, see *Appendix M*. In other words, even though different hyper-parameters or fine-tuning methods lead to different artificial entanglement profiling (*e.g.*, FIG. 2, FIG. 3, FIG. 16), from the resulting attention matrices those entanglement signature variations do not exhibit. Similar phenomena can be observed in the attention output operator $\mathcal{O} = XX^T$, see FIG. 6. We suggest two reasons might prompt these phenomena: First, the entanglement between text tokens is significantly low; Second, masking operation results in many values to be zero after the *softmax* operation due to the auto-regressive nature, further resulting in a relatively low entanglement structure. Regarding the second, we conduct an experiment to see how masking operation influences the results, see *Appendix I.1* for details.

Artificial Entanglement Profiling in MPS Adaptation. The MPS structure adopted above as an analysis tool can also serve directly as a PEFT parameterization, in which the low-rank factor of the update is represented as an MPS with bond dimension χ, termed *MPS adaptation*, see *Appendix G*. Combining the results, ΔW_Q and ΔW_V under MPS adaptation exhibit the same characteristic entanglement valley as observed in LoRA, the attention matrices retain the no-hair property, and increasing α again shifts the optimal learning rate to smaller values, confirming that these behaviors are not specific to the LoRA parameterization.

6 Conclusions

By quantifying *artificial entanglement* through MPS modeling, we show that LoRA and FFT induce **volume-law** internal entanglement profiles in the projection matrices ΔW_Q and ΔW_V with a distinctive "entanglement valley", whose depth is governed by the scaling parameter α and which separates the internal structure learned by LoRA from that of FFT. In contrast, the attention matrices exhibit an approximate **area-law** scaling, and the *external* attention outputs remain remarkably invariant despite these *internal* differences. Drawing an analogy to the No-Hair Theorem in black hole physics, we propose that the attention mechanism acts as an effective coarse-graining operator that suppresses redundant correlations, which explains why LoRA performs comparably to FFT with far fewer parameters. Our random matrix analysis and the MPS Adaptation extension further indicate that this no-hair behavior is not an artifact of LoRA but a more universal property. Together, our findings suggest that quantum-inspired analysis may offer a new foundation for the mechanistic interpretability of large language models, and open directions such as probing how entanglement evolves during pretraining and whether entanglement-based regularization can guide more efficient adaptation.

AI USE STATEMENT

In this work, we used generative AI tools for research ideation and for designing and providing feedback on the experimental methodology, for implementing methods in the experimental code, and for interpreting results in the data analysis. Additionally, we used generative AI tools for drafting and restructuring parts of the manuscript, editing for readability, identifying and summarizing relevant literature, and creating and editing analysis code. AI-generated suggestions were not treated as independent scientific evidence. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study collects no new human-subject data. It analyzes pretrained language models and existing instruction-tuning datasets and therefore inherits the limitations, biases, licensing conditions, and potential downstream risks of those resources. The term "artificial entanglement" denotes a classical mathematical diagnostic; it is not evidence that language models exhibit physical quantum behavior.

REPRODUCIBILITY STATEMENT

Appendix C specifies the tensorization, entropy computation, models, datasets, and optimization settings behind the profiling in the main text. *Appendix J* through *Appendix L* describe the matched-scope comparison, the singular-value transplant, and the factor-ordering and basis analyses, and *Appendix M* the attention measurements over all heads on unpadded inputs. *Appendix G* gives the MPS adaptation method, and *Appendix I* the additional models and datasets. Numerical summaries, per-seed results, and plotting scripts accompany the manuscript. *Appendix E* and *Appendix F* state the theoretical assumptions and give full proofs.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Luigi Amico, Rosario Fazio, Andreas Osterloh, and Vlatko Vedral. Entanglement in many-body systems. *Reviews of modern physics*, 80(2):517–576, 2008.

Afia Anjum, Maksim E Eren, Ismael Boureima, Boian Alexandrov, and Manish Bhattarai. Tensor train low-rank approximation (tt-lora): Democratizing ai with accelerated llms. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pp. 583–590. IEEE, 2024.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Aleksandr Berezutskii, Minzhao Liu, Atithi Acharya, Roman Ellerbrock, Johnnie Gray, Reza Haghshenas, Zichang He, Abid Khan, Viacheslav Kuzmin, Dmitry Lyakh, et al. Tensor networks for quantum computing. *Nature Reviews Physics*, 7(10):581–593, 2025.

Sourav Bhattacharya and Amitabha Lahiri. No hair theorems for positive\lambda. *arXiv preprint gr-qc/0702006*, 2007.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*, 2024.

Luca Bombelli, Rabinder K Koul, Joohan Lee, and Rafael D Sorkin. Quantum source of entropy for black holes. *Physical Review D*, 34(2):373, 1986.

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.

Pasquale Calabrese and John Cardy. Entanglement entropy and quantum field theory. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(06):P06002, 2004.

Pasquale Calabrese and John Cardy. Entanglement entropy and conformal field theory. *Journal of physics a: mathematical and theoretical*, 42(50):504005, 2009.

Brandon Carter. Axisymmetric black hole has only two degrees of freedom. *Physical Review Letters*, 26(6):331, 1971.

Min Chen, Minzhao Liu, Changhun Oh, Liang Jiang, Yuri Alexeev, and Junyu Liu. Towards symmetry-aware efficient simulation of quantum systems and beyond. *arXiv e-prints*, pp. arXiv–2303, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:10088–10115, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Jens Eisert, Marcus Cramer, and Martin B Plenio. Area laws for the entanglement entropy-a review. *arXiv preprint arXiv:0808.3773*, 2008.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

Daniel Harlow. Jerusalem lectures on black holes and quantum information. *Reviews of Modern Physics*, 88(1):015002, 2016.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models. In *International Conference on Machine Learning (ICML)*, pp. 17783–17806. PMLR, 2024.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Werner Israel. Event horizons in static vacuum space-times. *Physical review*, 164(5):1776, 1967.

Werner Israel. Event horizons in static electrovac space-times. *Communications in Mathematical Physics*, 8(3):245–260, 1968.

Damjan Kalajdjieski. Rank-stabilized lora: Unlocking the potential of lora fine-tuning. *arXiv preprint arXiv:2312.03732*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Dawid J Kopiczko, Tijmen Blankevoort, and Yuki M Asano. Vera: Vector-based random matrix adaptation. In *International Conference on Learning Representations (ICLR)*, 2024.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations (ICLR)*, 2018.

Vladislav Lialin, Namrata Shivagunde, Sherin Muckatira, and Anna Rumshisky. Relora: High-rank training through low-rank updates. *arXiv preprint arXiv:2307.05695*, 2023.

Minzhao Liu, Changhun Oh, Junyu Liu, Liang Jiang, and Yuri Alexeev. Simulating lossy gaussian boson sampling with matrix-product operators. *Physical Review A*, 108(5):052604, 2023.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

NE Mavromatos. Eluding the no-hair conjecture for black holes. *arXiv preprint gr-qc/9606008*, 1996.

Francesco Mezzadri. How to generate random matrices from the classical compact groups. *arXiv preprint math-ph/0609050*, 2006.

Ashley Milsted, Junyu Liu, John Preskill, and Guifre Vidal. Collisions of false-vacuum bubble walls in a quantum spin chain. *PRX Quantum*, 3(2):020316, 2022.

Román Orús. Tensor networks for complex quantum systems. *Nature Reviews Physics*, 1(9):538–550, 2019.

Don N Page. Average entropy of a subsystem. *Physical review letters*, 71(9):1291, 1993.

Yue Pan, Yijie Zhou, Feng Wu, and Zhen Zhang. Lora training in the ntk regime has no spurious local minima. *arXiv preprint arXiv:2402.11867*, 2024.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. In *OpenAI blog*, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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.

David C Robinson. Uniqueness of the kerr black hole. *Physical Review Letters*, 34(14):905, 1975.

Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers. In *Forty-second International Conference on Machine Learning*, 2024.

John Schulman and Thinking Machines Lab. Lora without regret. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250929. https://thinkingmachines.ai/blog/lora/.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. Lora vs full fine-tuning: An illusion of equivalence. *arXiv preprint arXiv:2410.21228*, 2024.

Mark Srednicki. Entropy and area. *Physical Review Letters*, 71(5):666, 1993.

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

Robert C Thompson. The behavior of eigenvalues and singular values under perturbations of restricted rank. *Linear Algebra and its Applications*, 13(1-2):69–78, 1976.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. Dylora: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 3274–3287, 2023.

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.

Guifré Vidal. Efficient classical simulation of slightly entangled quantum computations. *Physical review letters*, 91(14):147902, 2003.

Guifré Vidal. Efficient simulation of one-dimensional quantum many-body systems. *Physical review letters*, 93(4):040502, 2004.

Guifre Vidal. Entanglement renormalization. *Physical review letters*, 99(22):220405, 2007.

Guifré Vidal and Reinhard F Werner. Computable measure of entanglement. *Physical Review A*, 65 (3):032314, 2002.

Guifre Vidal, José Ignacio Latorre, Enrique Rico, and Alexei Kitaev. Entanglement in quantum critical phenomena. *Physical review letters*, 90(22):227902, 2003.

Longteng Zhang, Lin Zhang, Shaohuai Shi, Xiaowen Chu, and Bo Li. Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning. *arXiv preprint arXiv:2308.03303*, 2023a.

Mingyang Zhang, Hao Chen, Chunhua Shen, Zhen Yang, Linlin Ou, Xinyi Yu, and Bohan Zhuang. Loraprune: Structured pruning meets low-rank parameter-efficient fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3013–3026, 2024.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2023b.

Bojia Zi, Xianbiao Qi, Lingzhi Wang, Jianan Wang, Kam-Fai Wong, and Lei Zhang. Delta-lora: Fine-tuning high-rank parameters with the delta of low-rank matrices. *arXiv preprint arXiv:2309.02411*, 2023.

A NOTATIONS

Table 1 provides a list of notations.

Table 1: Notations

Symbol	Description
W_0	Pre-trained weight projection matrix.
ΔW	Weight update matrix.
ΔW_{LoRA}	LoRA weight update: $\Delta W_{\text{LoRA}} = \frac{\alpha}{r} BA$.
ΔW_{Full}	Full fine-tuning weight update.
W_Q, W_K, W_V	Query, key, value projection matrices.
W^O	Output projection matrix.
A, B	Low-rank matrices in LoRA decomposition ($A \in \mathbb{R}^{r \times d_{\text{in}}}$, $B \in \mathbb{R}^{d_{\text{out}} \times r}$).
A	Attention matrix (context-dependent).
$\mathcal{O} = XX^\top$	Attention output operator.
$d_{\text{in}}, d_{\text{out}}$	Input and output dimensions of weight projection matrices.
d_{model}	Token embedding dimension (model dimension).
d_k	Dimension of each attention head.
$d_L^{(k)}, d_R^{(k)}$	Dimensions of left and right subsystems at cut k.
$d_{\min}, d_{\max}$	Minimum and maximum dimensions ($d_{\min} \leq d_{\max}$).
d_k	Local dimension at site k (context-dependent).
T	Sequence length (number of tokens).
H	Number of attention heads.
B	Batch size.
M, N	Order of tensors (number of indices).
n, m	Number of sites in MPS factorization.
f_k, g_ℓ	Prime factors in factorization of d_{out} and d_{in}.
r	LoRA rank ($r \ll \min(d_{\text{in}}, d_{\text{out}})$).
α	LoRA scaling hyperparameter.
χ, χ_k	Bond dimension (at bond k).
$\chi_{\max}$	Maximum bond dimension (truncation threshold).
χ	Maximum bond dimension: $\chi := \max_k \chi_k$.
$\mathcal{T}^{[k]}$	k-th tensor in MPS decomposition.
$\mathcal{A}^{[k]}, \mathcal{B}^{[k]}$	MPS tensors for LoRA matrices A and B.
α_k	Virtual bond index (Schmidt index).
i_k	Physical index at site k.
$C^{(k)}$	Matrix at cut k in MPS construction.
$U^{(k)}, S^{(k)}, V^{(k)}$	SVD factors at cut k.
S, S_k	Von Neumann entanglement entropy (at cut k).
S_2	Rényi-2 entropy.
S_α	Rényi-α entropy.
$\rho_L^{(k)}$	Reduced density matrix of left subsystem at cut k.
$\lambda_{\alpha_k}, s_i^{(k)}$	Singular values (Schmidt coefficients).
$p_i^{(k)}$	Eigenvalues of reduced density matrix.
σ	Bulk spectrum spread parameter.
$\mathcal{C}_{\text{attn}}$	Effective Attention Charge: $\mathcal{C}_{\text{attn}} = \frac{\sigma^2}{1+\sigma^2}$.
r_{stable}	Stable rank: $r_{\text{stable}} = \frac{\text{Tr}(\Sigma^2)}{\lambda_1^2}$.
η	Stable rank excess: $\eta = r_{\text{stable}} - 1$.
λ_i	Eigenvalues (context-dependent).
$x_i = d_{\min} \cdot \lambda_i$	Normalized eigenvalues for MP distribution.
$\rho(x)$	Marchenko-Pastur density function.
c	Aspect ratio: $c = d_{\min}/d_{\max}$.
$\lvert\Psi\rangle$	Many-body quantum state.
$\lvert i_L\rangle, \lvert i_R\rangle$	Basis states for left and right subsystems.

$\mathcal{H}_L, \mathcal{H}_R$	Left and right Hilbert spaces.
$\mathcal{H}_A, \mathcal{H}_B$	Subsystem Hilbert spaces.
$\mathcal{X}_0$	Input tensor: $\mathcal{X}_0 \in \mathbb{R}^{B\times T\times d_{\text{model}}}$.
$\mathcal{A}(\mathcal{X}_0)$	Attention tensor: $\mathcal{A}(\mathcal{X}_0) \in \mathbb{R}^{B\times H\times T\times T}$.
Q_h, K_h, V_h	Query, key, value matrices for head h.
$A^{\perp}$	Centered attention matrix: $A^{\perp} = A - \frac{1}{T}\mathbf{1}\mathbf{1}^{\top}$.
X	Attention layer output: $X = AV$.
$\mathbb{R}, \mathbb{C}$	Real and complex number fields.
$\|\cdot\|_F$	Frobenius norm.
$\|\cdot\|_{\text{op}}$	Operator norm (largest singular value).
$\text{Tr}(\cdot)$	Trace operator.
$\text{vec}(\cdot)$	Vectorization operator.
$o(1)$	Asymptotically vanishing term.

B Detailed Tensor and Attention Background

Tensor, Tensor Network, and Matrix Product State (MPS). Tensors generalize matrices to multi-index arrays and provide a unified language for representing high-dimensional structures. An Order-N tensor $\mathcal{A}_{i_1\ldots i_N}$ has N indices, where each $i_k \in \{0, 1, \ldots, d_k - 1\}$ ranges over a local dimension d_k. Graphically as shown in FIG. 7 (a), one can visualize a tensor as a node with N legs, each leg corresponding to an index of the tensor. A fundamental operation between tensors is *tensor contraction*, which corresponds to summing over a pair of shared indices. In FIG. 7 (b), contraction is depicted by connecting the legs associated with the contracted dimensions. More elaborate contraction patterns naturally give rise to *tensor networks*, which provide structured factorization of large tensors into networks of smaller building blocks. Among various tensor network architectures, the *Matrix Product State* (MPS) is the most widely used one-dimensional tensor network. An MPS expresses a high-order tensor as a chain of Order-3 tensors connected by virtual bonds. In FIG. 7 (c), each site emits one physical leg (the physical index) and two virtual legs linking to its neighbors, forming a one-dimensional chain as depicted in the figure. Consider a general M-body pure quantum state on a local Hilbert space of dimension d:

$$|\Psi\rangle = \sum_{i_1,\ldots,i_M=0}^{d-1} c_{i_1,\ldots,i_M} |i_1, \ldots, i_M\rangle, \tag{7}$$

where $c_{i_1\ldots i_M}$ is an amplitude tensor of Order-M whose storage cost grows exponentially with M. The MPS representation approximates this tensor by factorizing it into a sequence of Order-3 tensors:

$$c_{i_1,\ldots,i_M} = \sum_{\alpha_1=1}^{\chi_1} \cdots \sum_{\alpha_{M-1}=1}^{\chi_{M-1}} \mathcal{T}^{[1]\,i_1}_{1,\alpha_1} \mathcal{T}^{[2]\,i_2}_{\alpha_1,\alpha_2} \cdots \mathcal{T}^{[M]\,i_M}_{\alpha_{M-1},1}, \tag{8}$$

where each $\mathcal{T}^{[k]}$ is an Order-3 tensor with one physical index i_k and two virtual indices α_{k-1}, α_k. The virtual indices are contracted, yielding the chain-like network structure. The parameter χ_k denotes the bond dimension at bond k, which may vary along the chain. The quantity $\chi := \max_k \chi_k$ is often referred to as the (maximum) *bond dimension*, which determines the expressive power of the MPS and controls the entanglement it can capture: larger χ allows more accurate approximations of c, while small χ corresponds to a compressed representation. In quantum physics, the effectiveness of the MPS formalism is deeply tied to the entanglement structure of the underlying state. For example, ground states of gapped one-dimensional systems satisfy an entanglement area law, which implies that they can be faithfully approximated by an MPS with modest bond dimension. This makes MPS not only a powerful conceptual framework but also a practical computational ansatz for representing high-dimensional quantum states with polynomial resources.

Forming an MPS from a high-order tensor. As depicted in FIG. 7 (c), an MPS representation can be systematically constructed from a high-order tensor by a sequence of reshaping and singular value decompositions (SVDs). Consider an Order-M tensor $c_{i_1 i_2 \ldots i_M}$, which may be viewed as the probability amplitude tensor of an M-body quantum state. The construction proceeds iteratively from left to right. In the first step, we group the first index i_1 into a "left" subsystem and the remaining

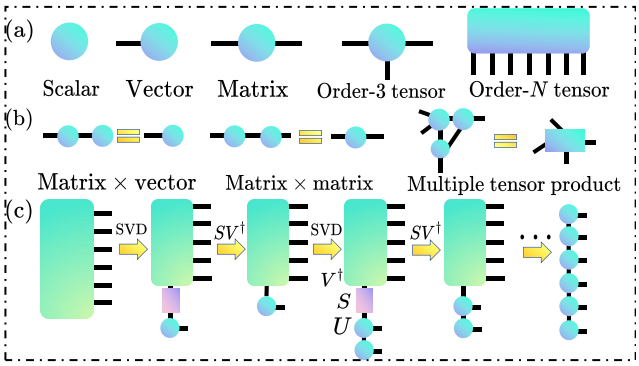

Figure 7: Overview of Tensor Diagrammatic Notation. (a) Visual representations of a scalar, vector, matrix, and an arbitrary order-N tensor. (b) Tensor contractions illustrating matrix multiplication and multiple tensor products. (c) The process of modeling an order-N tensor as a Matrix Product State.

indices $(i_2, \ldots, i_M)$ into a "right" subsystem, reshaping the tensor into a matrix

$$C^{(1)}_{i_1,\,(i_2 \ldots i_M)} \in \mathbb{C}^{d \times d^{M-1}}. \tag{9}$$

Performing a singular value decomposition,

$$C^{(1)} = U^{(1)} S^{(1)} V^{(1)\dagger}, \tag{10}$$

we absorb $U^{(1)}$ into the first MPS tensor $\mathcal{T}^{[1]}$, while the product $S^{(1)}V^{(1)\dagger}$ is passed to the next step. Explicitly,

$$\mathcal{T}^{[1]\,i_1}_{\alpha_0=1,\alpha_1} := U^{(1)}_{i_1,\alpha_1}, \tag{11}$$

where α_1 labels the singular vectors and defines the first virtual bond.

In the second step, the remaining object $S^{(1)}V^{(1)\dagger}$ is reshaped into a matrix by grouping (α_1, i_2) as the left index and $(i_3, \ldots, i_M)$ as the right index. Applying another SVD yields

$$C^{(2)}_{(\alpha_1 i_2),\,(i_3 \ldots i_M)} = U^{(2)} S^{(2)} V^{(2)\dagger}, \tag{12}$$

from which we define the second MPS tensor

$$\mathcal{T}^{[2]\,i_2}_{\alpha_1\alpha_2} := U^{(2)}_{(\alpha_1 i_2),\alpha_2}. \tag{13}$$

This procedure is repeated sequentially, each time isolating one physical index i_k and generating a new virtual bond α_k, until the final site is reached.

After $M-1$ such decompositions, the original tensor is factorized as

$$c_{i_1,\ldots,i_M} = \sum_{\alpha_1,\ldots,\alpha_{M-1}} \mathcal{T}^{[1]\,i_1}_{1,\alpha_1} \mathcal{T}^{[2]\,i_2}_{\alpha_1,\alpha_2} \cdots \mathcal{T}^{[M]\,i_M}_{\alpha_{M-1},1}, \tag{14}$$

which is precisely the MPS form. Each tensor $\mathcal{T}^{[k]}$ carries one physical index i_k and two virtual indices α_{k-1}, α_k, forming a one-dimensional chain as illustrated in the figure. In practice, the singular values in each $S^{(k)}$ may be truncated, retaining only the largest χ_k values. This truncation controls the bond dimension at each bond and yields an efficient compressed representation. Crucially, the maximum bond dimension required to accurately represent the state is directly related to its entanglement across bi-partitions.

Self Attention in Large Language Models. Generally, for an input $\mathcal{X}_0 \in \mathbb{R}^{B\times T\times d_{\text{model}}}$, the model employs Multi-Head Attention to capture diverse contextual dependencies. The mechanism utilizes projection matrices $W^Q, W^K, W^V \in \mathbb{R}^{d_{\text{model}}\times(Hd_k)}$ to generate query, key, and value representations, where d_k is the dimension of each head. These projected features are split into H parallel heads. Let Q_h, K_h, V_h denote the segments corresponding to head h. In modern LLMs such as LLaMA, the query and key vectors are typically augmented with *Rotary Position Embedding (RoPE)* (Su et al., 2024), which encodes positional information by applying rotational transformations

to the query and key vectors. This allows the model to capture relative positional relationships effectively. After applying RoPE, the attention weights are computed via scaled dot-product:

$$\mathcal{A}(\mathcal{X}_0)_h = \text{Softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_k}}\right) \in \mathbb{R}^{B\times T\times T} \tag{15}$$

The attention output for each head is then derived as $Head_h = \mathcal{A}(\mathcal{X}_0)_h V_h \in \mathbb{R}^{B\times T\times d_k}$. Finally, all heads are concatenated and projected by an output matrix $W^O \in \mathbb{R}^{(Hd_k)\times d_{\text{model}}}$ to form the final layer output.

C Detailed Methods

This section details the methodology for artificial entanglement analysis, especially we establish an analogy between the MPS factorization of matrices and the MPS representation of quantum many-body states: we here detail the interpretation of the MPS factorization of matrices as formally analogous to the amplitude tensor of a quantum many-body state $|\Psi\rangle$. We emphasize that the underlying systems are entirely classical: this is purely a mathematical analogy that provides a principled framework for quantifying entanglement-like correlations via Schmidt decomposition. See FIG. 8 and preliminaries of tensor analysis in *Appendix B*.

For a quantum state $|\Psi\rangle$ partitioned into two subsystems A and B via Schmidt decomposition:

$$|\Psi\rangle = \sum_i \lambda_i |i_A\rangle |i_B\rangle, \tag{16}$$

the entanglement entropy S is calculated as:

$$S_A = S_B = -\sum_i \lambda_i^2 \log \lambda_i^2 \tag{17}$$

For an MPS, the two subsystems are obtained by specifying a specific *cut* in the MPS chain, where we investigate the entanglement between the left and right subsystems after the bi-partition. We apply this framework to relevant matrices, including the updates of projection matrices ΔW_Q and ΔW_V, attention matrices A, and attention output operators $\mathcal{O} = XX^T$, by formalizing each matrix as an MPS.

As an example, we focus on one of the projection matrices as ΔW. While $\Delta W \in \mathbb{C}^{d_{\text{out}}\times d_{\text{in}}}$ is originally an Order-2 tensor, we reshape it into a higher-order tensor by factorizing its input and output dimensions into their prime components. Specifically, we iteratively divide each dimension by its smallest prime factor until the value reduces to one:

$$\begin{aligned} d_{\text{out}} &= \prod_{k=1}^{n} f_k, \qquad f_1 \le f_2 \le \cdots \le f_n, \\ d_{\text{in}} &= \prod_{\ell=1}^{m} g_\ell, \qquad g_1 \le g_2 \le \cdots \le g_m, \end{aligned} \tag{18}$$

where f_k and g_ℓ denote the prime factors sorted in non-decreasing order, and n and m denote the number of prime factors in d_{out} and d_{in}, respectively. This decomposition yields the finest-grained tensorization of the input and output spaces and provides a natural way to reshape ΔW into an Order-$(n+m)$ tensor amenable to tensor-network factorizations, including the MPS.

For LoRA, we denote $\mathcal{A}_i$ and $\mathcal{B}_i$ to be each local Order-3 tensor in the MPS on the same rank space $\mathcal{H}_r$, therefore ΔW_{LoRA} can be viewed as the probability amplitude tensor of a quantum wave function represented below:

$$\begin{aligned} |\Psi\rangle_{\text{LoRA}} &= \Delta W_{\text{LoRA}} |o_1 \cdots o_m\rangle_{\text{out}} \otimes |i_1 \cdots i_n\rangle_{\text{in}} \\ &= \sum_{\alpha_1,\ldots,\alpha_n}^{\chi} \mathcal{B}^{[1]d_{o_1}}_{\alpha_0=1,\alpha_1} \mathcal{B}^{[2]d_{o_2}}_{\alpha_1,\alpha_2} \ldots, \mathcal{A}^{[n+1]d_{i_1}}_{\alpha_{n,n+1}} \mathcal{A}^{[n+m]d_{i_m}}_{\alpha_{n+m-1},\,\alpha_{n+m}=1} \cdot \\ &\quad |o_1 \cdots o_m\rangle_{\text{out}} \otimes |i_1 \cdots i_n\rangle_{\text{in}}. \end{aligned} \tag{19}$$

Although the above formulation is written in terms of the decomposed factors $\mathcal{A}_i$ and $\mathcal{B}_i$, in practice we directly decompose the equivalent ΔW_{LoRA} itself into the MPS representation. For FFT, the expression is the similar, we express

$$
\begin{aligned}
&|\Psi\rangle_{\text{FFT}} = \Delta W_{\text{FFT}} |w_1 \cdots w_{m+n}\rangle \\
&= \sum_{\alpha_1,\cdots,\alpha_n}^{\chi} \mathcal{W}^{[1]d_{o_1}}_{\alpha_0=1,\alpha_1} \cdots \mathcal{W}^{[n+1]d_{i_1}}_{\alpha_{n,n+1}} \cdots \mathcal{W}^{[n+m]d_{i_m}}_{\alpha_{n+m-1},\alpha_{n+m}=1} \cdot \\
&|w_1 \cdots w_{m+n}\rangle.
\end{aligned} \tag{20}
$$

See *Appendix B* for more details about the formulation based on the factorization.

Further, we calculate the von Neumann entanglement entropy by performing an SVD at each bond during the MPS decomposition. At bond k (the k-th bond in the MPS chain), we obtain the singular values $\{\lambda_{\alpha_k}\}$, where α_k indexes the singular values (Schmidt indices) at bond k. The SVD is written in terms of each bipartite quantum state $|\Psi\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B$ component Ψ_{ij} below:

$$
\Psi_{ij} = \sum_{\alpha,\beta=1}^{\chi} U_{i\alpha}\, \Lambda_{\alpha\beta}\, (V^\dagger)_{\beta j} = \sum_{\alpha=1}^{\chi} U_{i\alpha}\, \lambda_\alpha\, (V^\dagger)_{\alpha j}, \tag{21}
$$

with at most χ non-zero singular values. The bi-partition at bond k corresponds to $\mathcal{H}_A = \mathcal{H}_{1:k}$ and $\mathcal{H}_B = \mathcal{H}_{k+1:(n+m)}$. Since the reduced density matrices $\rho_{1:k}$ and $\rho_{k+1:(n+m)}$ share the same non-zero spectrum, the bipartite entanglement entropy at bond k is

$$
S_k = S(\rho_{1:k}) = S(\rho_{k+1:(n+m)}) = -\sum_{\alpha_k=1}^{\chi} \lambda^2_{\alpha_k} \log\!\left(\lambda^2_{\alpha_k}\right), \tag{22}
$$

where S_k quantifies the entanglement entropy at bond k. We can calculate S_k for each bond k, and therefore we can obtain an S_k curve (artificial entanglement profile) with respect to the bond position.

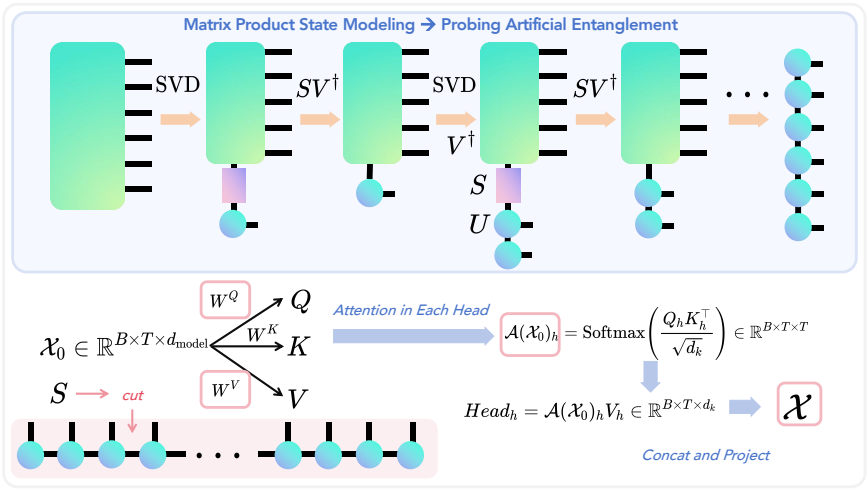

Figure 8: Overall framework for artificial entanglement analysis. (i) Reshaping matrices including the updates of weight projection matrices such as ΔW_Q and ΔW_V from FFT or LoRA, and attention-related matrices into higher-order tensors by factorizing input and output dimensions into prime components; (ii) Performing a sequence of SVDs to decompose the tensor into a chain of Order-3 tensors connected by virtual bonds, where virtual bond indices capture correlations between sites; (iii) Interpreting the MPS factorization as a mathematical formalism of many-body states, formally analogous to the MPS representation of the amplitude tensor of a quantum many-body state; (iv) Computing the von Neumann entanglement entropy (S) at each bond, yielding an artificial entanglement profile.

D Mathematical Foundations: Quantum Entanglement and Entropy Measures

This section establishes the mathematical foundations for computing and analyzing entanglement entropy in the context of MPS-decomposed weight projection matrices. We first explain how to compute entanglement entropy at each cut in the MPS construction, connecting the SVD procedure to the quantum information formalism of bipartite pure states and reduced density matrices. We then discuss important subtleties in understanding entanglement: why a spectral gap alone does not guarantee small entanglement, and how stable rank provides a quantitative measure of mass concentration that better characterizes the relationship between spectral structure and entanglement entropy. Finally, we review the definitions of von Neumann and Rényi entropies and their relationships for completeness.

D.1 Computing Entanglement Entropy at Each MPS Cut

Recall from the MPS construction procedure (*Appendix B*) that we systematically decompose a high-order tensor into an MPS by performing a sequence of SVDs. At each step k, we reshape the tensor to form a matrix $C^{(k)}$ that partitions the system into a "left" subsystem (indices $i_1, \ldots, i_k$ or virtual bonds $\alpha_1, \ldots, \alpha_{k-1}$ combined with i_k) and a "right" subsystem (remaining indices $i_{k+1}, \ldots, i_M$). This bi-partition defines a *cut* in the MPS, and the SVD at this cut,

$$C^{(k)} = U^{(k)} S^{(k)} V^{(k)\dagger}, \tag{23}$$

directly reveals the entanglement structure across this bi-partition. In this subsection, we explain how to compute the entanglement entropy at each such cut.

Bipartite Pure States from MPS Cuts. At cut k, the matrix $C^{(k)}$ has dimensions $d_L^{(k)} \times d_R^{(k)}$, where $d_L^{(k)}$ is the dimension of the left subsystem and $d_R^{(k)}$ is the dimension of the right subsystem. We view this matrix as representing a bipartite pure state between the left and right subsystems. To make this connection explicit, we fix orthonormal bases $\{|i_L\rangle\}_{i_L=1}^{d_L^{(k)}}$ for the left Hilbert space $\mathcal{H}_L \cong \mathbb{C}^{d_L^{(k)}}$ and $\{|i_R\rangle\}_{i_R=1}^{d_R^{(k)}}$ for the right Hilbert space $\mathcal{H}_R \cong \mathbb{C}^{d_R^{(k)}}$. The vectorization of $C^{(k)}$ embeds it into the tensor product space:

$$\mathrm{vec}(C^{(k)}) := \sum_{i_L=1}^{d_L^{(k)}} \sum_{i_R=1}^{d_R^{(k)}} C^{(k)}_{i_L, i_R} \, |i_L\rangle \otimes |i_R\rangle \,. \tag{24}$$

Normalizing yields the pure state

$$\left|\psi^{(k)}\right\rangle := \frac{\mathrm{vec}(C^{(k)})}{\|C^{(k)}\|_F} \in \mathcal{H}_L \otimes \mathcal{H}_R, \qquad \|C^{(k)}\|_F^2 = \sum_{i_L, i_R} |C^{(k)}_{i_L, i_R}|^2. \tag{25}$$

Reduced Density Matrix and Entanglement Spectrum. The reduced density matrix of the left subsystem is obtained by tracing out the right subsystem:

$$\rho_L^{(k)} := \mathrm{Tr}_R\left(\left|\psi^{(k)}\right\rangle \left\langle\psi^{(k)}\right|\right). \tag{26}$$

A direct calculation shows that

$$\rho_L^{(k)} = \frac{C^{(k)}(C^{(k)})^\dagger}{\mathrm{Tr}(C^{(k)}(C^{(k)})^\dagger)} = \frac{C^{(k)}(C^{(k)})^\dagger}{\|C^{(k)}\|_F^2}. \tag{27}$$

The entanglement spectrum is obtained directly from the SVD of $C^{(k)}$. Let $C^{(k)} = U^{(k)} S^{(k)} V^{(k)\dagger}$ with singular values $s_1^{(k)} \geq s_2^{(k)} \geq \cdots \geq s_r^{(k)} > 0$, where r is the rank of $C^{(k)}$. Since

$$C^{(k)}(C^{(k)})^\dagger = U^{(k)} \, \mathrm{diag}((s_1^{(k)})^2, \ldots, (s_r^{(k)})^2, 0, \ldots, 0) \, (U^{(k)})^\dagger, \tag{28}$$

the eigenvalues of the reduced density matrix $\rho_L^{(k)}$ are

$$p_i^{(k)} = \frac{(s_i^{(k)})^2}{\|C^{(k)}\|_F^2} = \frac{(s_i^{(k)})^2}{\sum_{j=1}^r (s_j^{(k)})^2}, \quad i = 1, \ldots, r, \qquad p_{r+1}^{(k)} = \cdots = p_{d_L^{(k)}}^{(k)} = 0. \tag{29}$$

These eigenvalues $\{p_i^{(k)}\}$ form the *entanglement spectrum* at cut k, and they directly determine the entanglement entropy.

Von Neumann Entanglement Entropy. Given the eigenvalues $\{p_i^{(k)}\}$ of the reduced density matrix $\rho_L^{(k)}$, the von Neumann entanglement entropy at cut k is defined as

$$S^{(k)} := -\sum_{i=1}^{d_L^{(k)}} p_i^{(k)} \log p_i^{(k)} = -\sum_{i=1}^{r} p_i^{(k)} \log p_i^{(k)}, \tag{30}$$

where we adopt the convention $0 \log 0 = 0$. This entropy quantifies the amount of entanglement across the bi-partition at cut k: $S^{(k)} = 0$ indicates a product state (no entanglement), while larger values indicate stronger entanglement. The maximum possible entropy is $S_{\max}^{(k)} = \log\Big(\min(d_L^{(k)}, d_R^{(k)})\Big)$, which occurs when all non-zero eigenvalues are equal (maximally entangled state). In our analysis of weight projection matrices ΔW, we reshape the flattened matrix into a tensor and construct an MPS representation. By computing $S^{(k)}$ at each cut k, we obtain an *entanglement profile* that reveals how entanglement varies across different bi-partitions of the weight projection matrix. This profile is central to our characterization of the entanglement structure in LoRA versus full fine-tuning.

For definitions and discussion of Rényi-α entropies (including Rényi-2) and their relationship to von Neumann entropy, see *Appendix D.3*.

D.2 Discussion: Spectral Gap, Stable Rank, and Entanglement Entropy

A "spectral gap" (Saada et al., 2024) indicates that one eigenvalue is separated from the rest, but entanglement entropy is small only when the largest eigenvalue is close to 1, *i.e.*, when the total tail mass $\sum_{i\geq 2} p_i$ is tiny. A large gap does not prevent $\Theta(\log T)$ entanglement: this occurs whenever a constant fraction of the probability mass is spread over $\Theta(T)$ small eigenvalues. To quantify this mass concentration more precisely, we review the definition of the *stable rank* (Saada et al., 2024) in the following.

For a PSD matrix $\Sigma \succeq 0$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_T \geq 0$, the stable rank is defined as

$$\mathrm{sr}(\Sigma) := \frac{\mathrm{Tr}(\Sigma^2)}{\lambda_1^2} = 1 + \sum_{i\geq 2} \left(\frac{\lambda_i}{\lambda_1}\right)^2. \tag{31}$$

The stable rank provides a quantitative measure of mass concentration: $\mathrm{sr}(\Sigma) \approx 1$ means the spectrum is sharply concentrated in the top eigenvalue, which strongly suggests small entanglement for $\rho = \Sigma/\mathrm{Tr}\Sigma$. In contrast, a large stable rank indicates that significant probability mass is distributed across many eigenvalues, which can lead to substantial entanglement even in the presence of a spectral gap. We formalize the connection between stable rank and entanglement entropy in Lemma F.2.

D.3 Von Neumann and Rényi Entropies: Definitions and Relationships

Throughout this work, we primarily use the von Neumann entropy to quantify artificial entanglement. However, for completeness, we also consider the Rényi-α family of entropies, which provides complementary perspectives on entanglement structure. This subsection reviews the definitions and discusses the relationships between these entropy measures.

Definitions. Given a reduced density matrix ρ with eigenvalues $\{p_i\}_{i=1}^T$ satisfying $\sum_i p_i = 1$, the von Neumann entropy is defined as

$$S(\rho) := -\sum_{i=1}^{T} p_i \log p_i, \tag{32}$$

while the Rényi-α entropy (for $\alpha > 0$, $\alpha \neq 1$) is defined as

$$S_\alpha(\rho) := \frac{1}{1-\alpha} \log\left(\sum_{i=1}^{T} p_i^\alpha\right). \tag{33}$$

In particular, the Rényi-2 entropy takes the form

$$S_2(\rho) = -\log \sum_{i=1}^{T} p_i^2 = -\log \operatorname{Tr}(\rho^2). \tag{34}$$

Relationships and Properties. We discuss Rényi-2 entropy in addition to von Neumann entropy because it depends only on the purity $\operatorname{Tr}(\rho^2)$ of the reduced state, making it particularly convenient for theoretical analysis and providing lower bounds on von Neumann entropy. Several of our main theorems provide explicit bounds or scaling for Rényi-2 entropy (see the discussion below). However, the von Neumann entropy captures more fine-grained information about the entanglement spectrum, as it weights each eigenvalue by its logarithm, making it more sensitive to the distribution of spectral weight across modes. The von Neumann entropy corresponds to the limit $\alpha \to 1$ of the Rényi-α entropy: $S(\rho) = \lim_{\alpha\to 1} S_\alpha(\rho)$. For any $\alpha > 0$, the Rényi entropies satisfy the monotonicity property: $S_\alpha(\rho) \geq S_\beta(\rho)$ for $\alpha < \beta$. In particular, we have $S(\rho) \geq S_2(\rho)$ for all density matrices ρ. Additionally, in our analysis of artificial entanglement in LLMs, we primarily use the von Neumann entropy because it is the standard measure of entanglement in quantum information theory, providing a direct connection to the quantum information perspective we adopt, it captures the full entanglement spectrum, making it more sensitive to subtle differences in correlation structures (*e.g.*, the entanglement valley), and it satisfies the subadditivity property, which is crucial for understanding how entanglement scales with system size (volume law vs area law).

Rényi-2 Entropy Results from the Theoretical Overview. For completeness, we collect here the Rényi-2 entropy results that accompany the von Neumann entropy results in Appendix E. For the attention matrix (Theorem E.1), the Rényi-2 entropy satisfies $S_2(A) = 2\log\big(1+\sigma^2\big) + o(1)$, where σ is the bulk spectrum spread parameter, providing a simpler characterization than the full von Neumann entropy scaling. For the stable rank bound (Lemma E.2), we have $S_2(X) \leq 2\log(1+\delta_1)$ where $\delta_1 \leq \sqrt{(T-1)\eta}$ and $\eta = r_{\text{stable}} - 1$. For the output entanglement collapse (Theorem E.3), the Rényi-2 entropy satisfies $S_2(X) = O(1/T) \to 0$, vanishing faster than the von Neumann entropy $S(X) = O((\log T)/T) \to 0$. See details in *Appendix F*.

E A Perspective Based on Random Matrix Theory

We study the artificial entanglement of the attention matrix and the attention output under an asymptotic model based on random matrix theory. Under assumptions consistent with (Saada et al., 2024), we explain the area law scaling with a logarithmic correction observed in the artificial entanglement of the attention matrix. Furthermore, we identify the conditions under which the attention output exhibits low entropy S as the number of tokens increases, and discuss the relationship between these findings and rank collapse in width (Saada et al., 2024). We sketch the theories here, and leave full preliminaries and details in *Appendix D and Appendix F*. While the main focus is von Neumann entropy that is the standard measure in quantum information theory, our theoretical results also include the analysis of Rényi-2 entropy that depends only on the purity $\operatorname{Tr}(\rho^2)$ of the reduced state, since Rényi-2 entropy provides a more tractable quantity for theoretical analysis and offers simpler closed-form expressions. Additionally, Rényi-2 entropy provides lower bounds on von Neumann entropy, useful for theoretical characterization. For a detailed discussion of the relationship between von Neumann and Rényi entropies, see *Appendix D.3*.

The Attention Cardy Formula: Criticality in the Bulk. Following the outlier–bulk law established for softmax attention at isotropic initialization (Saada et al., 2024), we decompose the attention matrix as $A = \frac{1}{T}\mathbf{1}\mathbf{1}^\top + A^\perp$. Here, the rank-1 component $\frac{1}{T}\mathbf{1}\mathbf{1}^\top$ captures the mean-field behavior, while the residual $A^\perp$ governs the fluctuations, converging to a quartercircular law (a deterministic distribution of singular values shaped like a quarter-circle in the large-T limit, see *Appendix F* for details). Below we establish a logarithmic scaling for entanglement entropy, mathematically analogous to the celebrated Cardy formula in conformal field theory (CFT) (Calabrese & Cardy, 2004; 2009):

Theorem E.1 (The Attention Cardy Formula)**.** *Under Assumption 1, let T be the sequence length (number of tokens). In the limit $T \to \infty$, the von Neumann entropy scales as:*

$$S(A) = \mathcal{C}_{\text{attn}} \log T + const + o(1), \tag{35}$$

*where the prefactor $\mathcal{C}_{\text{attn}}$ (termed as the **Effective Attention Charge**) is determined by the bulk spectrum spread parameter $\sigma > 0$ of the limiting quartercircular law Q_σ for the rescaled singular*

values of $\sqrt{T}\,A^{\perp}$*:*

$$\mathcal{C}_{\text{attn}} := \frac{\sigma^2}{1+\sigma^2}. \tag{36}$$

See Assumption 1 for details on σ*.*

Specifically, σ controls the width of the bulk spectrum distribution: larger σ corresponds to a more spread-out distribution of singular values in the bulk, leading to higher entanglement entropy. The parameter σ is determined by the second moment m_2 of the quartercircular law via $m_2 = \sigma^2$, which in turn relates to the Frobenius norm of $A^{\perp}$ as $\|A^{\perp}\|_F^2 \to \sigma^2$ in the large-T limit. For results on Rényi-2 entropy and its relationship to von Neumann entropy, see *Appendix D.3*.

Sketch of proof. The decomposition $\|A\|_F^2 = 1 + \|A^{\perp}\|_F^2$ combined with the quartercircular limit implies $\|A^{\perp}\|_F^2 \to \sigma^2$. Since the operator norm of $A^{\perp}$ scales as $O(T^{-1/2})$, the bulk singular values satisfy $s_i(A^{\perp}) = O(T^{-1/2})$ for $i \geq 2$, and the entanglement spectrum weights are $p_i = s_i(A)^2/\|A\|_F^2 = O(1/T)$ for $i \geq 2$. For the von Neumann entropy, the tail contribution $-\sum_{i\geq 2} p_i \log p_i$ is computed using the representation $s_i(A^{\perp})^2 = (x_i^{(T)})^2/T$ where $x_i^{(T)}$ are the rescaled singular values converging to the quartercircular law. Expanding $\log p_i = \log\big(s_i(A^{\perp})^2/\|A\|_F^2\big) = \log\Big(x_i^{(T)}\Big)^2 - \log T - \log\|A\|_F^2$, the $\log T$ term emerges from the sum $\sum_i (x_i^{(T)})^2/T \to \sigma^2$, yielding the prefactor $\sigma^2/(1+\sigma^2)$ in the logarithmic scaling. The convergence follows from treating the normalized sum as a Riemann sum that converges to the integral defined by the limiting quartercircular distribution. This spectral accumulation is structurally identical to the density of states in gapless quantum systems. Full details are in *Appendix F*.

Remark (From CFT to Attention Cardy Formula). We draw a parallel between our result and the Cardy formula ($S \sim \frac{c}{3}\log L$) for critical quantum systems. In physics, strictly gapped systems follows a strict "area law" where S saturates to a constant ($S \sim O(1)$), implying a finite correlation length. In contrast, Eq. (35) shows that the attention mechanism follows an **area law with a logarithmic correction** ($S \sim \log T$). This logarithmic divergence is the hallmark of a *critical phase* (Eisert et al., 2008), allowing the model to maintain non-vanishing dependencies between distant tokens. Therefore we refer to this critical scaling behavior as the "Attention Cardy Formula", highlighting its role in enabling long-context capabilities beyond the limitations of strict area-law saturation.

A bridge between stable rank and entanglement. To study the *output* $X = AV$, we first relate entanglement to the stable rank of $\Sigma = XX^{\top}$. The following lemma shows that when the stable rank (Saada et al., 2024) is close to 1, the entanglement must be small.

Lemma E.2 (Stable rank implies small entanglement)**.** *Let* $\Sigma = XX^{\top}$ *with eigenvalues* $\lambda_1 \geq \cdots \geq \lambda_T$ *and stable rank* $r_{\text{stable}} = \text{Tr}(\Sigma^2)/\lambda_1^2$*. Let* $\eta = r_{\text{stable}} - 1$ *and* $\delta_1 := \sum_{i\geq 2} \frac{\lambda_i}{\lambda_1}$ *(the sum of ratios of non-dominant to dominant eigenvalues). Then*

$$S(X) \;\leq\; h_2(\delta) + \delta\log(T-1), \tag{37}$$

where $\delta_1 \leq \sqrt{(T-1)\eta}$*,* $\delta = \min\{1, \delta_1\}$*, and* $h_2(u) := -u\log u - (1-u)\log(1-u)$*. For Rényi-2 entropy bounds, see Appendix D.3.*

Sketch of proof. By Cauchy–Schwarz, $(\sum_{i\geq 2}\lambda_i)^2 \leq (T-1)\sum_{i\geq 2}\lambda_i^2$. Dividing by λ_1^2 and using the definition of $\eta = r_{\text{stable}} - 1 = \sum_{i\geq 2}(\lambda_i/\lambda_1)^2$ gives $\delta_1 \leq \sqrt{(T-1)\eta}$. Since $\text{Tr}(\Sigma) = \lambda_1(1+\delta_1)$, the normalized density matrix $\rho = \Sigma/\text{Tr}(\Sigma)$ has top eigenvalue $\lambda_{\max}(\rho) = 1/(1+\delta_1)$, implying $1 - \lambda_{\max}(\rho) = \delta_1/(1+\delta_1) \leq \delta$. With $p_1 \geq 1-\delta$ and tail mass $\sum_{i\geq 2} p_i \leq \delta$, the von Neumann entropy is maximized when the tail mass is uniformly distributed across the remaining $T-1$ eigenvalues, yielding the bound $S(\rho) \leq h_2(\delta) + \delta\log(T-1)$. For Rényi-2 entropy, using $\text{Tr}(\rho^2) \geq p_1^2 = 1/(1+\delta_1)^2$ gives $S_2(\rho) \leq 2\log(1+\delta_1)$. *Appendix F* provides the full arguments.

Output entanglement collapse under stable rank collapse. Recent theory (Saada et al., 2024) shows that in width-limited regimes, the stable rank of $XX^{\top}$ may collapse to $1 + O(T^{-3})$. Plugging this into Lemma E.2 yields:

Theorem E.3 (Output entanglement collapse)**.** *If* $r_{\text{stable}}(XX^\top) - 1 = O(T^{-3})$ *with overwhelming probability*[5]*, then in the limit* $T \to \infty$*,*

$$S(X) = O\left(\frac{\log T}{T}\right) \to 0. \tag{38}$$

Thus the token–feature entanglement vanishes in the rank-collapsed regime. For Rényi-2 entropy scaling, see Appendix D.3.

Sketch of proof. Apply Lemma E.2 with $\eta = r_{\text{stable}}(\Sigma) - 1 = O(T^{-3})$. The bound $\delta_1 \le \sqrt{(T-1)\eta}$ from the lemma gives $\delta_1 = O(T^{-1})$, and hence $\delta = \min\{1, \delta_1\} = O(T^{-1})$. The lemma yields $S(X) \le h_2(\delta) + \delta \log(T-1)$. Since $h_2(\delta) = O(\delta \log(1/\delta))$ and $\delta = O(1/T)$, we have $h_2(\delta) = O((\log T)/T)$, while $\delta \log(T-1) = O((\log T)/T)$. Therefore $S(X) = O((\log T)/T) \to 0$. For Rényi-2 entropy, the lemma gives $S_2(X) \le 2\log(1+\delta_1) = O(\delta_1) = O(1/T) \to 0$, vanishing even faster. This reflects that almost all spectral weight concentrates on the top eigenvalue when the stable rank approaches 1. Full details are given in *Appendix F*.

F Theorems and Proofs

F.1 Setup: Attention Matrix and Output

We follow the assumptions from (Saada et al., 2024). Let $X_0 \in \mathbb{R}^{T\times d}$ have orthonormal rows, *i.e.*,

$$X_0 X_0^\top = I_T. \tag{39}$$

Define

$$Q := X_0 W_Q \in \mathbb{R}^{T\times d_{qk}}, \qquad K := X_0 W_K \in \mathbb{R}^{T\times d_{qk}}, \qquad V := X_0 W_V \in \mathbb{R}^{T\times d_v}, \tag{40}$$

softmax attention

$$A := A(X_0) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_{qk}}}\right) \in \mathbb{R}^{T\times T}, \tag{41}$$

and the single layer output

$$X := AV = A(X_0)\, X_0 W_V \ \in \ \mathbb{R}^{T\times d_v}. \tag{42}$$

F.2 Modeling Assumptions

We prove the theorems under a asymptotic model for softmax attention at isotropic initialization, as studied in (Saada et al., 2024).

Assumption 1 (Outlier + quartercircular bulk for attention (Saada et al., 2024))**.** *We note the fact that* A *is row-stochastic (rows sum to* 1*) due to the softmax function, and define*

$$A^\perp := A - \frac{1}{T}\mathbf{1}\mathbf{1}^\top, \tag{43}$$

where $\mathbf{1} \in \mathbb{R}^T$ *denotes the all-ones vector. Assume the following hold as* $T \to \infty$*:*

Outlier. *The top singular value satisfies* $s_1(A) \to 1$ *and the remainder obeys*

$$\|A^\perp\|_{\text{op}} = O(T^{-1/2}). \tag{44}$$

Here, $\|\cdot\|_{\text{op}}$ *denotes the operator norm, which equals the largest singular value of the matrix. The condition* $s_1(A) \to 1$ *arises directly from the row-stochastic nature of the softmax attention, where the all-ones vector* $\mathbf{1}$ *is an eigenvector with eigenvalue* 1*. The scaling* $\|A^\perp\|_{\text{op}} = O(T^{-1/2})$ *reflects the behavior of large random matrices: since the squared Frobenius norm of* $A^\perp$*, defined as* $\|A^\perp\|_F^2 := \sum_{i,j}(A^\perp_{ij})^2$*, remains bounded (approx.* $O(1)$*, see* ***Step 2*** *in Proof for* ***Theorem*** *F.1) and is spread roughly evenly across* T *dimensions, the strength of the noise along any specific direction scales inversely with the square root of the dimension.*

[5] Overwhelming probability means $1 - T^{-c}$ for every fixed $c > 0$, for all sufficiently large T.

Bulk law. *Let $x_1^{(T)}, \dots, x_T^{(T)}$ be the singular values of $\sqrt{T}\, A^{\perp}$,[6] their empirical measure $\nu_T := \frac{1}{T}\sum_{i=1}^{T} \delta_{x_i^{(T)}}$ converges weakly to the quartercircular law Q_σ with parameter $\sigma > 0$. Intuitively, this means that as the dimension T grows, the histogram of the singular values becomes a smooth, deterministic curve shaped like a quarter-circle. Moreover, the moments needed below are uniformly integrable:*

$$\begin{aligned} \frac{1}{T}\sum_{i=1}^{T}(x_i^{(T)})^2 &\to m_2, \\ \frac{1}{T}\sum_{i=1}^{T}(x_i^{(T)})^4 &\to m_4, \\ \frac{1}{T}\sum_{i=1}^{T}(x_i^{(T)})^2 \log\Big(x_i^{(T)}\Big)^2 &\to m_{2\log}, \end{aligned} \tag{45}$$

where for the quartercircular law one has $m_2 = \sigma^2$ and $m_4 < \infty$.

F.3 Entanglement of the Attention Matrix

Theorem F.1 (Entanglement of the attention matrix: generally $\Theta(\log T)$)**.** *Under Assumption 1 (Saada et al., 2024), let*

$$\rho_A := \rho_{\text{row}}(A) = \frac{AA^\top}{\text{Tr}(AA^\top)}. \tag{46}$$

Then, as $T \to \infty$, the von Neumann and Rényi-2 entanglement entropies satisfy

$$\begin{aligned} S(A) &= \frac{\sigma^2}{1+\sigma^2}\,\log T \;+\; C_A \;+\; o(1), \\ S_2(A) &= 2\log\big(1+\sigma^2\big) \;+\; o(1), \end{aligned} \tag{47}$$

where the constant term is

$$C_A = \log\big(1+\sigma^2\big) + \frac{\sigma^2}{1+\sigma^2}\log\big(1+\sigma^2\big) - \frac{m_{2\log}}{1+\sigma^2}. \tag{48}$$

In particular, $S(A)$ typically grows like $\log T$ even when A has a macroscopic singular-value outlier (equivalently, a macroscopic eigenvalue gap in $AA^\top$). For a discussion of the relationship between von Neumann and Rényi entropies, see Appendix D.3.

Proof. **Step 1: Orthogonality and Frobenius decomposition.** Because A is row-stochastic, $\sum_j A_{ij} = 1$ for each i, hence $\sum_j A^{\perp}_{ij} = 0$ (since the subtracted term $\frac{1}{T}\mathbf{1}\mathbf{1}^\top$ also has unit row sums) and therefore

$$\big\langle A^{\perp}, \tfrac{1}{T}\mathbf{1}\mathbf{1}^\top \big\rangle_F = \sum_{i,j} A^{\perp}_{ij}\cdot\frac{1}{T} = \frac{1}{T}\sum_i\sum_j A^{\perp}_{ij} = 0. \tag{49}$$

Thus,

$$\|A\|_F^2 = \Big\|\frac{1}{T}\mathbf{1}\mathbf{1}^\top\Big\|_F^2 + \|A^{\perp}\|_F^2 = 1 + \|A^{\perp}\|_F^2, \tag{50}$$

since $\|\frac{1}{T}\mathbf{1}\mathbf{1}^\top\|_F^2 = T^2\cdot(1/T^2) = 1$.

Step 2: Limit of $\|A^{\perp}\|_F^2$. Let $x_i^{(T)}$ be the singular values of $\sqrt{T}A^{\perp}$. Then

$$\|A^{\perp}\|_F^2 = \sum_{i=1}^{T} s_i(A^{\perp})^2 = \sum_{i=1}^{T}\frac{(x_i^{(T)})^2}{T} = \frac{1}{T}\sum_{i=1}^{T}(x_i^{(T)})^2 \;\to\; m_2 = \sigma^2, \tag{51}$$

[6] Note that the singular values of $A^{\perp}$ naturally decay as $s_i(A^{\perp}) \sim O(T^{-1/2})$, therefore by defining the rescaled values $x_i^{(T)} := \sqrt{T}s_i(A^{\perp})$, we bring them back to the macroscopic scale $O(1)$ that allows their empirical distribution to converge to a non-trivial, fixed shape (the quartercircular law) rather than collapsing to a Dirac delta at zero.

where we used Assumption 1(2) (Saada et al., 2024). Combining with (50) gives

$$\|A\|_F^2 \to 1 + \sigma^2. \tag{52}$$

Step 3: Entanglement spectrum from singular values. By the matrix–state correspondence reviewed in the preliminaries, the eigenvalues of ρ_A are

$$p_i^{(T)} = \frac{s_i(A)^2}{\|A\|_F^2}, \quad i = 1, \dots, T. \tag{53}$$

Assumption 1(1) states $s_1(A) \to 1$ and $\|A^\perp\|_{\mathrm{op}} = O(T^{-1/2})$, and hence for $i \geq 2$ we have the rank-one approximation bound

$$s_i(A) \leq \left\|A - \tfrac{1}{T}\mathbf{1}\mathbf{1}^\top\right\|_{\mathrm{op}} = \|A^\perp\|_{\mathrm{op}} = O(T^{-1/2}), \tag{54}$$

so $p_i^{(T)} = O(1/T)$ for $i \geq 2$. Moreover, (52) implies

$$p_1^{(T)} = \frac{s_1(A)^2}{\|A\|_F^2} \to \frac{1}{1+\sigma^2}. \tag{55}$$

Step 4: Rényi-2 entropy. We have

$$\mathrm{Tr}(\rho_A^2) = \sum_{i=1}^{T} (p_i^{(T)})^2 = (p_1^{(T)})^2 + \sum_{i\geq 2} (p_i^{(T)})^2. \tag{56}$$

For the tail term, using $p_i^{(T)} = s_i(A)^2/\|A\|_F^2$ and the bound $s_i(A) = O(T^{-1/2})$ for $i \geq 2$,

$$\sum_{i\geq 2} (p_i^{(T)})^2 = \frac{1}{\|A\|_F^4} \sum_{i\geq 2} s_i(A)^4 \leq \frac{T-1}{\|A\|_F^4} \|A^\perp\|_{\mathrm{op}}^4 = O\left(\frac{1}{T}\right), \tag{57}$$

where we used $\|A^\perp\|_{\mathrm{op}} = O(T^{-1/2})$ from Assumption 1 and $\|A\|_F^2 \to 1 + \sigma^2$ from (52). Therefore,

$$\mathrm{Tr}(\rho_A^2) = (p_1^{(T)})^2 + O(1/T) \to \frac{1}{(1+\sigma^2)^2}, \tag{58}$$

and hence

$$S_2(A) = -\log \mathrm{Tr}(\rho_A^2) \to 2\log\big(1+\sigma^2\big), \tag{59}$$

which matches (47). For definitions and properties of Rényi-2 entropy, see *Appendix D.3*.

Step 5: von Neumann entropy. Write

$$S(A) = -p_1^{(T)} \log p_1^{(T)} \; - \sum_{i\geq 2} p_i^{(T)} \log p_i^{(T)}. \tag{60}$$

The first term converges using (55):

$$-p_1^{(T)} \log p_1^{(T)} \to \frac{1}{1+\sigma^2} \log\big(1+\sigma^2\big). \tag{61}$$

For the tail, we use the representation $s_i(A^\perp) = x_i^{(T)}/\sqrt{T}$ and work directly at the level of normalized sums. Under Assumption 1 (uniform integrability of the relevant moments), replacing $\{s_i(A)\}_{i\geq 2}$ by $\{s_i(A^\perp)\}_{i=1}^T$ in the forthcoming Riemann–sum expressions incurs only an $o(1)$ error as $T \to \infty$. This is justified by the theory of finite-rank perturbations. Since A is a rank-1 perturbation of $A^\perp$, the rank inequalityfor singular value distributions implies that their empirical cumulative distribution functions differ by at most $1/T$ in Kolmogorov distance (defined as $\|F - G\|_\infty := \sup_{x\in\mathbb{R}} |F(x) - G(x)|$ for two distribution F and G):

$$\left\|F_A - F_{A^\perp}\right\|_\infty \leq \frac{\mathrm{rank}(A - A^\perp)}{T} = \frac{1}{T},^{7} \tag{63}$$

[7]Although the classical rank inequality (see Theorem A.43 in (Bai & Silverstein, 2010)) is stated for eigenvalues of Hermitian matrices, the corresponding result for singular values of arbitrary matrices follows

where the empirical singular value distribution (or empirical spectral distribution (ESD)) of A is defined as $F_A(x) := \frac{1}{T}\#\{i : s_i(A) \le x\}$ where $\#E$ denotes the cardinality of the set E, and similarly for $F_{A^\perp}$. Here $s_1(A) \ge \cdots \ge s_T(A)$ denote the singular values. Consequently, only $O(1)$ singular values can deviate by a non-negligible amount, which is negligible compared to the ambient dimension T[8]. Because our sums are normalized by $1/T$ (implicitly through the Frobenius-norm–based weights[9]), the contribution of these $O(1)$ mismatched indices vanishes at rate $O(1/T)$. The remaining sum over the bulk can thus be treated as a Riemann sum that converges to the integral defined by the limiting spectral distribution of $\sqrt{T}A^\perp$. Concretely,

$$\begin{aligned}\sum_{i\ge 2} p_i^{(T)} \log p_i^{(T)} &= \frac{1}{\|A\|_F^2} \sum_{i\ge 2} s_i(A)^2 \log\left(\frac{s_i(A)^2}{\|A\|_F^2}\right) \\ &= \frac{1}{\|A\|_F^2} \sum_{i=1}^{T} s_i(A^\perp)^2 \log\left(\frac{s_i(A^\perp)^2}{\|A\|_F^2}\right) + o(1).\end{aligned} \tag{64}$$

Now substitute $s_i(A^\perp)^2 = (x_i^{(T)})^2/T$:

$$\begin{aligned}\sum_{i\ge 2} p_i^{(T)} \log p_i^{(T)} &= \frac{1}{\|A\|_F^2} \sum_{i=1}^{T} \frac{(x_i^{(T)})^2}{T} \log\left(\frac{(x_i^{(T)})^2}{T\|A\|_F^2}\right) + o(1) \\ &= \frac{1}{\|A\|_F^2} \sum_{i=1}^{T} \frac{(x_i^{(T)})^2}{T} \left(\log\left(x_i^{(T)}\right)^2 - \log T - \log\|A\|_F^2\right) + o(1).\end{aligned} \tag{65}$$

Therefore,

$$\begin{aligned}-\sum_{i\ge 2} p_i^{(T)} \log p_i^{(T)} &= \frac{\log T}{\|A\|_F^2} \cdot \frac{1}{T}\sum_{i=1}^{T} (x_i^{(T)})^2 + \frac{\log\|A\|_F^2}{\|A\|_F^2} \cdot \frac{1}{T}\sum_{i=1}^{T} (x_i^{(T)})^2 \\ &\quad - \frac{1}{\|A\|_F^2} \cdot \frac{1}{T}\sum_{i=1}^{T} (x_i^{(T)})^2 \log\left(x_i^{(T)}\right)^2 + o(1).\end{aligned} \tag{66}$$

Using Assumption 1, $\frac{1}{T}\sum_i (x_i^{(T)})^2 \to m_2 = \sigma^2$ and $\frac{1}{T}\sum_i (x_i^{(T)})^2 \log\left(x_i^{(T)}\right)^2 \to m_{2\log}$, and using $\|A\|_F^2 \to 1+\sigma^2$, we obtain

$$-\sum_{i\ge 2} p_i^{(T)} \log p_i^{(T)} = \frac{\sigma^2}{1+\sigma^2}\log T + \frac{\sigma^2}{1+\sigma^2}\log(1+\sigma^2) - \frac{m_{2\log}}{1+\sigma^2} + o(1). \tag{67}$$

Combining with (61) yields (47) with the stated constant term. □

immediately from Hermitian dilation. Recall that the singular values of A are the eigenvalues of the Hermitian matrix $H(A) = \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}$. If $A = A^\perp + uv^\top$ with $\mathrm{rank}(uv^\top) = 1$, then $\mathrm{rank}\big(H(A) - H(A^\perp)\big) = 2$. Applying the rank inequality to $H(A)$ and $H(A^\perp)$ yields

$$\|F_A - F_{A^\perp}\|_\infty = \|F_{H(A)} - F_{H(A^\perp)}\|_\infty \le \frac{2}{2T} = \frac{1}{T}. \tag{63}$$

Equivalently, a rank-1 perturbation of a matrix can change at most one singular value index, so the empirical singular-value distribution can shift vertically by at most $1/T$.

[8] Formally, the "index-shift inequality" for singular values proved by (Thompson, 1976) (See Theorem 1, 3 in (Thompson, 1976)) guarantee stability of individual singular values, while the rank inequality ensures that $\|F_A - F_{A^\perp}\|_\infty \le \mathrm{rank}(A - A^\perp)/T = 1/T$. Therefore, the normalized counting measures of A and $A^\perp$ are asymptotically identical.

[9] The $1/T$ normalization arises because the Frobenius norm induces weights of the form $p_i = s_i(A)^2/\|A\|_F^2$, and bulk singular values satisfy $s_i(A^\perp) = O(T^{-1/2})$, so that $p_i = O(1/T)$. Consequently, any spectral sum of the form $\sum_i p_i\, \Psi(s_i)$ is effectively an average over T terms. Since a rank-1 perturbation can alter only $O(1)$ singular values, the total contribution of these mismatched indices is at most $O(1)\cdot(1/T) = O(1/T)$, which vanishes as $T \to \infty$.

The Attention Cardy Formula: Critical Scaling and Physical Interpretation. Here we provides the full proof of the Attention Cardy Formula stated in Theorem E.1 (see Appendix E). The complete derivation shows that $S(A) = \mathcal{C}_{\text{attn}} \log T + C_A + o(1)$, where the constant term C_A is explicitly given in (47). For the physical interpretation, connection to conformal field theory, and discussion of criticality and long-range correlations, see Appendix E and the Remark following Theorem E.1.

F.4 A General Bridge Between Stable Rank and Entanglement Bounds

To have **Theorem F.3**, we firstly have **Lemma F.2** shown below:

Lemma F.2 (Stable rank control implies small entanglement)**.** *Let $M \in \mathbb{R}^{T\times d}$, $\Sigma := MM^\top \succeq 0$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_T \geq 0$. Let $\rho := \Sigma/\mathrm{Tr}(\Sigma)$ and define the stable rank*

$$r_{\text{stable}}(\Sigma) := \frac{\mathrm{Tr}(\Sigma^2)}{\lambda_1^2} = 1 + \sum_{i\geq 2} \Big(\frac{\lambda_i}{\lambda_1}\Big)^2. \tag{68}$$

Let

$$\eta := r_{\text{stable}}(\Sigma) - 1 = \sum_{i\geq 2} \Big(\frac{\lambda_i}{\lambda_1}\Big)^2. \tag{69}$$

Then the tail mass satisfies

$$\delta_1 := \sum_{i\geq 2} \frac{\lambda_i}{\lambda_1} \;\leq\; \sqrt{(T-1)\eta}. \tag{70}$$

Consequently, letting $\delta := \min\{1, \delta_1\}$,

$$\begin{aligned} 1 - \lambda_{\max}(\rho) &\leq \frac{\delta_1}{1+\delta_1} \leq \delta, \\ S(M) = S(\rho) &\leq h_2(\delta) + \delta \log(T-1), \\ S_2(M) = S_2(\rho) &\leq 2\log(1+\delta_1), \end{aligned} \tag{71}$$

where $h_2(u) := -u\log u - (1-u)\log(1-u)$.

Proof. **Step 1: Tail ℓ_1 control from stable rank.** By Cauchy–Schwarz[10],

$$\Big(\sum_{i\geq 2} \lambda_i\Big)^2 \leq (T-1)\sum_{i\geq 2} \lambda_i^2. \tag{72}$$

Dividing by λ_1^2 gives

$$\delta_1^2 = \Big(\sum_{i\geq 2} \frac{\lambda_i}{\lambda_1}\Big)^2 \leq (T-1)\sum_{i\geq 2} \Big(\frac{\lambda_i}{\lambda_1}\Big)^2 = (T-1)\eta, \tag{73}$$

which proves (70).

Step 2: Bound the top eigenvalue of ρ. Since $\mathrm{Tr}(\Sigma) = \lambda_1 + \sum_{i\geq 2}\lambda_i = \lambda_1(1+\delta_1)$,

$$\lambda_{\max}(\rho) = \frac{\lambda_1}{\mathrm{Tr}(\Sigma)} = \frac{1}{1+\delta_1}. \tag{74}$$

Therefore,

$$1 - \lambda_{\max}(\rho) = 1 - \frac{1}{1+\delta_1} = \frac{\delta_1}{1+\delta_1} \leq \min\{1, \delta_1\} = \delta, \tag{75}$$

establishing the first inequality in (71).

Step 3: Von Neumann entropy upper bound. Let $p_1 \geq p_2 \geq \cdots \geq p_T$ be eigenvalues of ρ. From Step 2, $p_1 \geq 1-\delta$ and thus $\sum_{i\geq 2} p_i \leq \delta$ due to normalization condition. Among all distributions

[10] Applying the Cauchy-Schwarz inequality to the vector of tail eigenvalue ratios $u = (\lambda_2/\lambda_1, \ldots, \lambda_T/\lambda_1)$ and the all-ones vector $\mathbf{1} \in \mathbb{R}^{T-1}$ yields $(\sum_{i\geq 2}\lambda_i/\lambda_1)^2 = \langle u, \mathbf{1}\rangle^2 \leq \|u\|_2^2\|\mathbf{1}\|_2^2$. Since $\|u\|_2^2 = \eta$ and $\|\mathbf{1}\|_2^2 = T-1$, we have $\delta_1^2 \leq \eta(T-1)$.

on T points with tail mass at most δ, the Von Neumann entropy is maximized by $q_1 = 1-\delta$ and $q_2 = \cdots = q_T = \delta/(T-1)$. Hence $S(\rho) \le H(q)$, and

$$\begin{aligned} H(q) &= -(1-\delta)\log(1-\delta) - (T-1)\frac{\delta}{T-1}\log\left(\frac{\delta}{T-1}\right) \\ &= h_2(\delta) + \delta\log(T-1), \end{aligned} \tag{76}$$

which gives the second inequality in (71).

Step 4: Rényi-2 bound. Since $\mathrm{Tr}(\rho^2) = \sum_i p_i^2 \ge p_1^2$ and $p_1 = \lambda_{\max}(\rho) = 1/(1+\delta_1)$,

$$S_2(\rho) = -\log \mathrm{Tr}(\rho^2) \le -\log p_1^2 = 2\log(1+\delta_1), \tag{77}$$

which is the third inequality in (71). For definitions and properties of Rényi-2 entropy, see *Appendix D.3*. □

F.5 Output Entanglement Collapse from Stable Rank Collapse

Theorem F.3 (Output entanglement collapse (token–feature), conditional)**.** *Let $X = AV \in \mathbb{R}^{T\times d_v}$ with $V = X_0 W_V$, and define*

$$\Sigma := XX^\top. \tag{78}$$

Assume that with overwhelming probability[11]

$$r_{\mathrm{stable}}(\Sigma) - 1 = O(T^{-3}), \tag{79}$$

as established in rank-collapsed regimes in (Saada et al., 2024). Then, with overwhelming probability,

$$\begin{aligned} S(X) &= O\left(\frac{\log T}{T}\right) \to 0, \\ S_2(X) &= O\left(\frac{1}{T}\right) \to 0. \end{aligned} \tag{80}$$

For a discussion of Rényi-2 entropy and its relationship to von Neumann entropy, see Appendix D.3.

Proof. Apply Lemma F.2 with $\eta = r_{\mathrm{stable}}(\Sigma) - 1$. Under (79) we have $\delta_1 \le \sqrt{(T-1)\eta} = O(T^{-1})$ and hence $\delta = \min\{1, \delta_1\} = O(T^{-1})$. Lemma F.2 yields

$$S(X) \le h_2(\delta) + \delta\log(T-1). \tag{81}$$

Using $h_2(\delta) = O(\delta\log(1/\delta)) = O((\log T)/T)$ for $\delta = O(1/T)$ gives $S(X) = O((\log T)/T) \to 0$. Similarly, by Lemma F.2,

$$S_2(X) \le 2\log(1+\delta_1) = O(\delta_1) = O(1/T) \to 0. \tag{82}$$

For definitions and properties of Rényi-2 entropy, see *Appendix D.3*. □

Remark. Theorem F.3 is *conditional*. Under isotropic random initialization of W_V and $X_0 X_0^\top = I_T$, one often has $VV^\top \approx I_T$ and hence $\Sigma \approx AA^\top$ (Saada et al., 2024), so the output entanglement scaling can mirror that of A in Theorem F.1 rather than vanishing. The stable rank collapse condition $r_{\mathrm{stable}}(\Sigma) - 1 \ll 1$ should therefore be interpreted as a property that can emerge in a trained or otherwise constrained regime and can be validated empirically.

G Artificial Entanglement Profiling in MPS Adaptation

While the MPS modeling above is adopted for artificial entanglement profiling of LoRA and FFT (where we decompose their parameter updates into MPS representations to analyze entanglement structure), we can extend this analysis framework to a fine-tuning method that directly parameterizes weight updates using MPS structure, which is also parameter efficient and we term as *MPS adaptation*. The key distinction is that MPS adaptation is a PEFT method (a way to parameterize trainable updates), whereas MPS modeling is an analysis tool (a way to decompose and analyze parameter structures). By

[11] Overwhelming probability means $1 - T^{-c}$ for every fixed $c > 0$, for all sufficiently large T.

applying the same entanglement profiling framework to MPS adaptation, we can investigate whether it exhibits similar entanglement signatures (such as the entanglement valley observed in LoRA) and compare the entanglement structures across different PEFT paradigms. For a detailed explanation of MPS adaptation, its relationship to MPS modeling, and why it can be parameter-efficient, see *Appendix G.1*. To investigate this, we implement a simple version of MPS adaptation: We formalize $A \in \mathbb{R}^{r \times d_{\text{in}}}$ into two Order-3 tensors whose contraction reconstructs the effective matrix A, where we denote it as A_{MPS}. Specifically, we factorize the input dimension as $d_{\text{in}} = d_1 d_2$ and parameterize

$$A_{\text{MPS}} = \sum_{\alpha=1}^{\chi} \mathcal{A}^{[1]}_{r,\alpha,d_1} \mathcal{A}^{[2]}_{\alpha,1,d_2}, \tag{83}$$

where each $\mathcal{A}^{[i]}$ is an Order-3 tensor and the bond dimension χ controls the entanglement capacity. Analogous to LoRA's $\Delta W = \frac{\alpha}{r} BA$ where $B \in \mathbb{R}^{d_{\text{out}} \times r}$ and $A \in \mathbb{R}^{r \times d_{\text{in}}}$, MPS adaptation replaces the full matrix A with an MPS factorization A_{MPS}. The weight update is then:

$$\Delta W = \frac{\alpha}{r} B A_{\text{MPS}} = \frac{\alpha}{r} B \left(\sum_{\alpha=1}^{\chi} \mathcal{A}^{[1]}_{r,\alpha,d_1} \mathcal{A}^{[2]}_{\alpha,1,d_2} \right), \tag{84}$$

where the contraction over the bond index α reconstructs $A_{\text{MPS}} \in \mathbb{R}^{r \times d_{\text{in}}}$ from the tensor network factorization, and B remains the same matrix as in LoRA. This MPS parameterization of A can reduce the number of parameters compared to storing the full $r \times d_{\text{in}}$ matrix under specific settings, see a concrete example in *Appendix G.1*.

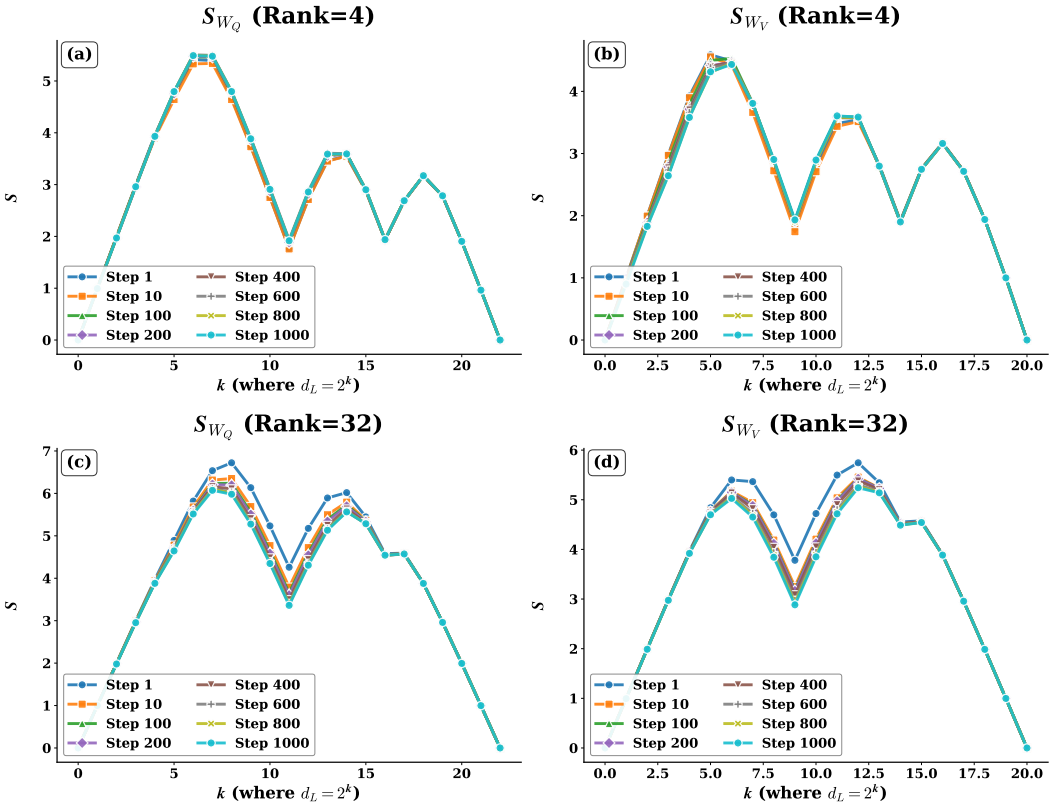

Figure 9: Artificial entanglement profiling of the MPS adaptation of ΔW for ranks $r \in \{4, 32\}$ with $\alpha = 16$ across training steps. Both rank settings exhibit a characteristic "entanglement valley" structure, while the rank-4 case displays a more pronounced *bi-modal* (dual-valley) profile. In contrast, the rank-32 curves form a smoother single-valley landscape.

Artificial Entanglement Profiling in MPS adaptation. We conduct below experiments:

(i) We conduct the artificial entanglement profiling of ΔW_Q and ΔW_V under different r (or χ), see FIG. 9. To simplify the problem, we set $r = \chi$, and therefore we do not set r or χ exceeds

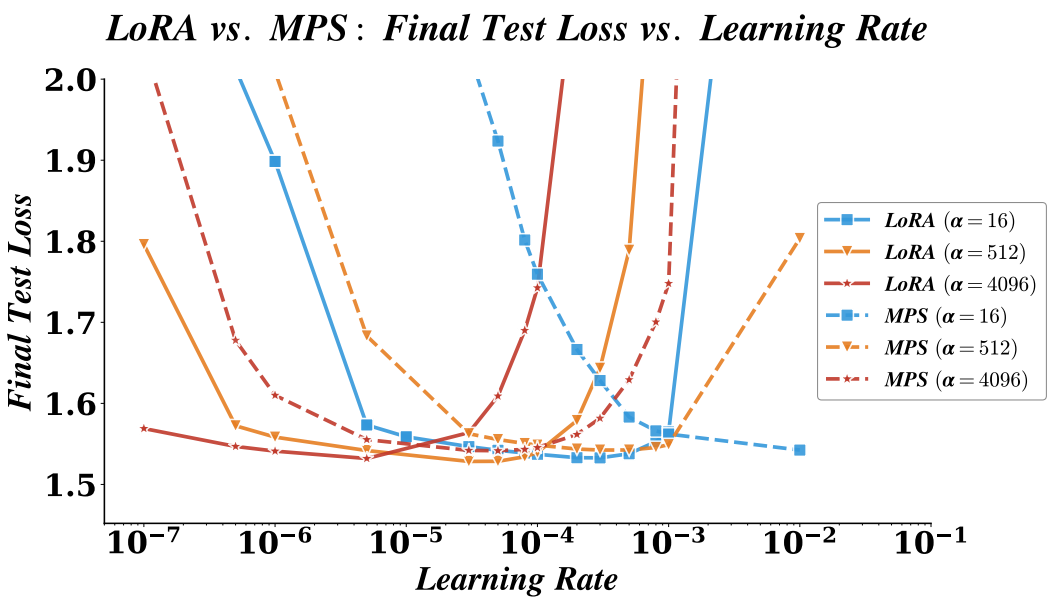

Figure 10: Final test loss of LoRA and MPS adaptation across multiple learning rates and scaling coefficients α. Compared to LoRA, the MPS adaptation exhibits an optimal learning-rate region that is consistently shifted toward larger learning rates.

$\min\{d_{\text{out}}, d_1, d_2\}$. Here we respectively set $r = \{4, 32\}$, and learning rate to be 2×10^2. Other settings follow **Default Experimental Setups** in Sec. 5.

(ii) We conduct the artificial entanglement profiling of S_A, see FIG. 11.

(iii) We vary α and learning rates to see the shift of the optimal learning rates, see FIG. 10.

Combining the results, in MPS adaptation we can observe rich entanglement profiling results in the updates of projection matrices (ΔW_Q and ΔW_V) and the no-hair property. Besides, similar to LoRA, when increasing α, the optimal learning rates are shifted to be smaller as the increase of α, approaching the solutions of LoRA and FFT. Besides, though possessing different entanglement profiling, MPS adaptation does not induce an obvious test loss decrease in the optimal setting, approximating the test performance of LoRA.

G.1 Method and Relationship to MPS Modeling

This section clarifies the distinction between *MPS modeling* (an analysis tool) and *MPS adaptation* (a PEFT method), explains their relationship, and demonstrates why MPS adaptation can be parameter-efficient.

MPS Modeling vs. MPS Adaptation. It is crucial to distinguish between two different uses of MPS in this work. **MPS Modeling (Analysis Tool):** We use MPS decomposition as an *analysis framework* to study the entanglement structure of parameter updates from LoRA and FFT. Specifically, we decompose the learned weight update matrices ΔW into MPS representations post-hoc to compute entanglement entropy profiles. This is purely for analysis purposes: the training process itself does not use MPS structure. **MPS Adaptation (PEFT Method):** We term and use MPS adaptation that directly parameterizes weight updates using MPS structure during training. Unlike LoRA which uses low-rank matrices BA, MPS adaptation uses tensor network structure as the fundamental parameterization, making the MPS structure intrinsic to the fine-tuning method rather than a tool for analysis.

Relationship and Connection. The connection between MPS modeling and MPS adaptation is that both utilize the same mathematical structure (MPS), but for different purposes: MPS modeling uses it to *analyze* existing methods, while MPS adaptation uses it to *define* a fine-tuning method. This

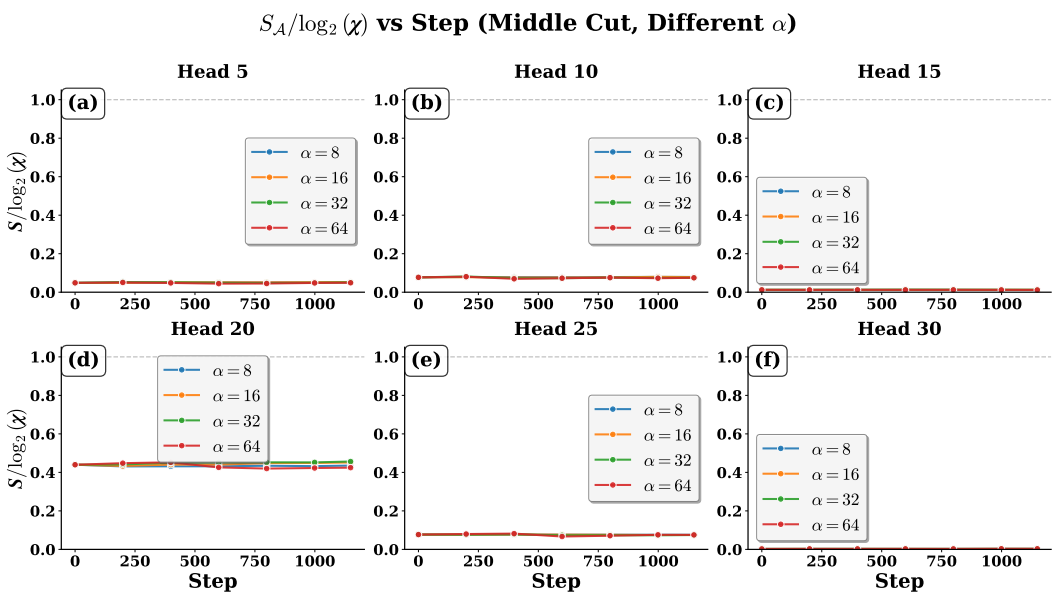

Figure 11: Entanglement entropy $S_A/\log(\chi)$ across training steps for several attention heads under scaling coefficients $\alpha \in \{8, 16, 32, 64\}$. Across all heads, the entanglement curves remain approximately invariant with respect to both α and the training step. This is similar to the behavior observed in LoRA (See FIG. 20).

allows us to apply the same entanglement analysis framework to MPS adaptation, enabling direct comparison of entanglement structures across different fine-tuning paradigms.

Parameter Efficiency of MPS Adaptation. To illustrate why MPS adaptation can be parameter-efficient, consider a concrete example. For a weight update matrix $\Delta W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ with $d_{\text{out}} = 4096$ and $d_{\text{in}} = 4096$, full fine-tuning requires $4096 \times 4096 = 16{,}777{,}216$ parameters.

LoRA with rank $r = 256$ parameterizes $\Delta W = \frac{\alpha}{r} BA$ where $B \in \mathbb{R}^{4096 \times 256}$ and $A \in \mathbb{R}^{256 \times 4096}$, requiring $4096 \times 256 + 256 \times 4096 = 2{,}097{,}152$ parameters.

MPS adaptation factorizes the input dimension as $d_{\text{in}} = d_1 \times d_2$ (e.g., $4096 = 64 \times 64$) and parameterizes $A \in \mathbb{R}^{r \times d_{\text{in}}}$ as an MPS with bond dimension χ, denoted as A_{MPS}:

$$A_{\text{MPS}} = \sum_{\alpha=1}^{\chi} \mathcal{A}^{[1]}_{r,\alpha,d_1} \mathcal{A}^{[2]}_{\alpha,1,d_2}, \tag{85}$$

where $\mathcal{A}^{[1]} \in \mathbb{R}^{r \times \chi \times d_1}$ and $\mathcal{A}^{[2]} \in \mathbb{R}^{\chi \times 1 \times d_2}$ are Order-3 tensors. The weight update is then $\Delta W = \frac{\alpha}{r} B A_{\text{MPS}}$ where $B \in \mathbb{R}^{d_{\text{out}} \times r}$ is the same matrix as in LoRA. With $r = 256$ and $\chi = 32$, this requires $256 \times 32 \times 64 + 32 \times 1 \times 64 = 524{,}288 + 2{,}048 = 526{,}336$ parameters for A_{MPS}, plus $4096 \times 256 = 1{,}048{,}576$ parameters for B, totaling approximately $1{,}574{,}912$ parameters—significantly fewer than LoRA while maintaining similar expressivity and test performance (see FIG. 10) through the tensor network structure. The parameter efficiency arises from the hierarchical factorization: instead of storing a full $r \times d_{\text{in}}$ matrix, we factorize it into smaller tensors whose total parameter count scales as $O(r\chi d_1 + \chi d_2)$ rather than $O(r d_{\text{in}})$, where typically $\chi \ll r$ and $d_1, d_2 \ll d_{\text{in}}$.

H Additional Main-Setting Results

H.1 Random-Matrix Benchmark

Volume law in a random Gaussian matrix. A tensor (including a matrix) with i.i.d. Gaussian entries, when reshaped into a vector in a tensor-product space and properly normalized, is statistically equivalent to sampling a Haar-random pure state (Mezzadri, 2006). Here, viewing the tensor as a

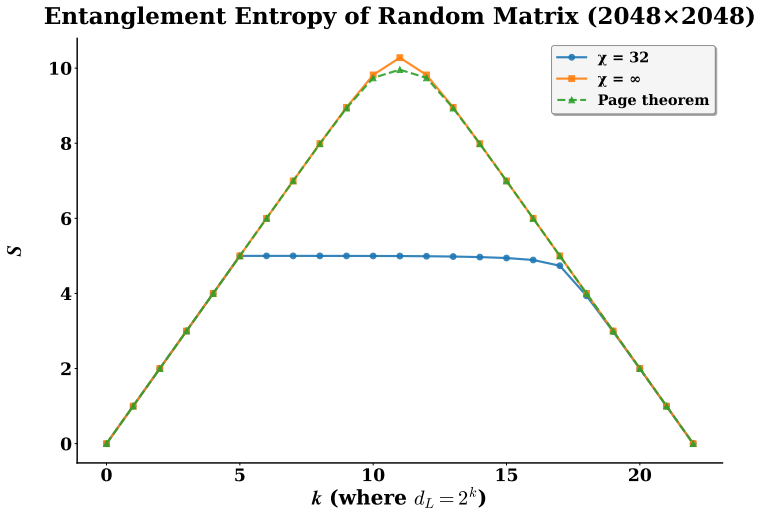

Figure 12: Artificial entanglement profiling of a random Gaussian matrix. The orange curve ($\chi = \infty$) corresponds to the full SVD at each bond and closely matches the Page-curve prediction for Haar-random states (green dashed line). The blue curve ($\chi = 32$) demonstrates the effect of truncating the MPS bond dimension the entropy saturates once the Schmidt rank exceeds χ, forming a plateau, while the true entropy (the orange) continues to increase toward the Page limit.

vector is natural, since an MPS is precisely a structured decomposition of such a high-dimensional vector via successive Schmidt decompositions. In the quantum domain, Haar-random pure states serve as a canonical model of maximally unstructured states, whose bipartite entanglement properties are well captured by Page theory (Page, 1993). Accordingly, the von Neumann entanglement entropy S obeys the Page law, which to leading order reads $S_{\text{Page}}(d_L) \simeq \log d_L - \frac{d_L}{2d_R \ln 2}$ for $d_L \leq d_R$, where d_L and d_R denote the dimensions of the left and right subsystems respectively after a bi-partition. This implies a *volume-law* scaling of entanglement, where S increases linearly with the size of the smaller subsystem. For Haar-random states, this linear growth further leads to a saturation near the mid cut. We show this profiling curve by sampling a real-valued random matrix $R \in \mathbb{R}^{m \times n}$ with i.i.d. Gaussian entries, with the total dimension $D = mn$ where we set $m, n = 2048$. We then factorize D into a list of local dimensions using *prime factorization*, where in this case the factor is 2. Then a similar construction of the MPS is conducted (See Sec. 4). At each cut, we retain at most χ_{max} singular values, where $\chi_{\text{max}} = 32$ or ∞. Therefore the exact $\chi = \min\{\chi_{\text{max}}, d_L, d_R\}$. FIG. 12 shows this artificial entanglement profiling curve, where we also draw S_{Page} as an analytic benchmark. FIG. 12 here also provides a benchmark profiling about how an artificial entanglement profiling is visualized in our study.

H.2 Hyperparameter Sensitivity and Optimization

H.3 Additional Attention-Head Profiles

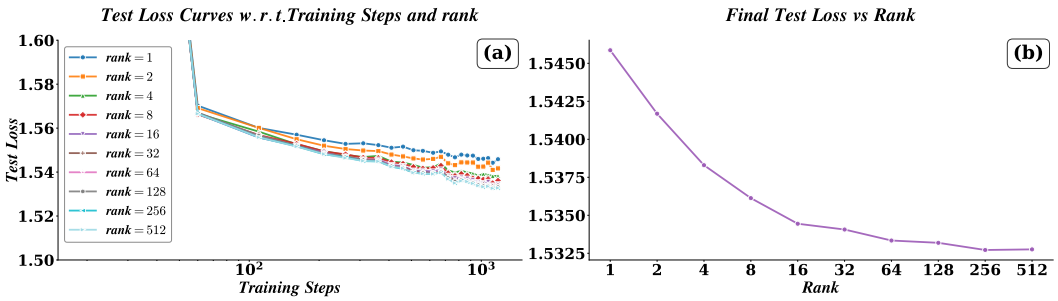

Figure 13: Test loss across training steps for different LoRA ranks r. When r is small, increasing r generally improves performance. While increasing to a specific r the test loss saturates and is not improved much. We keep other hyper-parameters fixed, such as the learning rate and α.

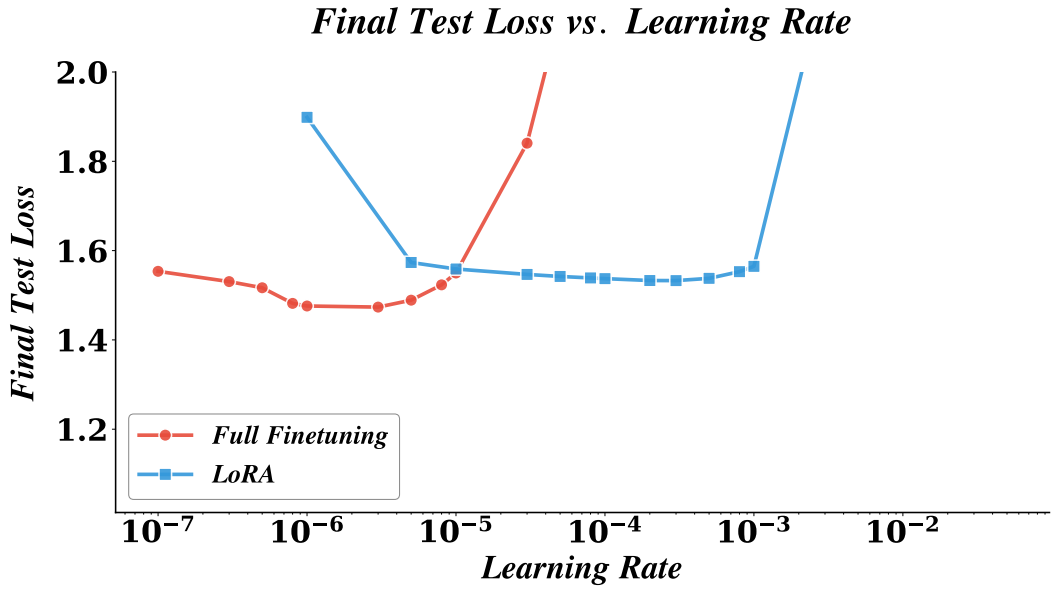

Figure 14: Final test loss as a function of learning rate for FFT and LoRA when the scaling parameter α is small (here $\alpha = 16$). LoRA requires a larger learning rate than FFT to achieve the best test performance, consistent with previous observations (Biderman et al., 2024; Schulman & Lab, 2025).

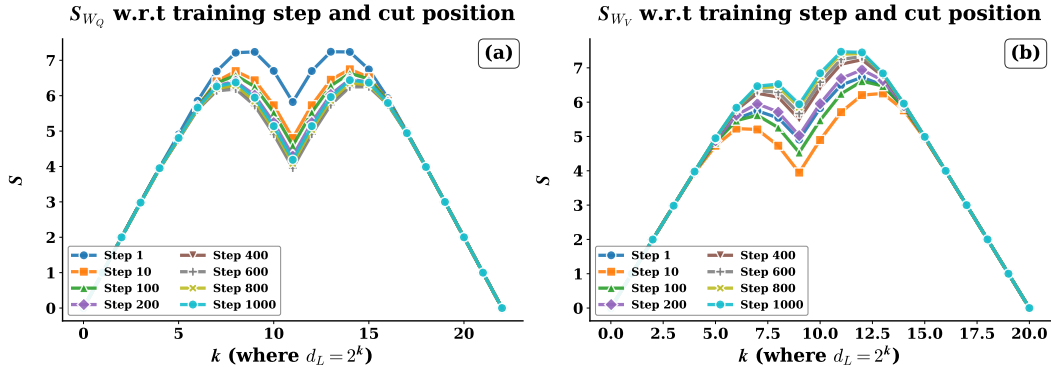

Figure 15: $S_{\Delta W_Q}$ and $S_{\Delta W_V}$ w.r.t the cut position k (with $d_L = 2^k$) across training steps when increasing $\alpha = 256$. Panels (a) and (b) show that at early time, the curves exhibit a pronounced "entanglement valley" structure centered near the middle cuts. As training progresses, the curves of ΔW_Q are less deepened as the cases in smaller $\alpha = 16$, while the curves of ΔW_V gradually lift and approach a saturated shape.

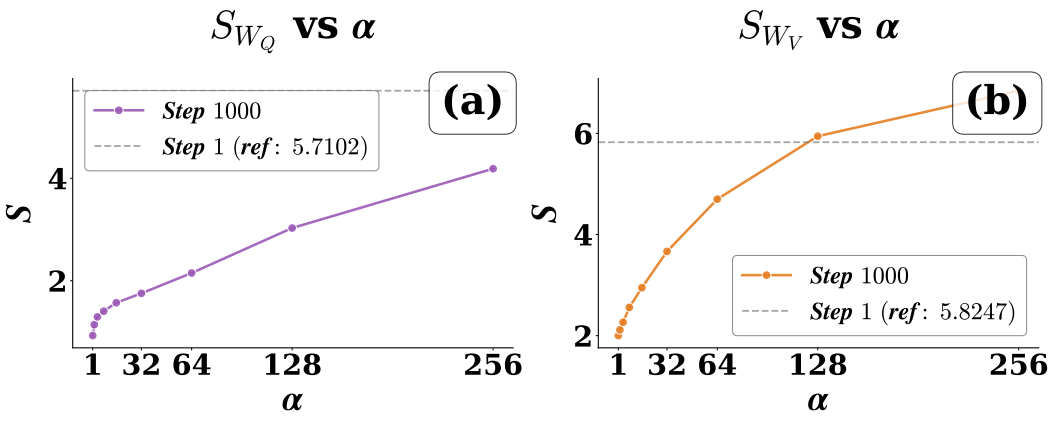

Figure 16: Entanglement entropy S in the middle bi-partition for different α. S in early time (Step = 1) serves as a baseline. By showing the converged solution at late time (Step = 1000), this demonstrates how $\alpha \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$ impacts the solution learned by LoRA.

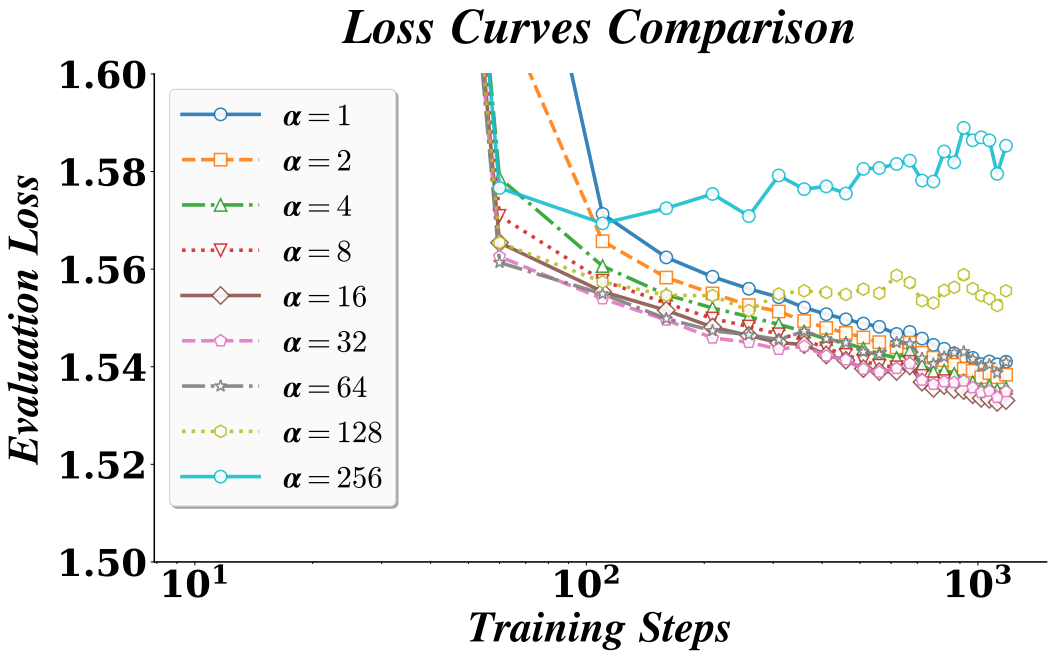

Figure 17: Test loss comparison across the scaling factor α in LoRA. X-axis is shown in log scale.

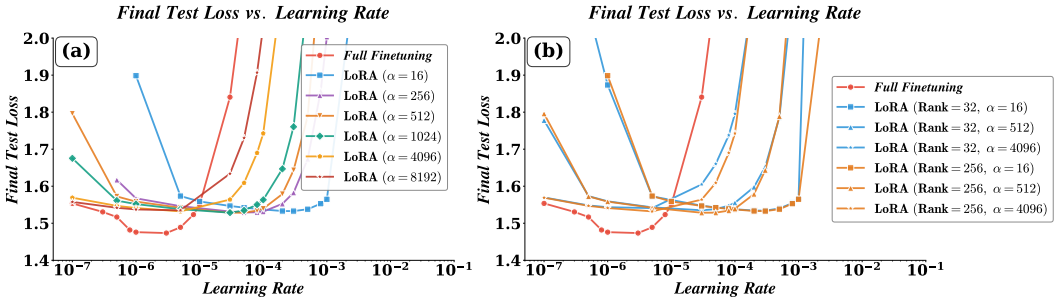

Figure 18: Final test loss with respect to learning rate under different LoRA configurations. (a) Comparison of full fine-tuning with LoRA at multiple α when rank $r = 256$. Larger α values shift the optimal learning rate to be smaller, approaching the solutions from full-fine-tuning. (b) Joint variation of r and α for LoRA. The settings $r = 32$ and $r = 256$ do not shift the optimal learning rate in different α.

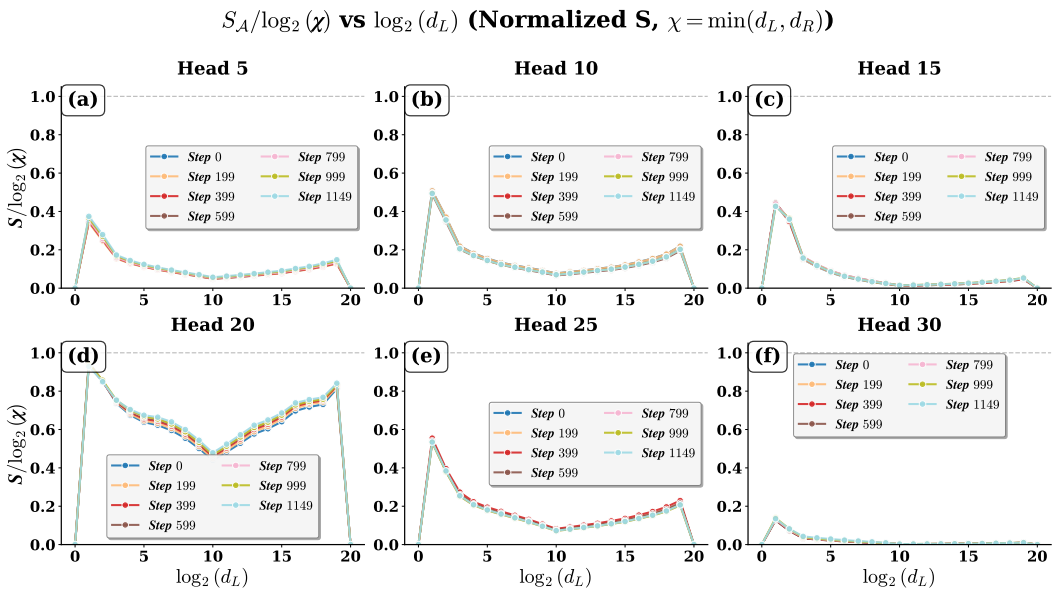

Figure 19: Normalized entanglement entropy $S_A/\log(\chi)$ of six representative attention heads across training steps, where $\chi = \min(d_L, d_R)$ denotes the bond dimension, and $\log(\chi)$ represents the theoretical maximum entanglement entropy. Most heads exhibit consistently mild scaling throughout training, far from the volume law, indicating that the effective correlations encoded by the attention mechanism remain far below the theoretical maximum.

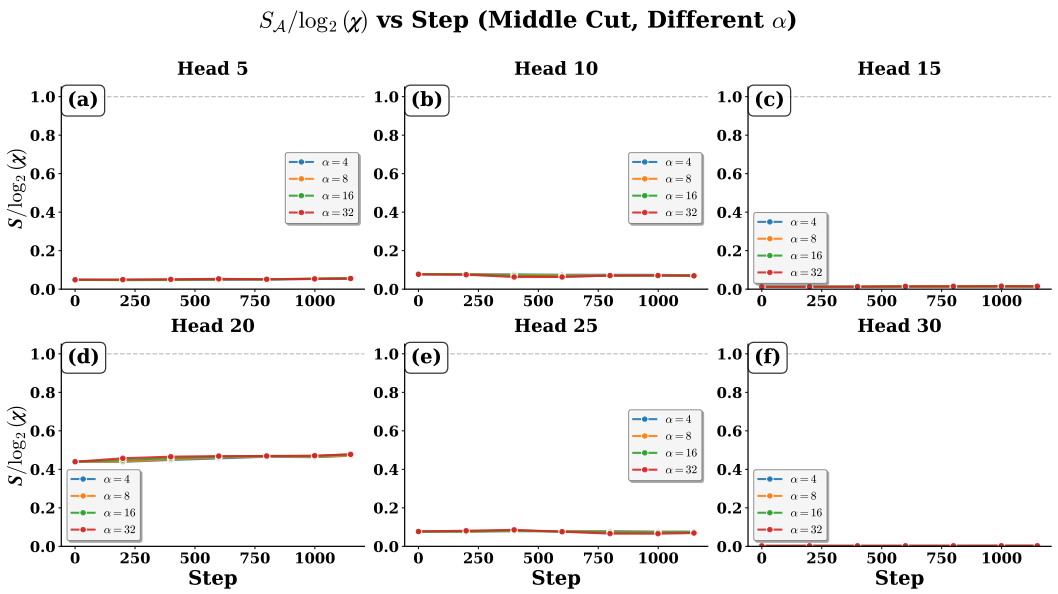

Figure 20: Normalized entanglement entropy $S_A/\log(\chi)$ of six representative attention heads across training steps and LoRA scaling coefficients α. S remains significantly below the theoretical maximum and shows negligible variation with either training time or the choice of α, suggesting a no-hair like behavior.

I MORE EXPERIMENTAL RESULTS

I.1 ABLATION STUDY ON CAUSAL MASKING IN AUTOREGRESSIVE ATTENTION

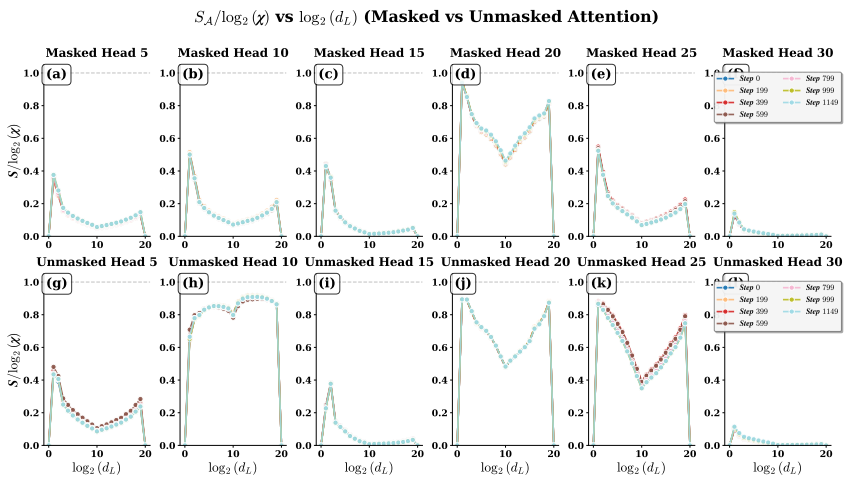

Figure 21: Comparison of masked and unmasked attention entanglement entropy across multiple heads. For each selected head, we plot the normalized entanglement entropy $S_A/\log(\chi)$ as with respect to the bi-partition during training. Removing the causal mask leads to markedly different behaviors across heads: some heads (Head 10, 25) exhibit a substantial increase in entanglement entropy when future positions become accessible, indicating a strong reliance on the causal constraint, while others remain almost unchanged, revealing intrinsically low-entanglement attention patterns that are insensitive to masking.

In autoregressive language models, the causal mask restricts each token to attend only to previous tokens, which may influence the entanglement structure we observe in attention matrices. To understand how the masking operation affects our entanglement analysis, we investigate whether the low entanglement entropy in attention matrices is intrinsic to the learned attention patterns or is an artifact of the causal constraint. To illustrate the contribution of the causal mask to the model's attention structure, we compute *unmasked attention matrices* during evaluation. We attach forward hooks (callback functions that intercept intermediate activations during the forward pass) to each attention layer to capture the pre-RoPE query and key projections,

$$Q_{\text{raw}} = XW_Q, \qquad K_{\text{raw}} = XW_K, \tag{86}$$

together with the corresponding position ids. The projections are reshaped into $[B, H, T, d]$ format with grouped-query attention handled accordingly, after which we reconstruct the rotary position embedding by computing the appropriate cosine and sine rotations, and applying them to each head. With the RoPE-rotated queries and keys, we recompute attention logits

$$S = \frac{QK^\top}{\sqrt{d}}, \tag{87}$$

using the same numerical conventions as the model but without applying the causal mask that normally eliminates future positions. A softmax over these logits yields the unmasked attention matrix

$$A_{\text{unmasked}} = \text{softmax}(S), \tag{88}$$

As shown in FIG. 21, removing the causal mask reveals a clear head-dependent entanglement structure. For some heads (*e.g.*, Head 10 and Head 25), the entanglement entropy increases substantially once future positions become accessible, indicating that the causal boundary had been suppressing a higher effective Schmidt rank. In contrast, other heads show nearly identical masked and unmasked entanglement, implying that their correlation patterns are intrinsically low-entangled and generate only weak entanglement even when the accessible Hilbert space is expanded. This contrast highlights a heterogeneity in the "entanglement capacity" of attention heads.

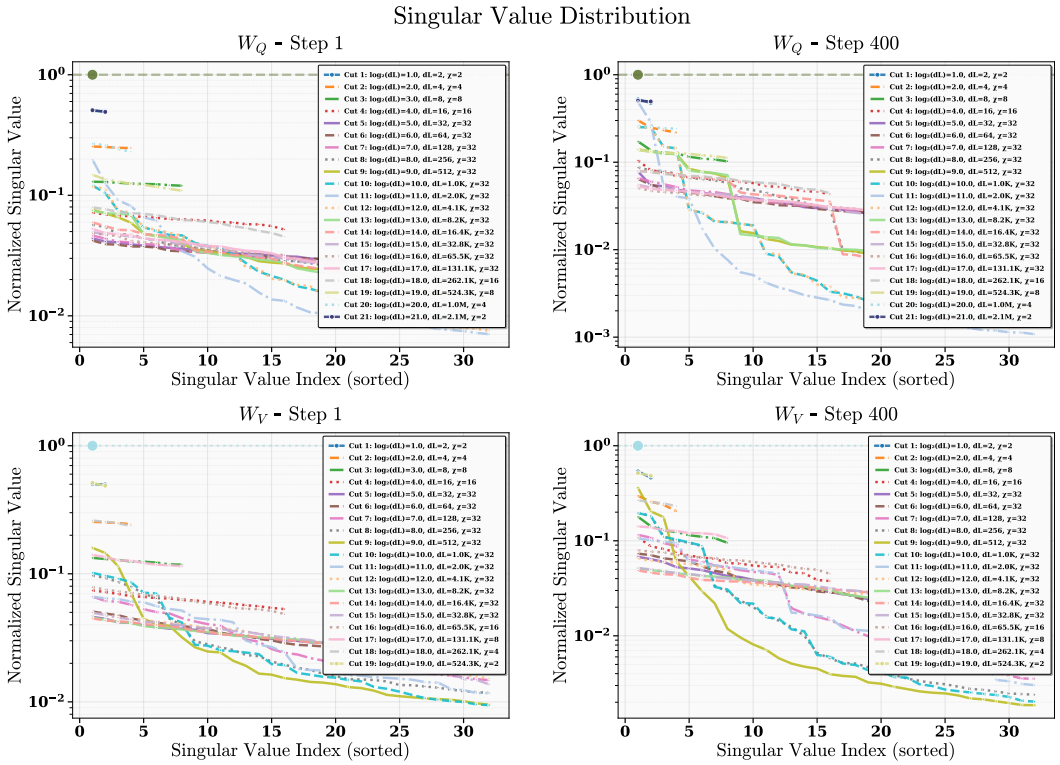

Figure 22: Normalized singular value spectra $\lambda_i^{(\ell)}$ at different cut positions ℓ for W_Q and W_V matrices at the first and last training steps. Each curve corresponds to a different cut position, with the singular values sorted in descending order and normalized by their ℓ_2 norm. The pronounced suppression of intermediate and small singular values near the mid-cut (row–column bi-partition) provides a direct spectral explanation for the entanglement valley observed in the entanglement entropy profile.

I.2 The Entanglement Valley

In the main text, we characterize the internal correlation structure of the updates of weight projection matrices (specifically, the updates of query and value projection matrices ΔW_Q and ΔW_V) by computing the entanglement entropy as a function of the bi-partition (cut) position. By modeling these update matrices as MPS and sweeping the cut from left to right, one obtains an artificial entanglement profile $S(\ell)$. Empirically, this profile exhibits a pronounced depression near the middle cut (*e.g.*, FIG. 3), which we refer to as the *entanglement valley*. For this valley, we consider the LoRA update matrix

$$\Delta W = \frac{\alpha}{r} BA \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}, \qquad B \in \mathbb{R}^{d_{\text{out}} \times r},\ A \in \mathbb{R}^{r \times d_{\text{in}}}, \tag{89}$$

at initialization, where the entries of A and B are i.i.d. with zero mean and finite variance, and $r \ll \min(d_{\text{out}}, d_{\text{in}})$. As discussed in our methods, we flatten ΔW into a vector using row-major order,

$$\text{vec}(\Delta W) \in \mathbb{R}^{d_{\text{out}} d_{\text{in}}}, \tag{90}$$

and factorize the total dimension into powers of two,

$$d_{\text{out}} = 2^m, \qquad d_{\text{in}} = 2^n, \tag{91}$$

(up to an inessential residual factor). We then reshape $\text{vec}(\Delta W)$ into an order-$(m+n)$ tensor with local dimension 2 at each site and perform a left-to-right MPS sweep. The bond ℓ corresponds to a bi-partition between the first ℓ binary indices and the remaining $m+n-\ell$ indices. In particular, the bond $\ell = m$ corresponds exactly to the *row–column bi-partition* $\mathbb{R}^{d_{\text{out}}} \otimes \mathbb{R}^{d_{\text{in}}}$, where the first m binary indices encode the row dimension (output space) and the remaining n indices encode the column dimension (input space), effectively separating the matrix into its row and column subsystems.

Lemma I.1 (Entanglement bound at the row–column cut)**.** *At the row–column cut $\ell = m$, the Schmidt rank of* $\mathrm{vec}(\Delta W)$ *is at most r. Consequently, the von Neumann and Rényi entanglement entropies satisfy*

$$S_{\ell=m}(\Delta W) \;\le\; \log r, \qquad S_{\alpha,\ell=m}(\Delta W) \;\le\; \log r, \quad \forall\, \alpha > 0. \tag{92}$$

For a discussion of Rényi entropies and their relationship to von Neumann entropy, see Appendix D.3.

Proof. Write the rank-r decomposition explicitly:

$$\Delta W = \frac{\alpha}{r} \sum_{a=1}^{r} b_a a_a^\top, \tag{93}$$

where $b_a \in \mathbb{R}^{d_{\mathrm{out}}}$ is the a-th column of B and $a_a^\top \in \mathbb{R}^{d_{\mathrm{in}}}$ is the a-th row of A. Using the identity $\mathrm{vec}(uv^\top) = u \otimes v$, we obtain

$$\mathrm{vec}(\Delta W) = \frac{\alpha}{r} \sum_{a=1}^{r} b_a \otimes a_a \;\in\; \mathbb{R}^{d_{\mathrm{out}}} \otimes \mathbb{R}^{d_{\mathrm{in}}}. \tag{94}$$

Thus $\mathrm{vec}(\Delta W)$ lies in the span of at most r product states. Hence its Schmidt rank across the row–column bi-partition is at most r. A pure state with Schmidt rank at most r has at most r nonzero Schmidt coefficients. The entanglement entropy is therefore maximized by the uniform distribution $1/r$, yielding the bound $S \le \log r$ for von Neumann entropy and, more generally, for all Rényi entropies. □

Proposition I.2 (Entanglement valley of ΔW at initialization)**.** *Assume $d_{\mathrm{out}} = 2^m$ and $d_{\mathrm{in}} = 2^n$ with $m, n \gg 1$. Let $S_\ell(\Delta W)$ denote the entanglement entropy at bond ℓ in the MPS sweep defined above. Then:*

Hard bottleneck. *At the row–column cut $\ell = m$,*

$$S_{\ell=m}(\Delta W) \;\le\; \log r. \tag{95}$$

Typical growth away from the bottleneck. *For cuts ℓ that are well inside the row region or the column region, the entanglement entropy typically follows a volume-law scaling. More precisely, for $\log r \ll \ell \ll m$ (row interior), where ℓ is large enough that the local Hilbert-space dimension 2^ℓ is not the limiting factor, yet sufficiently far from m so that the row–column low-rank structure is not fully exposed, and for $m \ll \ell \ll m + n - \log r$ (column interior), one expects with high probability over the random initialization of A and B that*

$$S_\ell(\Delta W) \approx \min(\ell,\, m + n - \ell) \log 2 - O(1), \tag{96}$$

and in particular $S_\ell(\Delta W) > S_{\ell=m}(\Delta W)$ since $S_{\ell=m}(\Delta W) \le \log r$. Consequently, the entanglement profile $\ell \mapsto S_\ell(\Delta W)$ exhibits a suppression near $\ell = m$, forming an entanglement valley. If $m \approx n$, this suppression appears near the middle of the sweep.

Proof. The hard bottleneck statement is exactly Lemma I.1, which gives $S_{\ell=m}(\Delta W) \le \log r$. For the typical growth statement, we give an intuition based on typical high-dimensional concentration (the phenomenon that random high-dimensional vectors exhibit predictable, concentrated statistical properties, as exemplified by Page's theorem (Page, 1993) for random quantum states). Consider first cuts $\ell < m$, which split only the m row bits. For each $a \in [r]$, the column vector $b_a \in \mathbb{R}^{2^m}$ has i.i.d. entries with zero mean and finite variance, hence after normalization it is an isotropic random vector. When reshaped into an order-m tensor with local dimension 2, such a random vector is well-approximated in typical entanglement statistics by a Haar-random pure state on m qubits. By Page-type concentration (Page, 1993), for any internal cut $1 \ll \ell < m$, one typically has

$$S_\ell(b_a) = \min(\ell, m - \ell) \log 2 - O(1). \tag{97}$$

Now consider $\Delta W = \frac{\alpha}{r} \sum_{a=1}^{r} b_a a_a^\top$. At a cut $\ell < m$, the bi-partition is entirely within the row subsystem, so the special row–column low-rank structure is not yet fully exposed as a *hard* bottleneck. Empirically and heuristically, in the interior regime $\log r \ll \ell \ll m$, the effective Schmidt rank available across the cut is dominated by the local Hilbert-space dimension 2^ℓ rather than by the global

factorization rank r, and the reduced spectrum is close to that of a typical high-dimensional state. This leads to a near-maximal, volume-law scaling

$$S_\ell(\Delta W) \approx \ell \log 2 - O(1), \qquad \log r \ll \ell \ll m, \tag{98}$$

and hence $S_\ell(\Delta W) > S_{\ell=m}(\Delta W)$ since $S_{\ell=m}(\Delta W) \leq \log r$. As ℓ approaches m from below, the cut begins to probe the global row–column factorization of ΔW, and the entanglement is progressively constrained, interpolating between the interior volume-law behavior and the bottleneck bound at $\ell = m$. An analogous argument applies to cuts $\ell > m$ inside the column subsystem, yielding

$$S_\ell(\Delta W) \approx (m + n - \ell) \log 2 - O(1), \qquad m \ll \ell \ll m + n - \log r. \tag{99}$$

Therefore the profile $\ell \mapsto S_\ell(\Delta W)$ is suppressed near $\ell = m$, forming an entanglement valley; when $m \approx n$, this suppression appears near the middle of the sweep. □

Why the specific valley position: a symmetry perspective. From a symmetry perspective, the appearance of the entanglement valley at a specific bi-partition can be understood as a consequence of an explicit structural symmetry breaking induced by the low-rank parameterization. For a generic full-rank matrix, the vectorized state $\mathrm{vec}(W)$ does not distinguish one internal bi-partition from another, and the entanglement profile is approximately invariant under translations of the cut position. In contrast, the factorized form $\Delta W = BA$ singles out a preferred decomposition between the row and column spaces, $\mathbb{R}^{d_{\text{out}}} \otimes \mathbb{R}^{d_{\text{in}}}$. The corresponding cut $\ell = m$ is the unique bi-partition that respects this structural decomposition, while all other cuts mix the two subspaces and are therefore insensitive to the low-rank constraint. As a result, the effective cut-translation symmetry is explicitly broken, and the entanglement bottleneck associated with the rank-r coupling becomes manifest only at the row–column cut, giving rise to a localized entanglement valley.

Spectral Interpretation of the Entanglement Valley. The entanglement entropy across a given cut is fully determined by the singular value spectrum of the corresponding bipartite decomposition. Let $\{\lambda_i^{(\ell)}\}_{i=1}^{r_\ell}$ denote the normalized singular values at cut ℓ, satisfying $\sum_i (\lambda_i^{(\ell)})^2 = 1$. The entanglement entropy is $S(\ell) = -\sum_i (\lambda_i^{(\ell)})^2 \log\left(\lambda_i^{(\ell)}\right)^2$, therefore directly reflects how spectral weight is distributed across modes: **flat spectra yield high S, whereas concentrated spectra yield low S.** To probe this mechanism, we visualize the singular value spectra at different cut positions. For each cut ℓ, we perform an SVD of the induced bipartite matrix, sort the singular values in descending order, and normalize them by their ℓ_2 norm. We then plot the normalized singular values $\lambda_i^{(\ell)}$ against their sorted index. See FIG. 22. Comparing these spectra across cuts, especially between boundary and middle regions, reveals a pronounced suppression of intermediate and small singular values near the mid-cut, providing a direct empirical spectral explanation for the observed entropy dip.

Measuring Distance from Random Matrix Behavior. To better understand the statistical properties of these singular value distributions and distinguish them from random matrix behavior, we compare them against theoretical predictions from random matrix theory. Specifically, we consider the *Marchenko-Pastur (MP) distribution* (Marčenko & Pastur, 1967), which describes the limiting eigenvalue distribution of large random matrices. For a random matrix $R \in \mathbb{R}^{d_{\min} \times d_{\max}}$ with i.i.d. entries of zero mean and finite variance, where $d_{\min} \leq d_{\max}$ and $c = d_{\min}/d_{\max}$, the normalized eigenvalues $x_i = d_{\min} \cdot \lambda_i$ of the sample covariance matrix $RR^\top/d_{\max}$ converge to the MP distribution as the dimensions grow. Here, λ_i denotes the normalized eigenvalues (or singular values squared) of the reduced density matrix, and the scaling $x_i = d_{\min} \cdot \lambda_i$ is the standard normalization in MP theory that makes the distribution shape depend only on the aspect ratio c rather than the absolute dimensions, ensuring the theoretical results are scale-invariant. The MP distribution $\rho(x)$ represents the probability density function describing how these scaled eigenvalues x_i are distributed. The MP distribution provides a natural baseline for understanding how the singular value spectra of our MPS-decomposed matrices deviate from purely random behavior. FIG. 23 shows the theoretical MP density curves $\rho(x)$ for different values of c, where the vertical axis represents the probability density and the horizontal axis represents the normalized eigenvalue $x = d_{\min} \cdot \lambda$. The figure illustrates how the distribution shape changes from highly concentrated (small c) to more spread out (large c, approaching $c = 1$). FIG. 24 compares the singular value distributions from our MPS-decomposed weight matrices across different training steps with the theoretical Marchenko-Pastur distribution. In this figure, we plot the empirical density of normalized eigenvalues λ (from the reduced density matrices at different cuts) against the theoretical MP density $\rho(x)$, where $x = d_{\min} \cdot \lambda$. A key observation

is that, while the entanglement entropy in regions far from the mid-cut follows a volume law and can reach the theoretical maximum (see FIG. 3), the internal structure of these matrices is not just that of a fully random matrix. The empirical distributions from W_Q and W_V exhibit deviations from the MP baseline, particularly in the tail regions and the overall shape of the spectrum. This indicates that despite achieving high entanglement entropy, the learned weight matrices have acquired structured correlations that distinguish them from purely random configurations. We include the random matrix distribution as a baseline reference to highlight these structural differences, demonstrating that the volume-law scaling observed in the entanglement profile does not imply random matrix statistics, but rather reflects learned structured patterns encoded in the weight matrices.

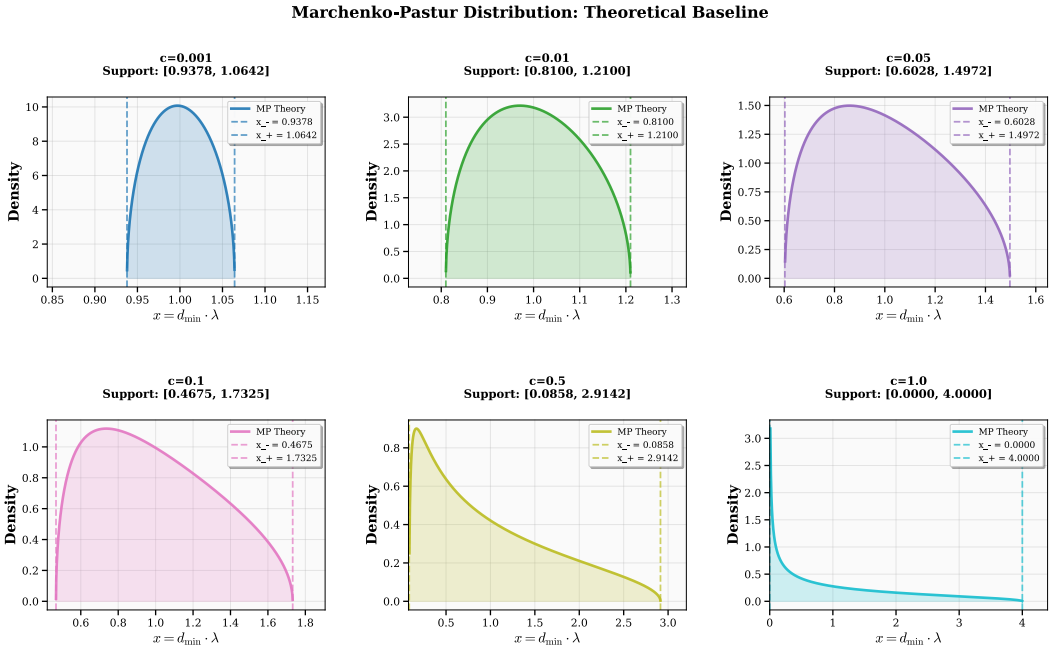

Figure 23: Marchenko-Pastur distribution: theoretical baseline. The figure shows the theoretical density curves for different values of the parameter $c = d_{\min}/d_{\max}$, where $d_{\min}$ and $d_{\max}$ are the smaller and larger dimensions of a random matrix, respectively. The MP distribution describes the limiting eigenvalue distribution of large random matrices with i.i.d. entries, providing a reference for comparing the singular value spectra observed in our MPS-decomposed weight projection matrices.

I.3 Limits of the No-Hair Property Under Extreme Scaling Coefficients

In the main text, we demonstrate that the attention matrices exhibit a no-hair property, where the entanglement entropy remains robust to variations in the scaling coefficient α within a moderate range (see FIG. 20 and FIG. 11). However, this robustness may break down when α deviates substantially from typical values used in practice. To probe the limits of this no-hair behavior, we investigate the entanglement structures of attention matrices under largely deviated scaling coefficients $\alpha \in \{16, 512, 4096\}$, extending beyond the regime explored in the main text. FIG. 25 shows the normalized entanglement entropy $S_A/\log(\chi)$ across training steps for several attention heads under these extreme α values. Compared to the moderate regime where $\alpha \in \{8, 16, 32, 64\}$ (FIG. 11), we observe that when α increases to 512, the entanglement curves remain approximately invariant, maintaining the no-hair-like property observed in the main text. However, when α scales further to 4096, pronounced differences emerge: for several attention heads, $\alpha = 4096$ produces substantially higher entanglement entropy S, with a sudden increase during early training steps, while $\alpha = 16$ and $\alpha = 512$ remain almost flat throughout training. These results highlight that the no-hair property observed in the main text holds within a certain range of α values, but breaks down under extreme scaling. Nevertheless, compared to the dramatic variations in entanglement structures observed in the token embedding space (see FIG. 16 and FIG. 15), where the embedding space exhibits substantial

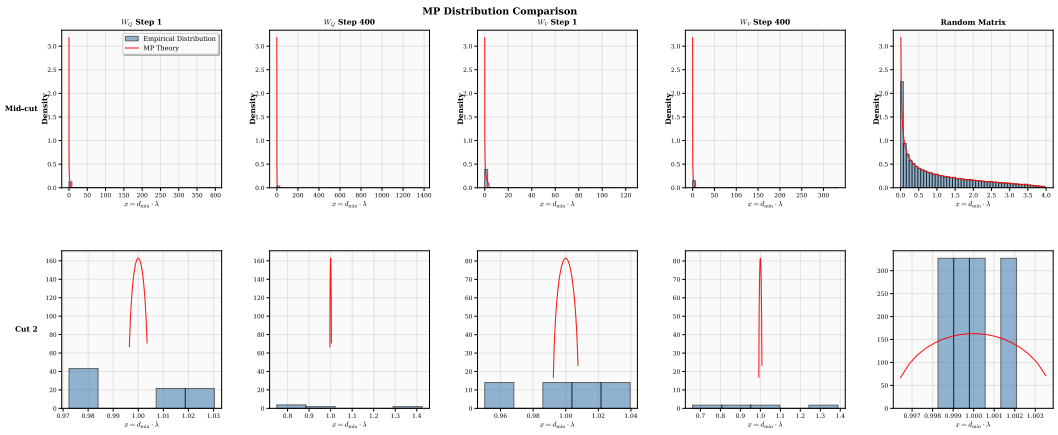

Figure 24: Comparison of normalized eigenvalue distributions from reduced density matrices at different cuts in MPS-decomposed weight projection matrices with the theoretical Marchenko-Pastur distribution. The figure shows empirical density distributions of normalized eigenvalues λ (from reduced density matrices at different cuts) from ΔW_Q and ΔW_V matrices, plotted against the theoretical MP density $\rho(x)$ where $x = d_{\min} \cdot \lambda$, illustrating how the observed eigenvalue spectra deviate from random matrix behavior.

sensitivity to α with entanglement profiles varying significantly across different scaling coefficients, we still consider the attention behavior as exhibiting a no-hair-like property.

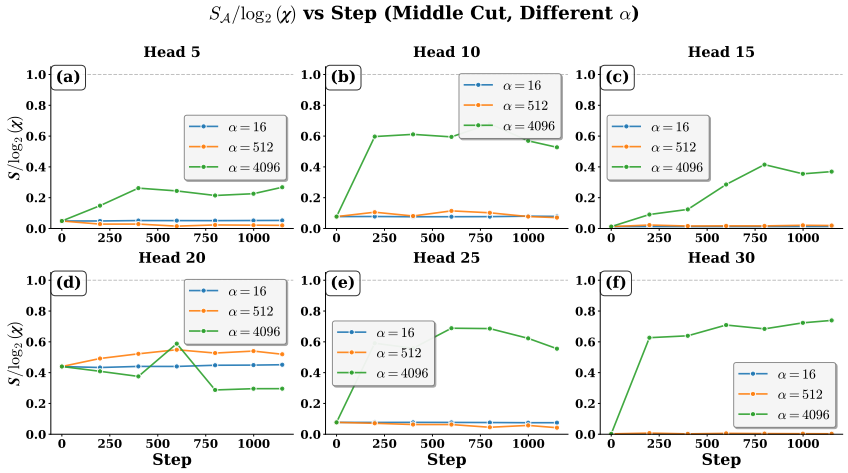

Figure 25: Normalized entanglement entropy $S_A/\log(\chi)$ for attention heads under extreme scaling coefficients $\alpha \in \{16, 512, 4096\}$, probing the limits of the no-hair property. Compared to the moderate regime (FIG. 11), when $\alpha = 512$ the entanglement curves remain approximately invariant, maintaining the no-hair-like property. However, when α scales further to 4096, pronounced differences emerge: for several attention heads, $\alpha = 4096$ produces substantially higher entanglement entropy S, with a sudden increase during early training steps, while $\alpha = 16$ and $\alpha = 512$ remain almost flat throughout training. These results demonstrate that the no-hair property holds within a certain range of α values but breaks down under extreme scaling.

I.4 Artificial Entanglement Profiling: LLaMA-3.2-1B, OpenThoughts3 Dataset

We extend our empirical analysis to the OpenThoughts3 dataset. The results verify that the entanglement structures and no-hair phenomena observed in the Tulu3 dataset (Sec. 5) remain consistent across different data distributions.

Artificial entanglement profiling in Attention matrix. FIG. 26, 27, and 28 demonstrate that the attention matrices on OpenThoughts3 exhibit the same approximate area-law scaling with logarithmic correction as observed on Tulu3. The normalized entanglement entropy $S_A/\log(\chi)$ remains significantly below the theoretical maximum and robust to variations in the scaling coefficient α, confirming that the no-hair property holds across datasets. Similarly, FIG. 29 shows that the output operator $\mathcal{O} = XX^\top$ maintains entanglement significantly below the theoretical maximum, consistent with the filtering behavior observed in the main text.

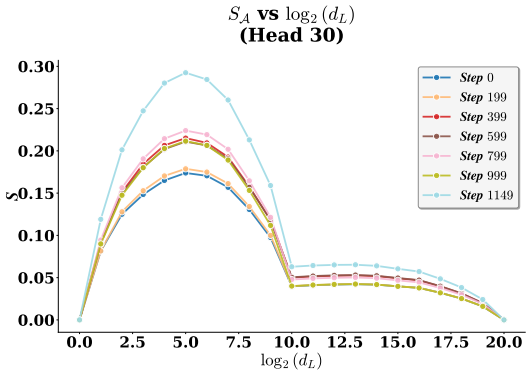

Figure 26: Entanglement entropy S_A of the attention matrix (Head 30) with respect to the bi-partition position across training steps for LLaMA-3.2-1B fine-tuned on OpenThoughts3 dataset.

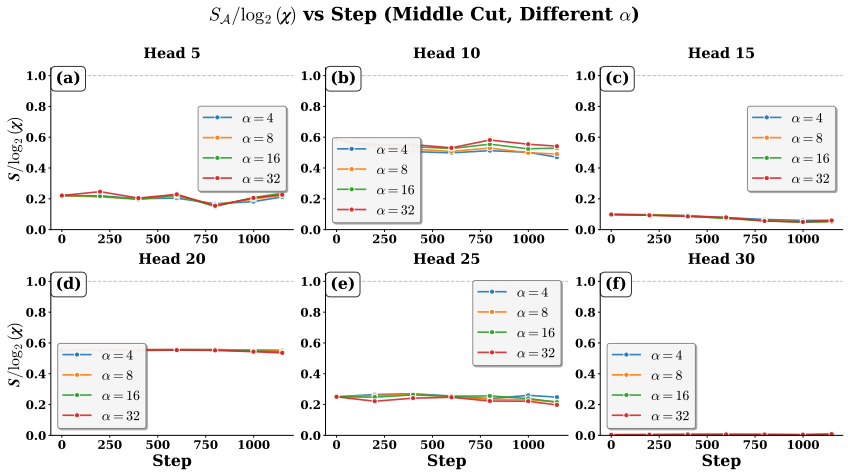

Figure 27: Entanglement entropy $S_A/\log(\chi)$ across training steps for several attention heads under less-deviated scaling coefficients $\alpha \in \{4, 8, 16, 32\}$ for LLaMA-3.2-1B fine-tuned on OpenThoughts3 dataset. See FIG. 20 for details.

Token embedding space artificial entanglement profiling: volume-law and entanglement valley. FIG. 30 shows that MPS adaptation on OpenThoughts3 exhibits the same characteristic entanglement valley structure as observed in the main text, confirming that the volume-law behavior in embedding space is robust across datasets.

Optimization landscape comparison. FIG. 31 and 32 demonstrate that the optimization dynamics and relative performance of LoRA and MPS adaptation on OpenThoughts3 mirror the observations on Tulu3. The optimal learning rate regions and the shift in optimal learning rates with α follow the same patterns as described in Sec. 5, confirming the consistency of these phenomena across datasets.

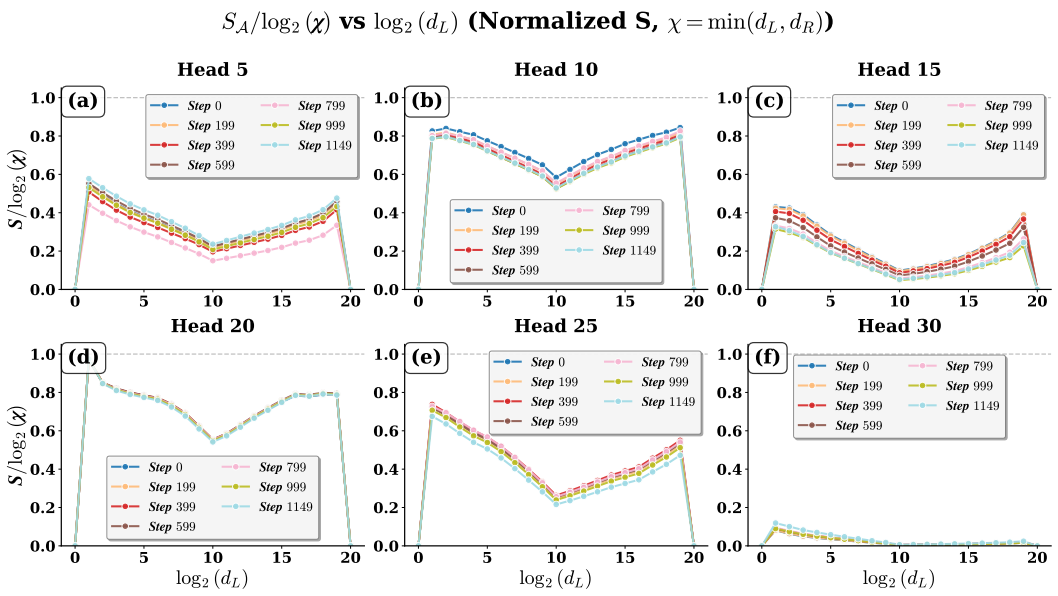

Figure 28: Normalized entanglement entropy $S_A/\log(\chi)$ of six representative attention heads across training steps for LLaMA-3.2-1B fine-tuned on OpenThoughts3 dataset, where $\chi = \min(d_L, d_R)$ denotes the maximal entanglement capacity allowed by the bi-partition.

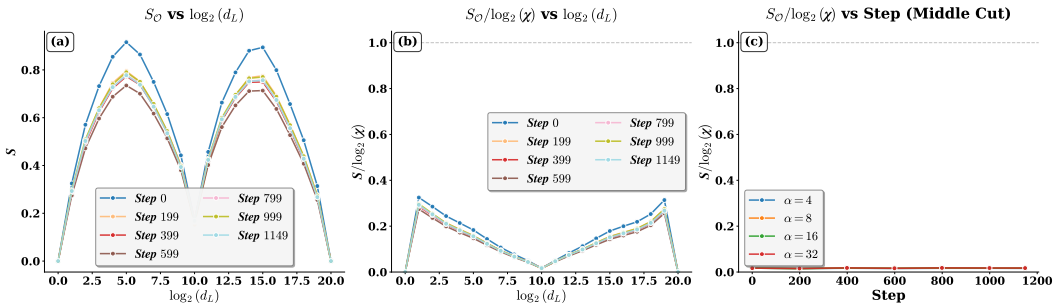

Figure 29: Entanglement entropy $S_\mathcal{O}$ of the output operator $\mathcal{O} = XX^\top$ across bi-partition positions and training steps for LLaMA-3.2-1B fine-tuned on OpenThoughts3 dataset.

I.5 Artificial Entanglement Profiling: LLaMA-3.1-8B, Tulu3 Dataset

We extend our analysis to the larger LLaMA-3.1-8B model fine-tuned on the Tulu3 dataset to validate the universality of our findings across model scales. The results demonstrate that the entanglement structures and no-hair phenomena observed in the 1B model (Sec. 5) persist consistently in larger architectures.

Artificial entanglement profiling in Attention matrix. FIG. 33, 35, and 34 show that the attention matrices in the 8B model exhibit the same approximate area-law scaling with logarithmic correction as observed in the 1B model. The normalized entanglement entropy $S_A/\log(\chi)$ remains significantly below the theoretical maximum and robust to variations in the scaling coefficient α, confirming that the no-hair property holds across model scales. Similarly, FIG. 36 demonstrates that the output operator $\mathcal{O} = XX^\top$ maintains entanglement significantly below the theoretical maximum, consistent with the filtering behavior observed in smaller models.

Token embedding space artificial entanglement profiling: volume-law and entanglement valley. FIG. 37 and 39 reveal that the embedding space (W_Q and W_V) in the 8B model exhibits the same volume-law behavior with a characteristic entanglement valley as observed in the 1B model. The valley deepens during training in LoRA and shows sensitivity to the scaling coefficient α,

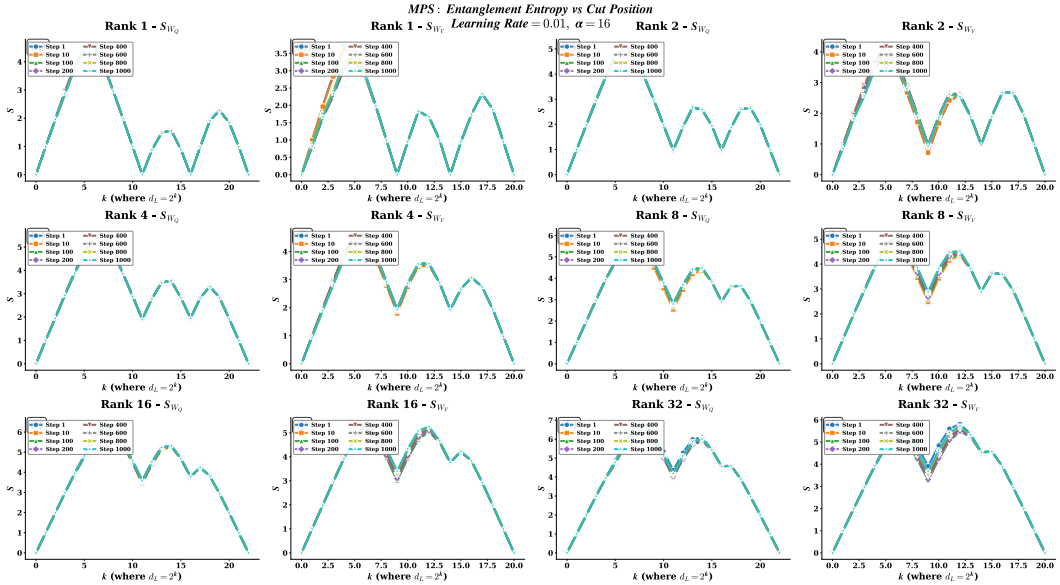

Figure 30: Artificial entanglement profiling of the MPS adaptation of ΔW for ranks $r \in \{1, 2, 4, 8, 16, 32\}$ with $\alpha = 16$ across training steps for LLaMA-3.2-1B fine-tuned on OpenThoughts3 dataset.

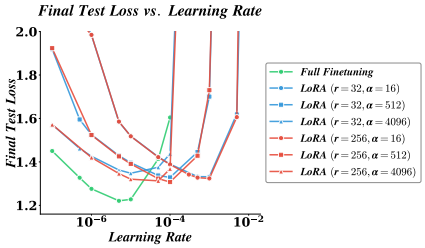

Figure 31: Final test loss with respect to learning rate under different LoRA configurations for LLaMA-3.2-1B fine-tuned on OpenThoughts3 dataset, including a comparison with full fine-tuning.

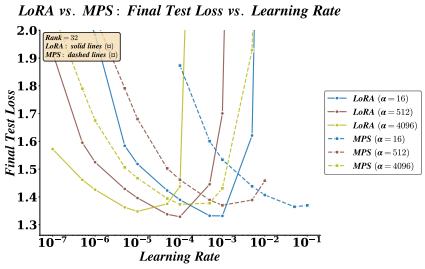

Figure 32: Final test loss of LoRA and MPS adaptation across multiple learning rates and scaling coefficients α for LLaMA-3.2-1B fine-tuned on OpenThoughts3 dataset.

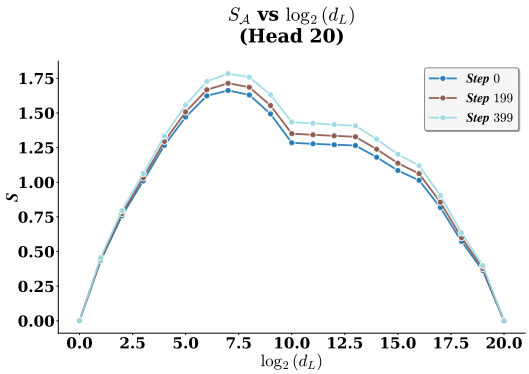

Figure 33: Entanglement entropy S_A of the attention matrix (Head 20) with respect to the bi-partition position across training steps for LLaMA-3.1-8B fine-tuned on Tulu3 dataset.

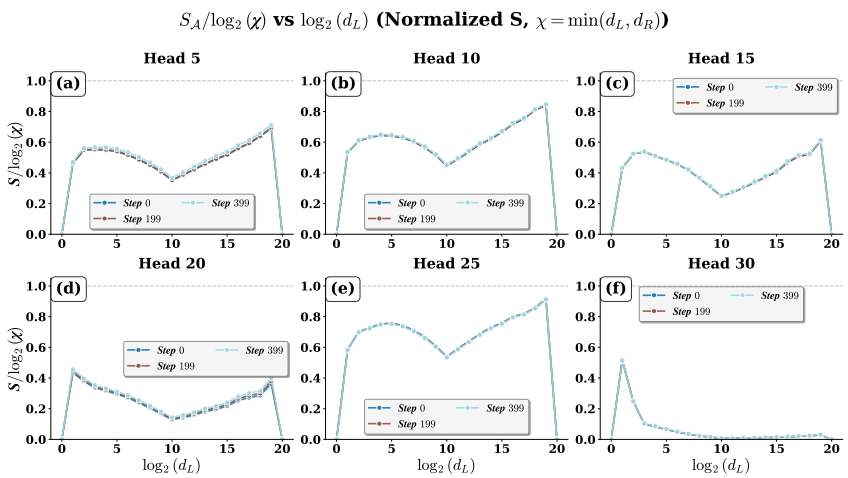

Figure 34: Normalized entanglement entropy $S_A/\log(\chi)$ of six representative attention heads across training steps for LLaMA-3.1-8B fine-tuned on Tulu3 dataset, where $\chi = \min(d_L, d_R)$ denotes the maximal entanglement capacity allowed by the bi-partition.

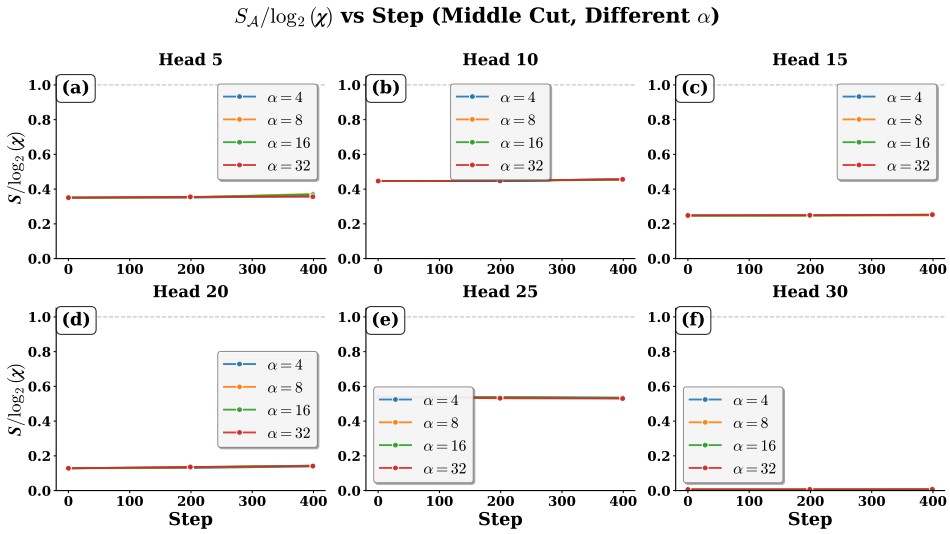

Figure 35: Normalized entanglement entropy $S_A/\log(\chi)$ across training steps for several attention heads under different scaling coefficients α for LLaMA-3.1-8B fine-tuned on Tulu3 dataset.

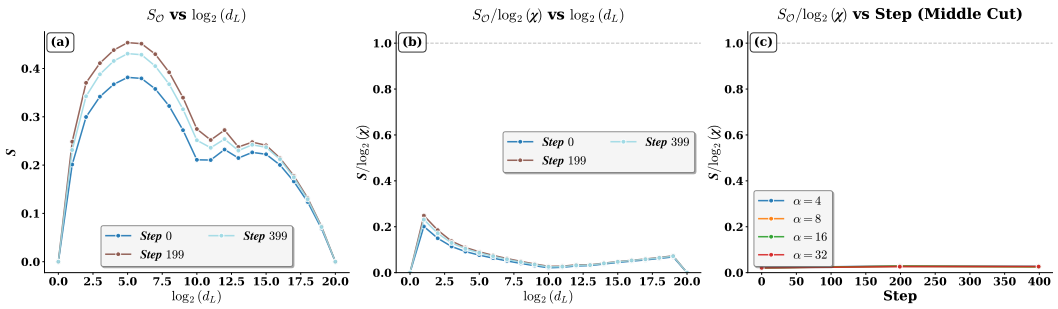

Figure 36: Entanglement entropy $S_{\mathcal{O}}$ of the output operator $\mathcal{O} = XX^\top$ across bi-partition positions and training steps for LLaMA-3.1-8B fine-tuned on Tulu3 dataset.

consistent with the findings in Sec. 5. This confirms that the volume-law entanglement profile and its hyperparameter sensitivity are fundamental properties independent of model scale.

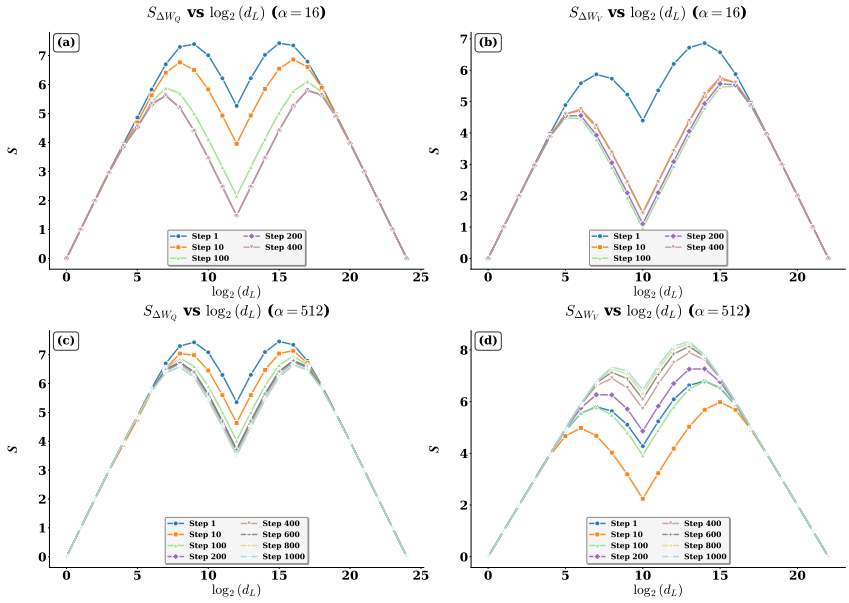

Figure 37: Artificial entanglement profiling of ΔW_Q and ΔW_V with respect to the cut position across training steps and scaling coefficients $\alpha \in \{16, 512\}$ for LLaMA-3.1-8B fine-tuned on Tulu3 dataset.

Optimization landscape comparison. FIG. 38 compares LoRA and MPS adaptation strategies for the 8B model. Both methods achieve comparable performance, with optimal learning rates shifting as α increases, mirroring the behavior observed in the 1B model (Sec. 5). This consistency further validates the generalizability of our theoretical framework across model scales.

I.6 Artificial Entanglement Profiling: LLaMA-3.1-8B, OpenThoughts3 Dataset

We further validate our findings by analyzing the LLaMA-3.1-8B model fine-tuned on the OpenThoughts3 dataset. The results confirm that the entanglement structures and no-hair phenomena are robust across both model scales and dataset distributions, reinforcing the universality.

Artificial entanglement profiling in Attention matrix. FIG. 40 and 41 demonstrate that the attention matrices in the 8B model on OpenThoughts3 exhibit the same area-law scaling with logarithmic correction and no-hair property as observed in both the 1B model (Sec. 5) and the 8B model on Tulu3. The normalized entanglement entropy remains significantly below the theoretical maximum

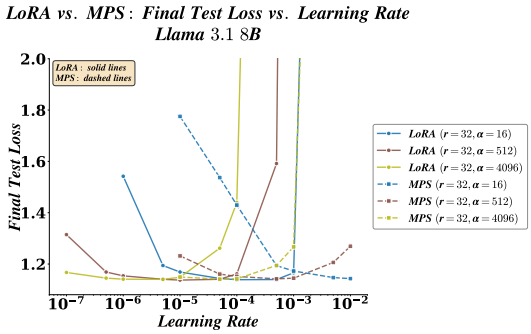

Figure 38: Final test loss of LoRA and MPS adaptation across multiple learning rates and scaling coefficients α for LLaMA-3.1-8B fine-tuned on Tulu3 dataset.

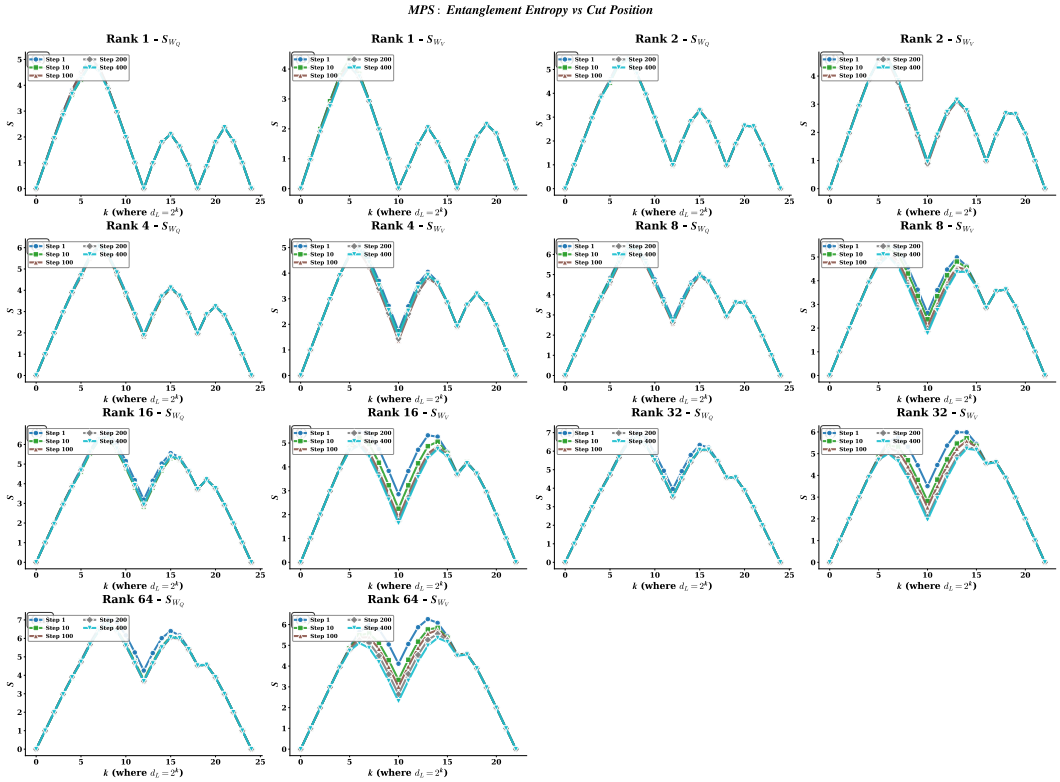

Figure 39: Artificial entanglement profiling of the MPS adaptation of ΔW for ranks $r \in \{1, 2, 4, 8, 16, 32\}$ with $\alpha = 16$ across training steps for LLaMA-3.1-8B fine-tuned on Tulu3 dataset.

and robust to variations in α, confirming that these phenomena are universal across different datasets and model scales.

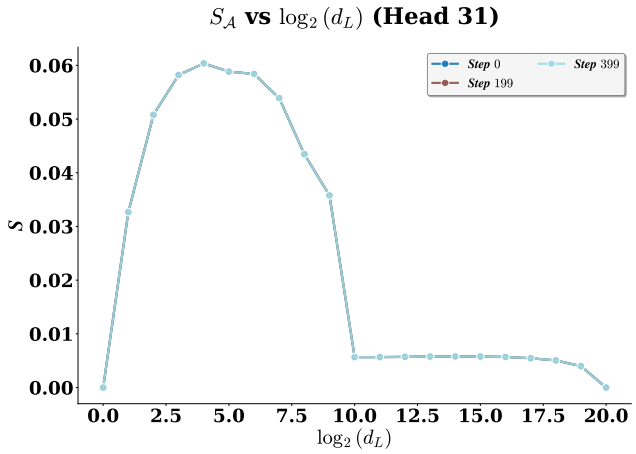

Figure 40: Entanglement entropy S_A of the attention matrix (Head 31) with respect to the bi-partition position across training steps for LLaMA-3.1-8B fine-tuned on OpenThoughts3 dataset.

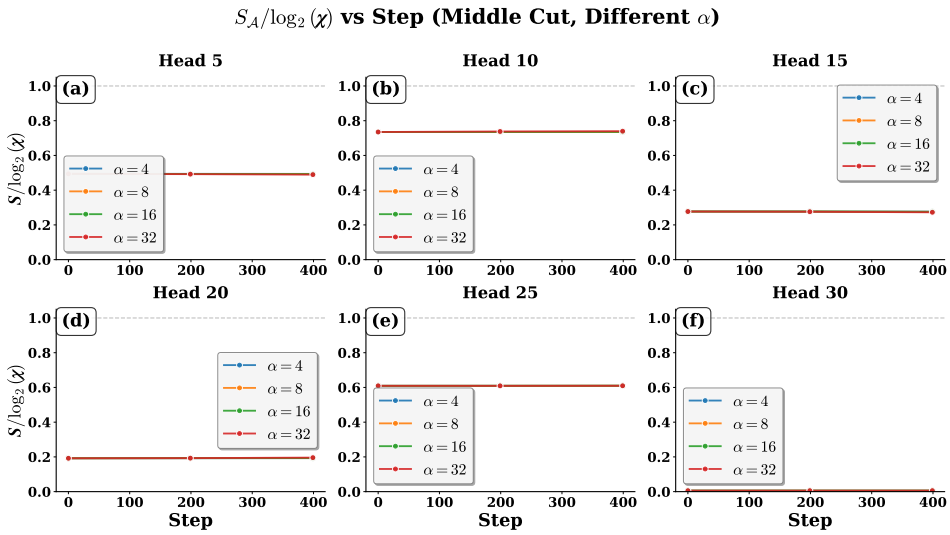

Figure 41: Normalized entanglement entropy $S_A/\log(\chi)$ across training steps for several attention heads under different scaling coefficients α for LLaMA-3.1-8B fine-tuned on OpenThoughts3 dataset.

Token embedding space artificial entanglement profiling: volume-law and entanglement valley. FIG. 42 reveals that the embedding space (W_Q and W_V) in the 8B model on OpenThoughts3 exhibits the same volume-law behavior with a characteristic entanglement valley as observed in both the 1B model (Sec. 5) and the 8B model on Tulu3. The valley deepens during training in LoRA and shows sensitivity to the scaling coefficient α, consistent with the findings in Sec. 5. This confirms that the volume-law entanglement profile and its hyperparameter sensitivity are fundamental properties independent of both model scale and dataset distribution.

Optimization landscape comparison. FIG. 43 compares LoRA and MPS adaptation strategies for the 8B model on OpenThoughts3. The consistent performance patterns—where both methods achieve comparable results and optimal learning rates shift with α—mirror the observations in other model-dataset combinations, providing strong evidence for the generalizability of tensor-network-based adaptation methods and the robustness of the no-hair phenomenon across diverse experimental settings.

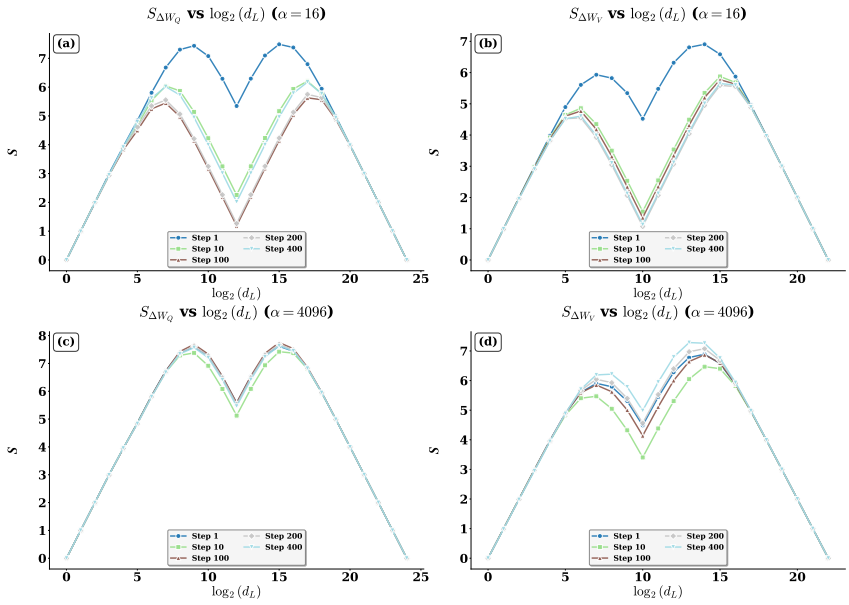

Figure 42: Artificial entanglement profiling of ΔW_Q and ΔW_V with respect to the cut position across training steps and scaling coefficients $\alpha \in \{16, 4096\}$ for LLaMA-3.1-8B fine-tuned on OpenThoughts3 dataset.

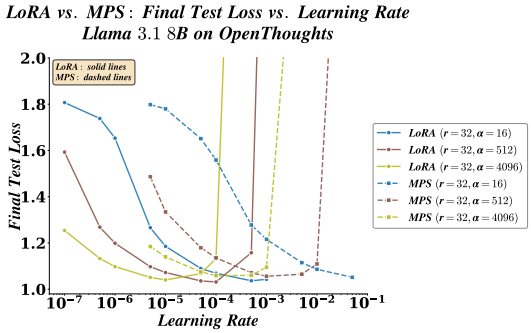

Figure 43: Final test loss of LoRA and MPS adaptation across multiple learning rates and scaling coefficients α for LLaMA-3.1-8B fine-tuned on OpenThoughts3 dataset.

J Dense and LoRA Fine-Tuning of the Same Q/V Projections

The comparison between FFT and LoRA in Sec. 5 differs in which parameters are trained: FFT updates the full model, whereas LoRA adapts only the query and value projections. Here we ask whether the difference in internal artificial entanglement profiling persists when both methods train the same projections. We compare dense updates with LoRA updates to every W_Q and W_V in the model, keeping all other weights frozen. We call the first method *Dense Q/V* to distinguish this control from the original FFT experiment.

Experimental Setups for the Matched-Scope Comparison. The comparison uses `meta-llama/Llama-3.2-1B-Instruct` and `allenai/tulu-3-sft-mixture`. Both methods target all Q/V projections in all 16 layers. LoRA Q/V uses rank $r = 256$ and scaling coefficient $\alpha = 16$.

Selection of the Shared Learning Rate. We run both methods for 200 optimizer steps at each of three candidate learning rates: at every rate one method improves less than the other, and we keep the rate at which that lagging method improves most, see Table 2. This procedure selects 3×10^{-5}. The same analysis examples serve both this selection and the evaluation below.

Table 2: Selection of the shared learning rate at step 200. Entries are final-to-initial analysis-NLL ratios, and the last column is the worse ratio across the two methods. Lower is better.

Learning rate	Dense Q/V	LoRA Q/V	Worst-method ratio
3×10^{-5}	0.578358	0.590534	**0.590534**
1×10^{-4}	0.597033	0.581250	0.597033
3×10^{-4}	0.713133	0.578082	0.713133

Artificial Entanglement Profiling and the Signed Valley Depth.

We profile the updates at layers $0, 8, 15$, representing the first, middle and last layers, at steps $\{1, 10, 100, 200, 400, 600, 800, 1000\}$. The Dense update is $W_t - W_0$, whereas the LoRA update is the effective scaled product $(\alpha/r)BA$. Each matrix is profiled without SVD truncation.

The profiling uses the von Neumann entropy of Sec. 4, which we report in bits and without any normalization that varies across bi-partition positions. We denote it as $S_{\mathrm{raw}}(k)$. The *central partition* k_0 separates all output factors from all input factors, so that $d_L(k_0) = d_{\mathrm{out}}$, though it need not be the midpoint of the MPS chain. To quantify the *central dip* relative to both sides, we define the signed valley depth

$$D = \min\left\{ \max_{k<k_0} S_{\mathrm{raw}}(k),\ \max_{k>k_0} S_{\mathrm{raw}}(k) \right\} - S_{\mathrm{raw}}(k_0). \tag{100}$$

Consequently, $D > 0$ indicates a two-sided valley, whereas $D \leq 0$ means that the central partition is not below both side maxima. A negative depth does not necessarily imply a central peak.

We compute each measurement separately for every seed, module, layer, and saved step. Means and sample standard deviations use the three training seeds, because Q/V matrices, layers, bi-partition positions and checkpoints are not independent training replicates. Method differences are paired within each seed before their mean and standard deviation are computed.

Artificial Entanglement Profiling of $S_{\Delta W_Q}$ and $S_{\Delta W_V}$ under Matched Trainable Scope. We compare the artificial entanglement profiling of Dense Q/V and LoRA Q/V across bonds under the shared learning rate, see FIG. 44. At step 1000, LoRA retains a pronounced *entanglement valley* in every module and layer combination, whereas Dense exhibits no two-sided dip at all: its signed depth remains negative throughout. Table 3 reports the corresponding depths together with the entropy contrasts at the central partition, where the paired LoRA-minus-Dense depth ranges from 3.544 to 4.167 bits across the six combinations and the LoRA entropy at the central partition is lower throughout.

Evolution of the Valley under Matched Trainable Scope. The contrast is not confined to the final checkpoint, see FIG. 45: across all eight saved steps the paired depth remains positive for every seed, module and layer, with the smallest individual difference at 0.934 bits. Meanwhile, the valley in LoRA deepens as training progresses while the Dense depths stay near or below zero, reproducing

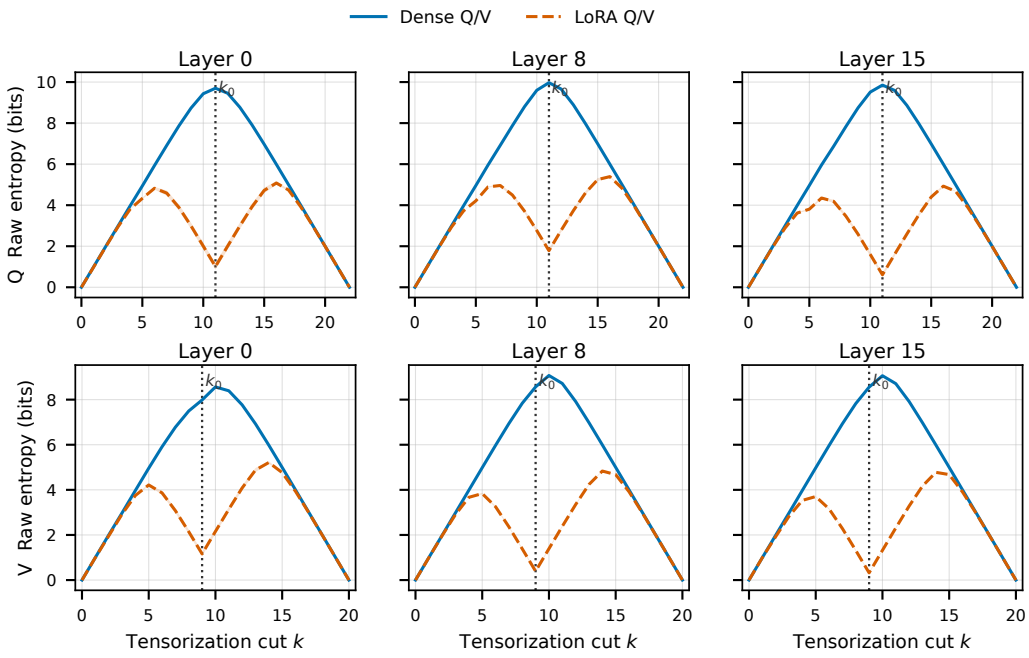

Figure 44: Artificial entanglement profiling of ΔW_Q and ΔW_V at step 1000, when Dense and LoRA update the same Q/V projections under one shared learning rate. Rows separate W_Q and W_V, and columns show layers $0, 8, 15$. Curves and bands are mean $\pm$ sample standard deviation over the three paired training seeds. Dotted lines mark the central partition k_0, and no normalization varying across bi-partition positions is applied.

Table 3: Signed valley depths at step 1000 and the entropy contrasts at the central partition, in bits. Entries are mean $\pm$ sample standard deviation over three seeds. Both ΔD and $\Delta S(k_0)$ are paired LoRA-minus-Dense differences. Negative Dense depths indicate no two-sided valley, not necessarily a central peak.

Module	Layer	Dense D	LoRA D	ΔD	$\Delta S(k_0)$
Q	0	-0.266 ± 0.004	3.816 ± 0.091	4.082 ± 0.095	-8.689 ± 0.181
Q	8	-0.369 ± 0.003	3.177 ± 0.035	3.547 ± 0.032	-8.186 ± 0.085
Q	15	-0.341 ± 0.001	3.741 ± 0.039	4.082 ± 0.040	-9.245 ± 0.090
V	0	-0.487 ± 0.002	3.056 ± 0.059	3.544 ± 0.058	-6.824 ± 0.098
V	8	-0.725 ± 0.004	3.442 ± 0.005	4.167 ± 0.003	-8.173 ± 0.029
V	15	-0.729 ± 0.001	3.382 ± 0.017	4.111 ± 0.018	-8.233 ± 0.030

under matched trainable scope the time-dependent contrast between FIG. 3 and FIG. 2. The bands describe variation across training seeds, not variation across layers or bi-partition positions.

At step 1000 the analysis NLL is 1.551992 ± 0.000193 for Dense Q/V and 1.568809 ± 0.000354 for LoRA Q/V, a paired difference of 0.016817 ± 0.000358, so Dense reaches the lower loss under this shared setting.

Combining these results, **the difference in adaptation scope in the original FFT–LoRA comparison is not required for the entanglement valley to appear**: the valley persists once both methods update exactly the same projections.

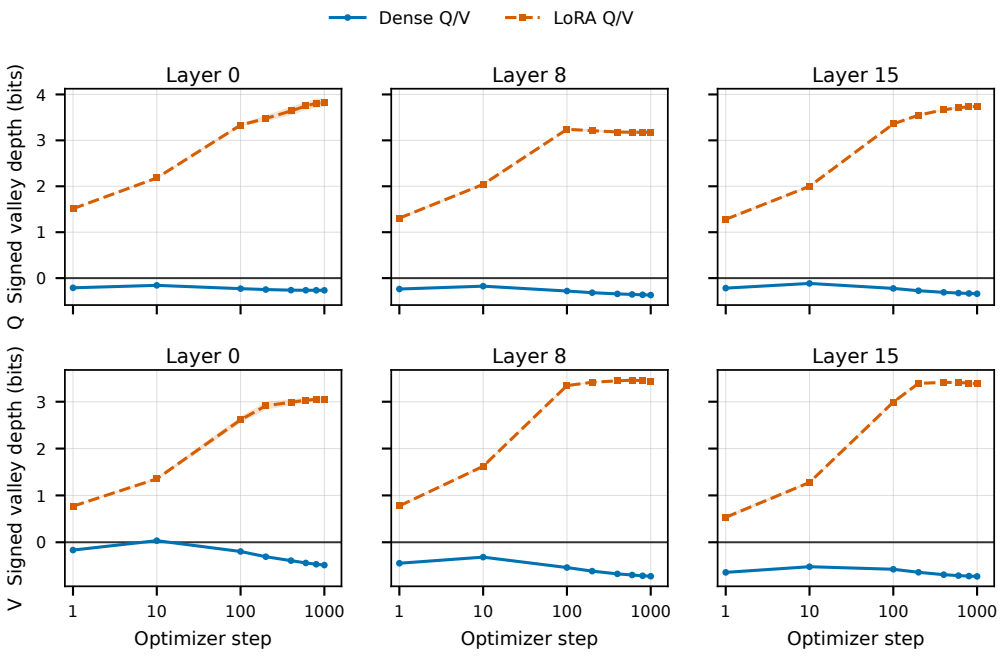

Figure 45: Signed valley depth during training for the same Q/V comparison. Curves and bands are seed mean ± sample standard deviation over the three paired training seeds. The zero line separates a two-sided valley ($D > 0$) from no two-sided valley ($D \leq 0$). Optimizer steps are shown on a logarithmic axis.

K Swapping Singular Values and Singular Vectors

Appendix J shows that the entanglement valley survives when Dense and LoRA update the same Q/V projections, yet it does not tell us which part of the learned update carries the valley. An update is determined by its singular values together with its singular vectors, and either could in principle produce the *central dip*. Here we separate the two by transplanting one of them across methods, without retraining.

Construction of the Hybrid Updates. We reuse the stored step 1000 matrices of the matched-scope comparison, which gives 18 paired strata over three seeds and the six W_Q–W_V and layer combinations. For each matrix we compute an economy SVD and truncate every factor to $r = 256$, so that all four arms share the same nominal rank. Denoting the spectrum source by the first subscript and the singular-vector source by the second, we construct

$$\begin{aligned} M_{DD} &= U_D \Sigma_D V_D^\top, \quad M_{LD} = U_D \Sigma_L V_D^\top, \\ M_{DL} &= U_L \Sigma_D V_L^\top, \quad M_{LL} = U_L \Sigma_L V_L^\top, \end{aligned} \tag{101}$$

so that M_{LD} carries the LoRA spectrum on Dense singular vectors, whereas M_{DL} carries the Dense spectrum on LoRA singular vectors. Taking M_{LL} as the reference and M_{DD} as the opposite end, we report for each hybrid its residual against M_{LL} and the fraction of the M_{DD}–M_{LL} gap that it closes, averaged equally over the six strata within each seed.

Artificial Entanglement Profiling of the Hybrid Updates. Table 4 reports the two transplants. Placing the LoRA spectrum on Dense singular vectors closes 97.7% of the off-center profile gap and 98.0% of the signed-depth gap, whereas the reverse transplant closes essentially none of either: its profile closure is -0.0113 ± 0.0004 and its depth closure is 0.0249 ± 0.0009. Consequently, **the concentrated singular values of the LoRA update, rather than its singular vectors, account for the entanglement valley in this diagnostic**.

Table 4: Fraction of the M_{DD}–M_{LL} gap closed by each transplant, together with the residual measured against M_{LL} in bits. Entries are mean ± sample standard deviation over three seed-level averages.

Arm	Profile gap closed	Depth gap closed	Profile residual	Depth residual
M_{LD} (LoRA spectrum)	0.9771 ± 0.0004	0.9798 ± 0.0042	0.0675 ± 0.0021	0.0635 ± 0.0133
M_{DL} (LoRA vectors)	-0.0113 ± 0.0004	0.0249 ± 0.0009	2.9859 ± 0.0375	3.0740 ± 0.0324

L Robustness of the Entanglement Valley to Factor Ordering and Basis

The artificial entanglement profiling of Sec. 4 reshapes ΔW into a chain of factors, and the resulting profile depends on how those factors are ordered. A natural concern is therefore whether the *central dip* is a property of the learned update or an artifact of the ordering we happen to adopt. Here we reprofile the same matrices under many alternative orderings and bases, without retraining any model.

Tested Transformations. We reuse the stored step 1000 matrices of the matched-scope comparison in *Appendix J* and reprofile them in two ways. First we reorder the factors of the chain, keeping all output factors on one side and all input factors on the other, so that the central partition stays in place. We try 21 such orders: the native one, four that reverse the output factors, the input factors, both, or place the even-indexed factors first, and 16 drawn at random and then held fixed. Second we change the basis, using four signed permutations and four structured orthogonal maps. Each Dense Q/V and LoRA Q/V matrix is reprofiled under all of them, at both projections, all three layers and all three training seeds. The seed remains the statistical unit, because these transformations reanalyze the same three runs instead of adding new ones.

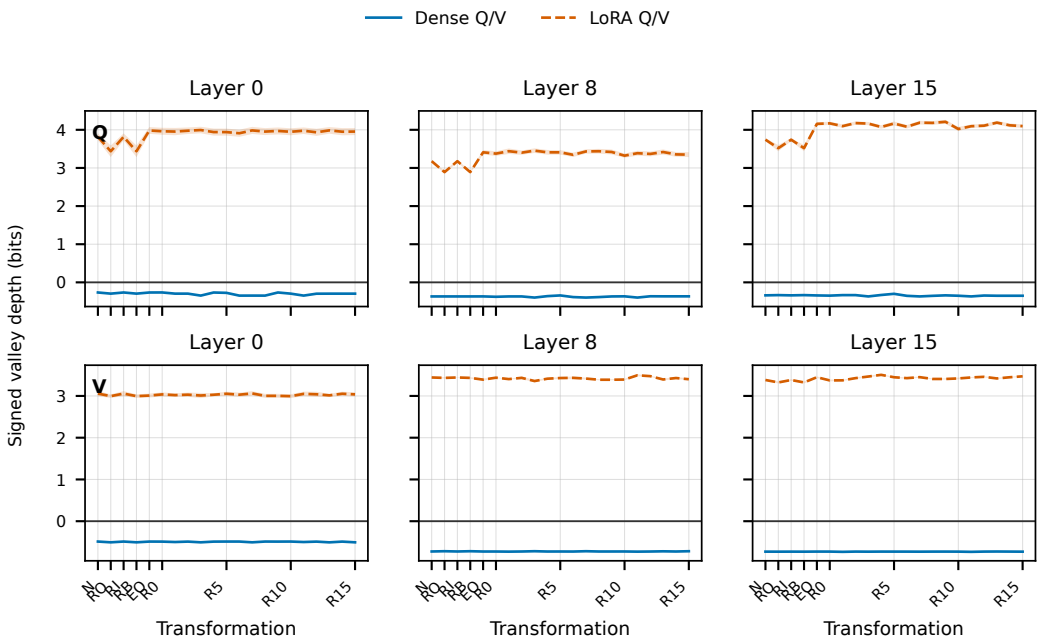

Figure 46: Signed valley depth at step 1000 under the 21 block-preserving factor orders. N is the native order, RO, RI, RB and EO reverse the output factors, the input factors, both, or place even-indexed factors first, and R0–R15 are fixed-random orders with selected labels shown. Curves and bands are mean ± sample standard deviation over the three paired training seeds at each fixed module and layer, and the connected categorical points are not a training trajectory.

Four validation controls accompany the analysis. Reprofiling under the identity transform reproduces the stored profiles exactly, and the entropy at the central partition, which is invariant under these transformations by construction, is recovered to within 7.4×10^{-4} bits under every transformation. Scaling and site-local orthogonal invariances are likewise recovered to within 5.7×10^{-5} and 2.1×10^{-4} bits.

Artificial Entanglement Profiling under Alternative Orderings. FIG. 46 shows the signed valley depth across the 21 orderings. Under every ordering, module, layer and seed, LoRA Q/V exhibits the same characteristic entanglement valley as observed in the main text, with depth between 2.852 and 4.269 bits, whereas Dense Q/V remains below zero, between -0.735 and -0.260 bits. The paired LoRA-minus-Dense difference stays positive throughout, between 3.223 and 4.610 bits. Under the eight basis transforms the behavior is the same, with LoRA Q/V between 2.953 and 4.386 bits and paired differences between 3.436 and 4.679 bits. The magnitudes nevertheless shift with the transformation, so the comparison is delimited by the orderings and bases we tested rather than holding for arbitrary coordinate changes. Within that range, these results confirm that **the entanglement valley and its contrast with Dense are robust to the choice of factor ordering and basis**.

M External Artificial Entanglement across Inputs and Attention Heads

The no-hair property reported in Sec. 5 rests on the profiling of a single attention head, evaluated in normalized form. Since the internal profiling of the projection updates differs by several bits between methods, the property deserves a measurement that spans heads, inputs and sequence lengths. Here we profile the attention matrices of matched checkpoints over all heads, on nontraining inputs, and without padding.

Measured Checkpoints and Inputs. Three trained arms enter the comparison: *Dense-add*, which freezes the base weights and trains a zero-initialized full additive matrix for each projection, *Projected-r256*, which projects that additive matrix back to rank 256 after every optimizer step, and *LoRA-r256*, which trains the factors with $r = 256$. The three arms share a rank budget of 256 in the latter two cases and are compared at checkpoints matched on calibration loss. For each arm we profile the causal attention matrices at layers $0, 8, 15$ over all 32 heads, on 100 nontraining inputs, and with every sequence kept at its actual length rather than padded. Heads and inputs are averaged within each of the three paired training seeds, which remain the statistical unit.

Artificial Entanglement Profiling of S_A under Matched Checkpoints. FIG. 47 gives the paired distributions. The LoRA-minus-Projected entropy differences at the central partition are -0.0112 ± 0.0008, $+0.0113 \pm 0.0311$ and $+0.0065 \pm 0.0272$ bits at layers 0, 8 and 15. In the two deeper layers the sign varies across seeds, so the ordering between the arms is not stable there. Consequently, **a concentrated update spectrum does not consistently correspond to lower attention entropy**: the internal profiling differences of *Appendix J* and *Appendix K* span several bits, whereas the attention entropies of these matched checkpoints agree to within hundredths of a bit. This measurement extends the no-hair property from one head to all 32 heads and from padded to unpadded inputs.

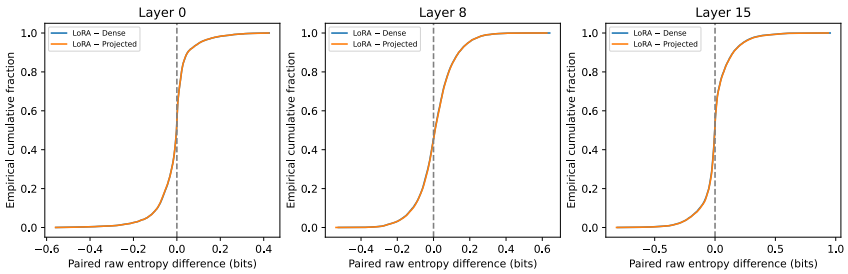

Figure 47: Paired per-head distributions of the attention entropy at the central partition, for matched Dense-add, Projected-r256 and LoRA-r256 checkpoints at layers $0, 8, 15$. Each distribution collects all 32 heads over 100 nontraining unpadded inputs, and the three paired training seeds are the statistical unit.