A Random Matrix Perspective on the Learning Dynamics of Multi-Head Latent Attention

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2025

Abstract

In this work, we study how multi-head latent attention (MLA), a popular strategy for compressing key/value memory, affects a transformer's internal capacity during pretraining. Using a lightweight suite of Marchenko–Pastur (MP) diagnostics, we analyze the spectrum of the $QK^\top$ Gram matrix throughout training, comparing three variants: the standard multi-head attention (MHA) baseline, MLA-PreRoPE with rotary applied before compression, and MLA-Decoupled, which shares a single rotary sub-vector across all heads. Our random matrix analysis reveals **three key findings**. First, capacity bottlenecks emerge locally: both MHA and MLA-PreRoPE exhibit sharp, early spikes in specific layers that persist and propagate, disrupting the balance between bulk and outlier directions. Second, these spikes coincide with rank collapse, concentrating the model's expressivity into narrow subspaces. Third, only the decoupled variant prevents this cascade, maintaining broad spectral support and suppressing outlier formation across layers. These results underscore that *how* rotary embeddings are applied is just as critical as *where* compression occurs. Sharing rotary components across heads mitigates spectral fragmentation and preserves representational capacity.

1. Introduction

Modern large language models (LLMs) increasingly face inference latency bottlenecks, not due to computational limitations, but primarily because of memory-bound key/value (KV) cache operations. To address this, recent architectures like DeepSeek-V2 and V3 have adopted Multi-Head Latent Attention (MLA) [10, 11, 14, 21], which compresses queries and keys into lower-dimensional latent representations before computing attention scores. This approach significantly reduces the KV cache size, often by over 50%, while maintaining competitive performance across various tasks.

Despite these practical advancements, the implications of latent-space compression on the internal dynamics of attention mechanisms remain underexplored. In particular, understanding how MLA influences the learning dynamics of attention layers is crucial for highlighting its inductive biases and potential limitations. While prior studies have applied Random Matrix Theory (RMT) to analyze weight matrices, co-variance matrix, and nonlinear layers in neural networks [2–7, 9, 13, 15–20], the spectral behavior of the attention Gram matrix, especially post-MLA compression, has not been thoroughly investigated.

To bridge this gap, we aim to address these key questions: **RQ1:** Where do MLA-induced spikes arise? Are they localized to specific layers or heads, or uniformly distributed? **RQ2:** Is width compression alone sufficient to suppress spectral spikes, or does rotary-vector sharing play a critical role? **RQ3:** What impact do residual spikes have on rank collapse and latent space utilization?

While prior studies have quantified the memory efficiency of MLA, they often overlook the underlying spectral dynamics that contribute to its performance. Our work addresses this shortfall through a rigorous spectral analysis.

Contributions We summarize our contributions as follows.

1. *RMT-based attention diagnostic framework.* We introduce a lightweight framework that analyze the squared singular value spectrum of $QK^\top$ matrix using four metrics derived from the Marchenko–Pastur (MP) law: MP-Gap, outlier count and energy, and MPSoft and stable rank.
2. *The first spectrum-resolved analysis of MLA training.* We benchmark classical MHA, MLA with rotary applied before up-projection (MLA-PreRoPE), and a decoupled variant with rotary shared across heads (MLA-Decoupled), integrating them in LLaMA architecture.
3. *Identification of a mid-layer spike cascade.* We discover that spectral spikes emerge early and persist in MHA and MLA-PreRoPE models, causing severe rank collapse. PreRoPE partially mitigates but does not eliminate these effects.
4. *Demonstration of rotary-sharing as a mitigation Strategy.* The MLA-Decoupled variant maintains a low MP-Gap, suppresses the growth of outliers, and preserves stable rank. This highlights that the method of sharing rotary embeddings is as critical as latent width compression in preventing spectral collapse.

By integrating memory-efficient attention mechanisms with heavy-tailed RMT-MP analysis, our work provides both a practical diagnostic tool and a design insight: shared rotary embeddings are essential in preventing spectral collapse where width compression alone is insufficient.

2. Method

To investigate how latent compression and rotary embedding strategies reshape the spectral structure of the attention kernel during training, we log four Marchenko–Pastur (MP) diagnostics—MP-Gap, outlier count, outlier energy, and soft/stable rank—on the $QK^\top$ Gram matrix at each transformer layer [1], tracked across training with high temporal resolution. All measurements are conducted under identical training conditions across three attention configurations, enabling controlled comparison of their spectral dynamics.

At each logging step, we compute the scaled Gram matrix as follows:

$$G = \frac{1}{d_{\text{model}}}(QK^\top)(QK^\top)^\top, \quad G \in \mathbb{R}^{m\times m}, \quad m = H \cdot d_k,$$

and extract its eigenvalues via a single SVD on GPU.

Marchenko–Pastur edge. For aspect ratio $\gamma = m/d_{\text{model}}$, the theoretical bulk spectrum is bounded by [12]

$$\lambda_\pm = \sigma^2(1 \pm \sqrt{\gamma})^2.$$

We apply *empirical variance correction* by estimating $\sigma^2 = \text{Var}([(QK^\top)_{ij}])$ at each step, aligning the MP bulk with observed data, essential for accurately resolving spectra in finite-width networks.

We summarize the four Marchenko–Pastur (MP) diagnostics that form the backbone of our spectral analysis in Table 1. MP-Gap (Δ) quantifies the strength of the dominant spike: a value of zero implies the spectrum is entirely confined to the MP bulk, while larger values indicate that a leading eigenvalue has detached from random noise. Outlier Count captures how many eigenvalues exceed the MP upper edge, revealing the prevalence of spiking within a layer. Outlier Energy measures the proportion of spectral mass captured by those spikes, translating the count into a

fractional energy *budget*. Finally, MPSoft- and Stable-Rank translate spike behavior into usable capacity metrics: soft-rank (ρ) quantifies the relative distance of the top spike from the bulk edge, while stable-rank (r_s) measures how much of the bulk dimension remains after excluding spikes.

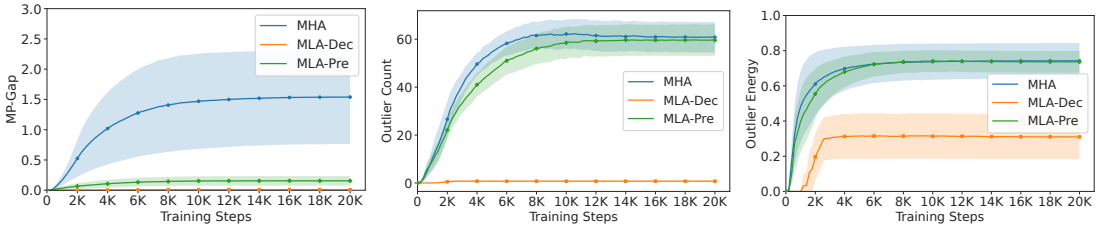
Table 1: Summary of spectral measures and their interpretations

Name	**Formula**	**Interpretation**
MP-Gap	$\Delta = \lambda_1 - \lambda_+$	Spike strength; $\Delta = 0 \Rightarrow$ no detachment
Outlier Count	$\#\{\lambda_i > \lambda_+\}$	Spike population
Outlier Energy	$\frac{\sum_{\lambda_i > \lambda_+} \lambda_i}{\sum_i \lambda_i}$	Spectral mass lost to spikes
MPSoft Rank [13]	$\rho = \frac{\lambda_1}{\lambda_+}$	Normalized spike distance
Stable Rank [13]	$r_+ = \sum \frac{\lambda_i}{\lambda_1}$	Residual capacity

Computational cost. All four diagnostics are computed from a single forward-pass SVD on the $QK^\top$ Gram matrix, imposing less than 1% runtime overhead. For instance, an SVD on a 768×768 matrix costs only 3.6M FLOPs, negligible compared to a transformer forward pass. The method scales efficiently to larger models via subsampling of heads or layers, and leaves the backward graph untouched, ensuring that training throughput is virtually unaffected.

3. Experimental Results

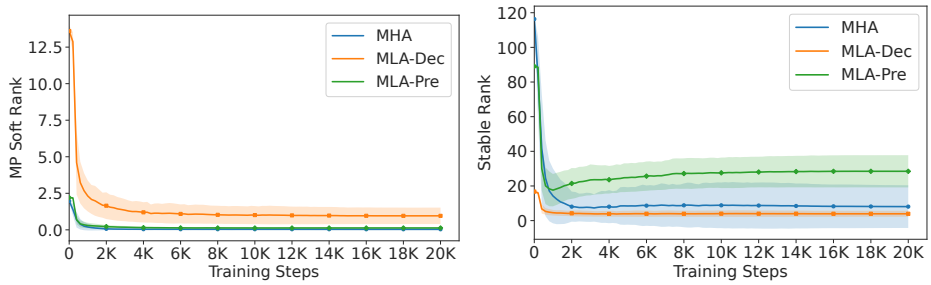
Setup. All three variants are trained from scratch for 20K steps on 2.2B tokens with a context length of 256 on C4 dataset, following the hyper-parameter settings from [8] for LLaMA-130M model. Training is performed with a global batch size 512, on 2 RTX 3090 (24 GB) GPUs. We use a compression ratio of two, reducing the latent dimension from 64 to 32 in both MLA variants.



(*a*) MP-Gap Comparison (*b*) Upper Counts Comparison (*c*) Upper Energy Comparison

Figure 1: **Spectral-spike dynamics.** We report three key spectral metrics: (a) MP-Gap, (b) outlier count, and (c) outlier energy, for MHA (blue), MLA-Dec (orange), and MLA-Pre (green). Curves represent mean values across layers; shaded bands indicate ± 1 standard deviation in LLaMA-130M model. Together, these metrics quantify the emergence and strength of spectral outliers in the $QK^\top$ Gram spectrum during training.

Decoupled MLA eliminates spectral spikes Figure 1(a) presents the MP-Gap, the distance between the largest eigenvalue (i.e., squared singular value) of the $QK^\top$ Gram matrix and the Marchenko-Pastur (MP) bulk edge. In classical MHA, this gap rapidly surges to ≈ 2 within the first 5k steps and then plateaus, indicating the emergence of a persistent, high-magnitude spectral spike. Pre-RoPE MLA follows a similar trajectory but saturates at roughly one-fifth the amplitude,



(*a*) MP-SoftRank Comparison (*b*) Stable Rank Comparison

Figure 2: **Spectral-Capacity Dynamics.** (a) MP-Soft-Rank (largest eigenvalue divided by the Marchenko–Pastur edge) and (b) Stable Rank ($\sum \lambda/\lambda_1$) are shown for LLaMA-130M model with MHA (blue), MLA-Pre (orange), and MLA-Dec (green). Curves show layer means; shaded regions indicate ± 1 standard deviation across 12 layers. Higher MP-Soft-Rank signals sharper spectral spikes; higher Stable Rank indicates better bulk direction usage. MLA-Dec converges rapidly to ideal values, while MHA and MLA-Pre exhibit persistent spikes and rank degradation.

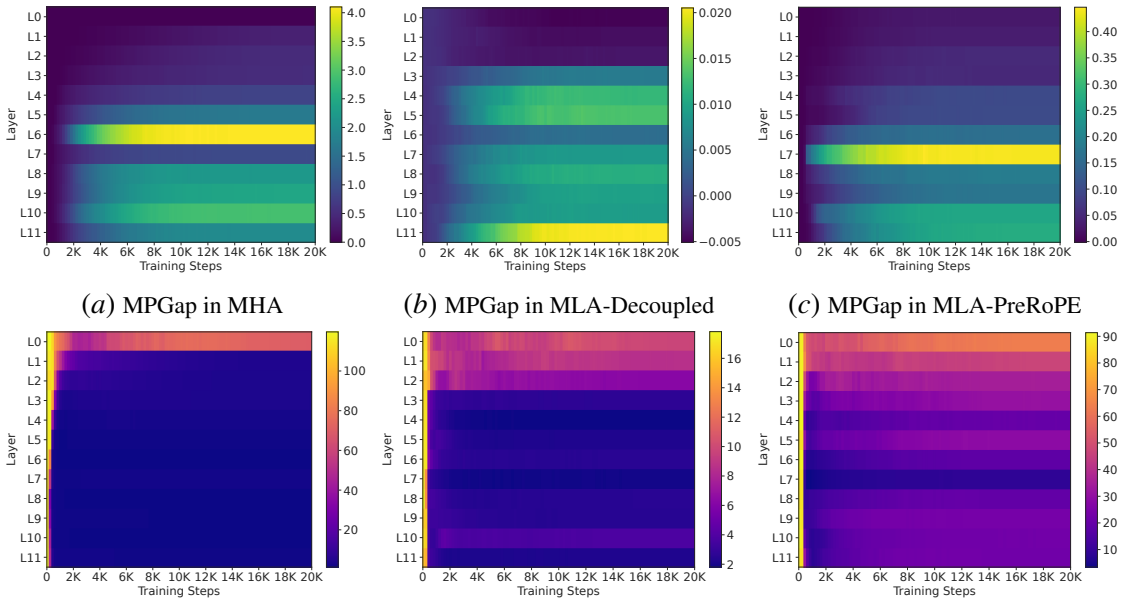
suggesting that compressing latent dimensions alone is insufficient to eliminate spike formation. In contrast, Decoupled MLA maintains an MP-Gap near zero throughout, reflecting that its shared rotary sub-vector across heads keeps all singular values inside the MP bulk.

Figure 1(b) shows the the number of outlier eigenvalues, exceeding the MP upper edge. MHA and Pre-RoPE both stabilize at 60 to 65 outliers per layer (roughly 5 to 6 per head), while Decoupled MLA consistently exhibits zero, empirically confirming the absence of spectral outliers and validating the collapsed MP-Gap. Figure Figure 1(c) quantifies outlier energy, the proportion of total spectral energy carried by these spikes. MHA and Pre-RoPE MLA channel nearly 70% of the spectrum into the spike subspace, signaling severe rank collapse. In contrast, Decoupled MLA re-distributes this energy back into the bulk, dropping outlier energy below 30%.

This shift reflects a broader effective rank and reinforces a key insight: *how rotary embeddings are applied matters*. Head-shared rotary embeddings suppress spike formation entirely, whereas conventional key/query compression schemes do not.

Decoupled MLA preserves substantially higher usable dimensionality Figure 2 complements the spike metrics from Figure 1 by examining how spectral spikes influence usable dimensionality. Initially, the MP-Soft-Rank ratio (left) is high across all models due to random initialization placing the largest eigenvalue significantly above the bulk distribution. Within the first thousand training steps, MHA stabilizes around a ratio of ≈ 2, Pre-RoPE MLA settles around ≈ 1.5, whereas decoupled MLA rapidly converges closer to ≈ 1.1. This indicates that decoupling effectively anchors the largest eigenvalue near the Marchenko-Pastur bulk edge, whereas the other two methods consistently maintain substantial separation.

The Stable-Rank plot (right) provides complementary insight. Following an initial transient phase, decoupled MLA maintains an effective rank around 30, approximately three times higher than both MHA and Pre-RoPE MLA, which plateau below 10. Taken together with Figure 1, these observations clearly demonstrate that employing a shared rotary sub-vector across attention heads not only mitigates spectral spikes but also preserves the global representational capacity of the model. In contrast, classical MHA and basic key/query compression both experience persistent rank collapse. See Appendix A for a detailed discussion.



(*a*) MPGap in MHA (*b*) MPGap in MLA-Decoupled (*c*) MPGap in MLA-PreRoPE

(*d*) StableRank in MHA (*e*) StableRank in MLA-Decoupled (*f*) StableRank in MLA-PreRoPE

Figure 3: **Layerwise spectral dynamics**: (Top row)MP-Gap and (bottom row)StableRank heatmaps. MHA exhibits strong mid-layer concentration in MPGap and declining StableRank in later layers, while MLA-based methods spread representational changes more evenly, maintaining higher stable ranks across depths

Capacity bottlenecks: Mid-layer spikes vs. uniform utilization The MP Gap and Stable Rank heatmaps (Figure 3) reveal a clear capacity bottleneck in the MHA model, where a persistent band of spectral spikes emerges in mid layers 6 to 8 and gradually spreads deeper. This spike pattern reflects a strong separation between bulk and edge singular values and coincides with a collapse in stable rank beyond layer 5. Pre RoPE MLA shows a much weaker version of this effect, with about 20 percent spike intensity and partial recovery in deeper layers, but still suffers capacity loss. In contrast, the Decoupled MLA model avoids spike formation entirely and sustains more than 60 percent normalized rank across all layers and training steps. The absence of edge singularities in Decoupled MLA suggests that redistributing rotary vectors across heads enforces uniform spectral behavior and preserves stable representational capacity throughout the network.

4. Conclusion and Limitations

Our RMT analysis shows that sharing rotary embeddings across heads eliminates spectral spikes, maintains MP Gap at the noise level with outliers close to one, and preserves over 60 percent stable rank in MLA decoupled mode. In contrast, classical MHA and MLA Pre RoPE remain spike dominated and lose around 70 percent of spectral energy to a few dominant directions. These findings are based on a 12 layer LLaMA-130M trained for 20K steps on 2.2 billion C4 tokens.

References

[1] Han Bao, Ryuichiro Hataya, and Ryo Karakida. Self-attention networks localize when QK-eigenspectrum concentrates. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[2] Florent Bouchard, Ammar Mian, Malik Tiomoko, Guillaume Ginolhac, and Frederic Pascal. Random matrix theory improved fréchet mean of symmetric positive definite matrices. In *Forty-first International Conference on Machine Learning*, 2024.

[3] Romain Couillet, Gilles Wainrib, Hafiz Tiomoko Ali, and Harry Sevi. A random matrix approach to echo-state neural networks. In *International Conference on Machine Learning*, 2016.

[4] Vasilii Feofanov, Malik Tiomoko, and Aladin Virmaux. Random matrix analysis to balance between supervised and unsupervised learning under the low density separation assumption. In *International Conference on Machine Learning*, 2023.

[5] Aymane El Firdoussi, Mohamed El Amine Seddik, Soufiane Hayou, Reda ALAMI, Ahmed Alzubaidi, and Hakim Hacid. Maximizing the potential of synthetic data: Insights from random matrix theory. In *The Thirteenth International Conference on Learning Representations*, 2025.

[6] Romain Ilbert, Malik Tiomoko, Cosme Louart, Ambroise Odonnat, Vasilii Feofanov, Themis Palpanas, and Ievgen Redko. Analysing multi-task regression via random matrix theory with application to time series forecasting. *Advances in Neural Information Processing Systems*, 2024.

[7] Noam Levi and Yaron Oz. The underlying scaling laws and universal statistical structure of complex datasets. *arXiv preprint arXiv:2306.14975*, 2023.

[8] Pengxiang Li, Lu Yin, and Shiwei Liu. Mix-LN: Unleashing the power of deeper layers by combining pre-LN and post-LN. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[9] Zhenyu Liao and Romain Couillet. The dynamics of learning: A random matrix approach. In *International Conference on Machine Learning*, 2018.

[10] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

[11] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[12] VA Marchenko and Leonid A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.(NS)*, 72(114):4, 1967.

[13] Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 2021.

[14] Fanxu Meng, Zengwei Yao, and Muhan Zhang. Transmla: Multi-head latent attention is all you need. *arXiv preprint arXiv:2502.07864*, 2025.

[15] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International conference on machine learning*, 2017.

[16] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. *Advances in neural information processing systems*, 30, 2017.

[17] Max Staats, Matthias Thamm, and Bernd Rosenow. Locating information in large language models via random matrix theory. *arXiv preprint arXiv:2410.17770*, 2024.

[18] Matthias Thamm, Max Staats, and Bernd Rosenow. Random matrix theory analysis of neural network weight matrices. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024.

[19] Malik Tiomoko, Romain Couillet, Florent Bouchard, and Guillaume Ginolhac. Random matrix improved covariance estimation for a large class of metrics. In *International Conference on Machine Learning*, 2019.

[20] Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *International conference on machine learning*, 2022.

[21] Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Huazuo Gao, Jiashi Li, Liyue Zhang, Panpan Huang, Shangyan Zhou, Shirong Ma, et al. Insights into deepseek-v3: Scaling challenges and reflections on hardware for ai architectures. *arXiv preprint arXiv:2505.09343*, 2025.

Appendix A. Outlier Energy Distribution Shows Spectral Compression vs. Spread

The violin plot (Figure 4) depicts the distribution of outlier energy across all 12 layers at the final training steps. In both the classical MHA and Pre-RoPE MLA variants, the distribution peaks around 0.70 with wide tails, indicating that approximately 70% of the spectral energy remains concentrated in a few dominant directions. This suggests persistent rank compression even at convergence.

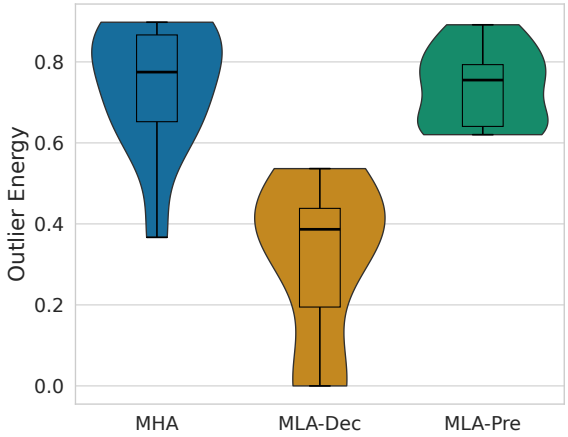


Figure 4: Upper Energy Violin Plot

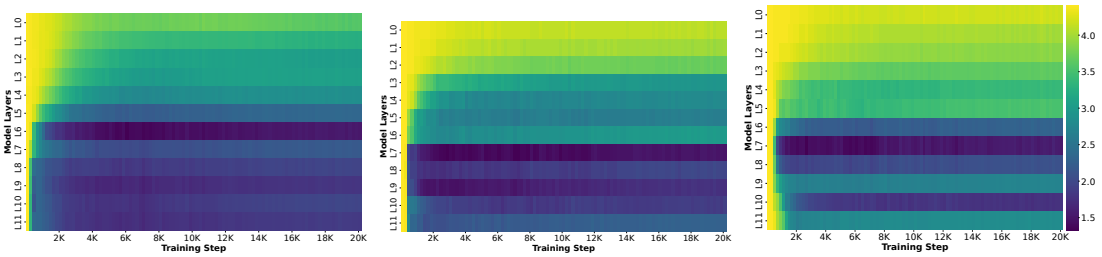
In contrast, the Decoupled MLA variant exhibits tightly clustered violin plots centered near 0.30, reflecting a substantial shift of spectral mass away from outliers and toward the bulk. Moreover, this pattern is consistent across all layers, suggesting a structurally broader use of representational capacity. These distributions align with the averaged outlier trends reported in Figure 1 (where MHA and Pre-RoPE plateau near 0.7 and MLA-Decoupled stabilizes near 0.3), and reinforce the heatmap observations from Figure 3, which showed suppressed MP-Gap and sustained stable rank in the decoupled design.

In summary, the violin plots confirm that *only the decoupled architecture effectively returns spike energy to the bulk*, thereby preserving a broader and more effective rank across layers.

Appendix B. Attention Entropy Distribution in MHA and MLA

MLA improves information flow and stability across layers Our entropy analysis in Figure 5 highlights significant differences in information flow across three attention mechanisms. Vanilla MHA displays a clear bifurcation: early layers (L0-L3) quickly reach an entropic-overload state ($>$4 bits), while a deep entropy drop around layer 5 plunges below 1.5 bits and fails to fully recover. This rigid stratification suggests rich information flow at the network's start but significant starvation in the middle and deeper layers.

MLA-Decoupled softens these extremes, moderating both overload and starvation. MLA-PreRoPE further improves the entropy distribution: the middle-layer entropy dip nearly disappears, deeper layers recover rapidly within the first 5,000 steps, and the overall stack stabilizes twice as quickly as MHA. Thus, combining latent compression with pre-RoPE positional embeddings yields a more uniform and rapidly converging information flow, highlighting how nuanced architectural adjustments can significantly enhance transformer performance.



(*a*) Attention entropy MHA (*b*) MLA-Decoupled (*c*) MLA-PreRoPE

Figure 5: Attention entropy patterns in classical MHA and MLA variants (decoupled and Pre-RoPE)