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

Nandan Kumar Jha NJ2049@NYU.EDU
New York University

Brandon Reagen BJR5@NYU.EDU
New York University

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 $W_Q W_K^\top$ 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: i)** 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; **ii)** these spikes coincide with rank collapse, concentrating the model's expressivity into narrow subspaces; **iii)** 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) continue to grow in scale and capability; however, their practical utility is constraint by increasing inference latency stems from memory-bound key/value (KV) cache operations, rather than compute. To address this, recent architectures such as DeepSeek-V2/V3 have introduced Multi-head Latent Attention (MLA) [10, 11, 14, 21], which compresses queries and keys into lower-dimensional latent representations before attention computation. This design reduces KV cache size by over 50% while maintaining strong performance.

Despite these practical gains, the fundamental question remains unanswered: *How does latent compression impacts the spectral dynamics of attention, and what are its implications for learning and generalization?* While Random Matrix Theory (RMT) has emerged as a powerful tool to study neural network internal dynamics [2–7, 9, 13, 15–20], its application to the spectral behavior of attention mechanisms under latent-space compression remains largely unexplored.

This gap limits our understanding of MLA's inductive biases and potential pitfalls. For instance, do spectral pathologies such as rank collapse persist under MLA? Can width compression alone prevent outlier growth, or do we need additional design choices, such as applying rotary embeddings before compression? While prior studies have quantified the memory efficiency of MLA, they often overlook the underlying spectral dynamics that contribute to its efficiency.

© N.K. Jha & B. Reagen.

In this work, we aim to bridge this gap by investigating following research questions:

RQ1: Where do MLA-induced spectral spikes emerge? Are they layer- or head-specific?

RQ2: Is latent compression alone sufficient to suppress outliers, or rotary-vector sharing matters?

RQ3: What impact do residual spikes have on rank collapse and latent space utilization?

We summarize our **contributions** as follows.

1. *RMT-based diagnostic framework for attention.* We develop a lightweight tool to analyze the squared singular values spectrum of $W_Q W_K^\top$ Gram matrix, using four Marchenko-Pastur (MP) metrics [12]: MP-Gap, outlier count and energy, MPSoft rank, and stable rank.
2. *First spectral analysis of MLA during training.* We apply our framework to benchmark classical MHA and two MLA variants: MLA-PreRoPE, where rotary embeddings are applied before up-projection, and MLA-Decoupled, which uses a shared rotary vector across heads. For consistency and fair evaluation, all experiments are conducted within the 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. *Rotary-vector sharing as a key mitigation strategy.* The MLA-Decoupled variant suppresses spectral outliers: maintains a low MP-Gap, and preserves stable rank across layers. This highlights that rotary-vector sharing is crucial for maintaining a good spectral behavior.

2. Experimental Setup

To investigate how latent compression and rotary embedding strategies influence the spectral dynamics of the attention mechanism, we perform RMT analytics on $W_Q W_K^\top$ Gram matrix [1] at each attention layer, using four MP metrics: MP-gap, outlier count and energy, and soft and stable rank.

We integrate all three variants into a LLaMA-130M architecture and train them from scratch for 20K steps on 2.2B tokens from the C4 dataset, using a context length of 256. We adopt the down-scaled architectural settings and training setup for LLaMA-130M from [8]. Training is performed with a global batch size of 512 on two RTX 3090 GPUs (24 GB each). In both MLA variants, we apply a compression ratio of 2, reducing the latent dimension from 64 to 32.

3. Spectral Analysis of Latent Compression in MLA via Random Matrix Theory

Notations W_Q and W_K denote the query and key weight matrices, respectively; d_{model} is the model embedding dimension; H is the number of attention heads; and d_k is the per-head dimension.

Cross-Gram construction For each attention layer we consider the learned query and key projection weights $W_Q, W_K \in \mathbb{R}^{m \times d_{\text{in}}}$, where $m = H \cdot d_k$ is the total head dimension and d_{in} is the input dimension of the projection matrices. At every logging step we form the *cross-Gram* matrix

$$G = \frac{1}{d_{\text{in}}} W_Q W_K^\top \in \mathbb{R}^{m \times m}, \tag{1}$$

and compute its eigenvalues via singular-value decomposition (SVD).

Rationale For MHA, the input dimension to the query/key projection is $d_{\text{in}} = d_{\text{model}}$ (e.g., 768 in LLaMA-130M). In MLA variants, this projection is factorized into a shared down-projection ($W^{\downarrow}$) and a head-specific up-projection ($W_Q^{\uparrow}, W_K^{\uparrow}$). We analyze *only the up-projection*, setting $d_{\text{in}} = 32$, to *isolate* the effects of latent compression and rotary embedding design on the attention spectrum,

while keeping the shared $W^{\downarrow}$. In the *decoupled* setting, we further isolate the RoPE branch, with $d_{\text{in}} = 32$ and row dimension $m = \frac{1}{2}Hd_k$, to focus purely on the query/key structure without confounding from the value pathway.

Marchenko–Pastur (MP) metrics Dividing by d_{in} sets the expected entry variance of G to one under the i.i.d. null model. With aspect ratio $\gamma = m/d_{\text{in}}$, the MP bulk edges are therefore

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

Table 1 summarize the MP metrics used for spectral analysis. `MP-Gap` quantifies the strength of the dominant spike: a value of zero indicates that the bulk spectrum lies entirely within the MP bulk, while larger values reflect a detached leading eigenvalue. `Outlier Count` measures how many eigenvalues exceed the MP upper edge, capturing the prevalence of spiking. `Outlier Energy` quantifies the spectral mass of these spikes, translating it into a fractional energy *budget*. Finally, `MPSoft-` and `Stable-Rank` turn spike behavior into capacity metrics: soft-rank measures the relative distance of the top spike from the bulk edge, while stable-rank captures how much of the bulk dimension remains after excluding spikes.

Metric	Formula	Interpretation
MP-Gap	$\Delta = \lambda_1 - \lambda_+$	Spike strength
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

Table 1: Summary of spectral measures and their interpretations (λ_1 is largest eigen value)

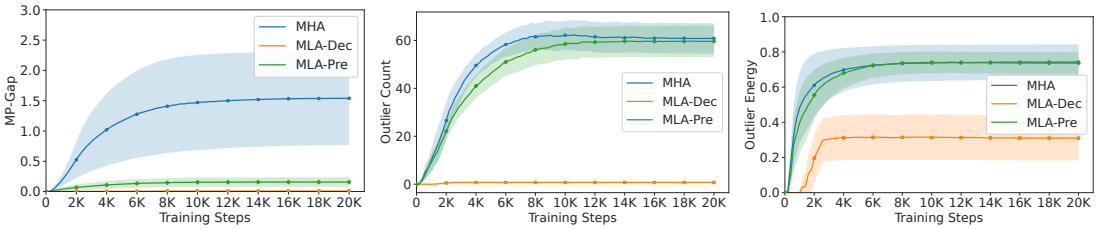
4. Experimental Results

Decoupled MLA substantially reduces spectral spikes As shown in 1(a), MP-Gap in classical MHA rises to ≈ 2 within the first 5K steps and then plateaus, indicating a persistent, high-magnitude spectral spike. Pre-RoPE MLA exhibits a similar trend but saturates at a lower amplitude, suggesting that latent compression alone does not suppress spike formation. In contrast, Decoupled MLA keeps the MP-Gap near-zero throughout training, indicating that its shared rotary sub-vector prevents singular values from escaping 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 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 the spike formation substantially, whereas conventional key/query compression schemes do not.

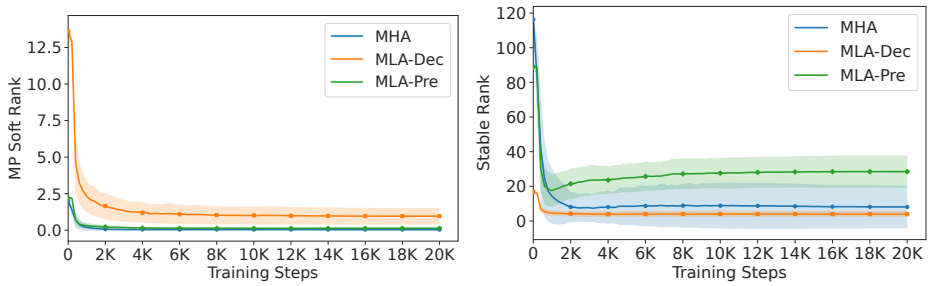
Compression-regularization trade-off: Decoupled MLA suppresses outliers while Pre-RoPE MLA maximizes capacity Figure 2 illustrates two sides of the spectral trade-off. First, MP-Soft-Rank which measures how far the largest eigenvalue lies above the MP bulk. After just 1K steps, the Decoupled MLA drives this ratio to ≈ 1.0, substantially reducing spectral spikes (outliers). In contrast, classical MHA and Pre-RoPE MLA stabilize at ≈ 1.2 and ≈ 1.5, respectively, indicating persistent spike formation.



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

Figure 1: **Spectral-spike dynamics:** (a) MP-Gap, (b) outlier count, and (c) outlier energy, for MHA (blue), MLA-Dec (orange), and MLA-Pre (green). Curves show layer-wise means and shaded bands denote ± 1 standard deviation in LLaMA-130M. Together, these metrics capture the emergence and strength of spectral outliers in the $W_Q W_K^\top$ spectrum.

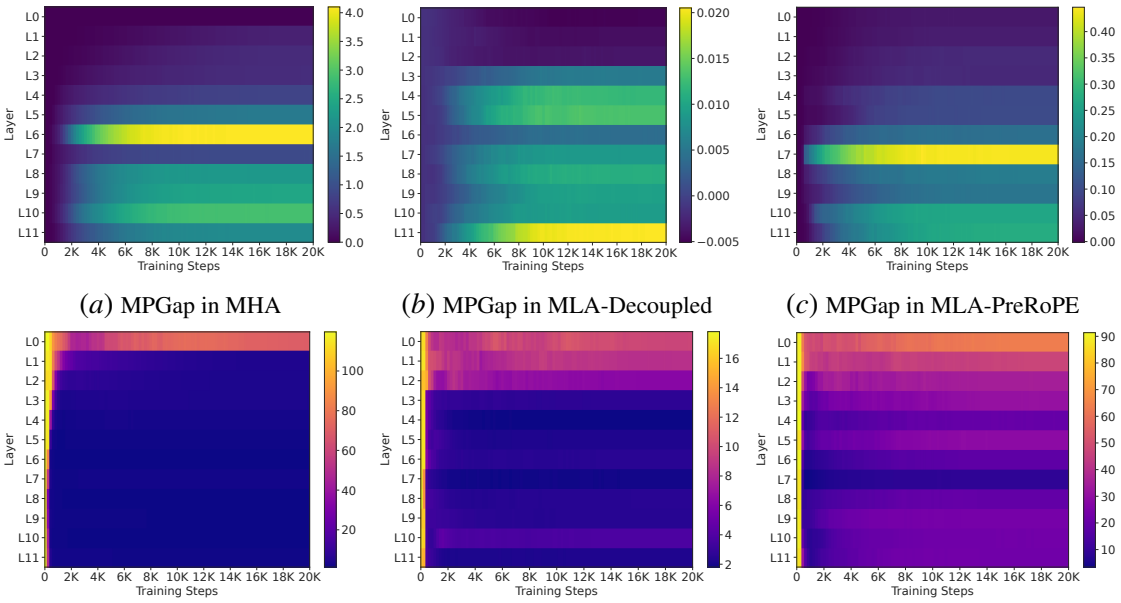
Nonetheless, on the the flip side, the Stable-Rank, a proxy for usable dimensionality shows a reverse order: MLA-Pre retains the highest capacity (~ 45), MHA saturates at off around ~ 12, while MLA-Dec collapses to ~ 5. This trade-off is expected. The decoupled variant reduces the row dimension (by isolating the RoPE branch), shares a single rotary sub-vector across heads, and applies a strong compression bottleneck ($d_{\text{in}} = 32$). These factors suppress the top singular value—effectively taming spikes—but also reduce the Frobenius norm more rapidly than the spectral norm, leading to lower stable rank. By contrast, MLA-Pre retains the full row dimension and latent energy, preserving many directions even while tolerating a moderate spike.



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

Figure 2: **Spectral-Capacity Dynamics:** (a) MP-Soft-Rank, and (b) Stable Rank 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 excels at suppressing outliers, while MLA-Pre offers the highest representational capacity. MHA remains in between on both metrics.

Capacity bottlenecks: Mid-layer spikes vs. uniform utilization The MP-Gap heatmaps (top row of Figure 3) show a sharp *hot band* in MHA: the mid layer (L6) reaches to a gap of ≈ 4 within the first 5k steps and the spike then diffuses into deeper layers. Pre-RoPE MLA shows the same localization but the peak magnitude is roughly one-tenth that of MHA, suggesting that latent compression (by a factor of 2) dampens the spikes but does not remove them completely. By contrast, Decoupled MLA remains essentially flat (< 0.02 throughout), demonstrating that spreading each rotary sub-vector across heads suppresses edge singular values and prevents spike formation.



(*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

The Stable-Rank heat-maps (bottom row) complete the picture. MHA starts with a high rank (~ 120) in the first few layers but collapses below $\sim$20 after layer $\sim$5, mirroring the MP-Gap spike. Pre-RoPE MLA partially recovers in deeper layers ($\sim$20–30%) and retains $\sim$40% rank in earlier layers, though it still underperforms. In contrast, Decoupled MLA consistently sustains $>$60% normalized rank across all layers and training steps, indicating stable representational capacity.

Outlier energy distribution shows spectral compression vs. spread

The violin plot in Figure 4 summarizes the outlier energy across all 12 layers at convergence. For both the baseline MHA and Pre-RoPE MLA, the distribution is centered around ≈ 0.75, with a long tail extending to ~ 0.60. This pattern suggests that approximately 75% of the spectral energy remains concentrated in a few dominant directions—evidence of persistent rank compression. In contrast, the Decoupled MLA exhibits a significantly different trend: its distribution is both shifted downward and narrowed, with a median around ≈ 0.40 and most of the mass concentrated between 0.20 and 0.55. This shift indicates that a substantial portion of the outlier energy has been redistributed into the MP bulk.

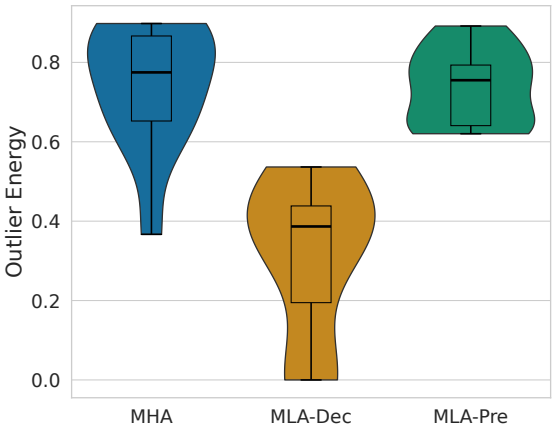


Figure 4: Upper Energy Violin Plot

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.

Perplexity Comparison Table 2 summarizes the final perplexity values across MHA and MLA variants with different RoPE configurations. The Decoupled MLA achieves the best perplexity (26.86), closely matching the baseline MHA despite its architectural modifications. The Pre-RoPE variant performs slightly worse, while the NoPE variant, lacking any positional encoding, exhibits a substantial degradation in performance, underscoring the importance of positional structure in MLA-based models.

Recall that Decoupled MLA (Figure 3) eliminates the mid-layer MP-Gap spike and sustains more than 60% Stable Rank across depth. Thus, both the spectral diagnostics and perplexity converge on a consistent message: suppressing spike-induced rank collapse preserves (and may even slightly improve) predictive performance without increasing the parameter budget, while also significantly reducing the KV cache size.

Table 2: Perplexity comparison across MHA and MLA variants with different RoPE configurations (LLaMA-130M). MLA without RoPE (NoPE) shows substantially degraded performance

	MHA	MLA-Pre	MLA-Dec	MLA-NoPE
Eval PPL (↓)	26.89	27.72	26.86	31.54

5. Conclusion

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.

Broader impact Our work aims to bridges architectural efficiency with spectral interpretability. By combining memory-efficient attention mechanisms with RMT-based diagnostics, we uncover critical design insights for building future LLMs that are not only faster but also spectrally robust.

Limitations These findings are based on a 12 layer LLaMA-130M trained for 20K steps on 2.2B training tokens from C4 corpus. Heavier models, longer training schedules, or additional spectral metrics such as tail alpha may reveal new behaviors. Extending the logger to billion scale models and correlating spectral properties with downstream quality remain open directions for future work.

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.

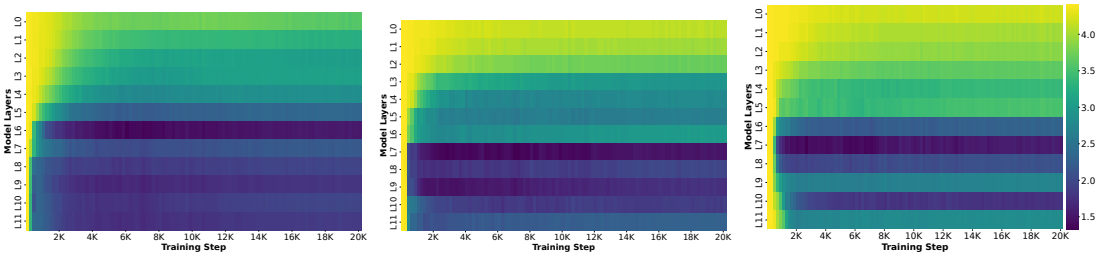
[19] Malik Tiomoko, Romain Couillet, Florent Bouchard, and Guillaume Ginolhac. Random matrix improved covariance estimation for a large class of metrics. In *International Conference on Machine Learning*, 2019.

[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. 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.



(*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)

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.

Appendix B. Computational Cost

All four diagnostics are computed from a single forward-pass SVD on the $W_Q W_K^\top$ weight 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.