

Spectral Outliers Reveal Dominant Learned Structure in Transformer Attention

Kasun Dewage*, Marianna Pensky*, Suranadi De Silva*, and T. H. Bandara†

*University of Central Florida

Orlando, Florida, USA

{KasunTharuka.Dewage, Marianna.Pensky, su966204}@ucf.edu

†Virtusa

thbandara@virtusa.com

Abstract—We apply Marchenko–Pastur (MP) random matrix theory to pre-trained attention weights in order to separate each projection matrix into a random-like bulk and a set of spectral outliers. We validate this decomposition causally: zeroing the MP-identified outliers (signal) in Mistral-7B drives HellaSwag, MMLU, and PIQA close to random-chance performance, whereas zeroing a count-matched subset of bulk singular values causes smaller but non-negligible degradation. Across 11 pre-trained transformers we identify five recurring patterns: spectral outliers encode a dominant component of the learned structure; Q projections carry the most outliers; V projections under grouped-query attention lack a clean signal/noise separation; entry-level outliers form structured row-bands in Q and column-bands in O; and specific residual-stream dimensions persist as band outliers across layers in K and O. We close by outlining how these observations could inform parameter-efficient fine-tuning and structured pruning.

Index Terms—random matrix theory, transformer interpretability, attention mechanism, Marchenko–Pastur distribution, model pruning, LoRA, large language models

I. INTRODUCTION

Modern pretrained transformers [1] contain billions of parameters, yet not all of them carry the same informational content. Mechanistic interpretability seeks to identify which components encode learned behaviors [2], [3], while parameter-efficient methods such as LoRA [4] exploit low-rank structure in weight updates but offer limited guidance on *which* projections or layers should receive the largest adaptation budget.

Random matrix theory (RMT) provides a principled framework for this problem. The Marchenko–Pastur law [5] characterizes the eigenvalue distribution of large random covariance matrices: values exceeding the MP upper edge are unlikely under a purely random model and can therefore indicate structured signal. Prior work [6], [7] focuses mainly on aggregate spectral statistics rather than on the spatial distribution of structure, and typically does not causally validate whether the identified structure is functionally important. Martin and Mahoney [6] showed that state-of-the-art DNNs often exhibit heavy-tailed spectral distributions rather than the clean MP

bulk assumed by the spiked model; our use of the MP framework should therefore be read as an approximation that nonetheless yields empirically useful signal/noise decompositions (Section VII).

We go beyond scalar summaries and introduce: (1) **entry-level outlier heatmaps** revealing row–column positions of anomalously large weights across Q, K, V, O; (2) **cross-layer residual-stream alignment analysis** identifying residual-stream dimensions persistently acting as band outliers; and (3) **targeted ablation experiments** providing causal evidence that the detected structures are functionally important. We analyze 11 pretrained models spanning encoder-only (BERT [8], RoBERTa [9]) and decoder-only (OPT [10], LLaMA 1/2/3 [11]–[13], Mistral [14], Qwen 2.5 [15], Phi-3 [16]) designs.

Our contributions are:

- 1) MP analysis across six architecture families, revealing recurring spectral patterns in attention weights.
- 2) Entry-level outlier heatmaps showing a consistent spatial organization: row-bands in Q and column-bands in O.
- 3) Evidence that V projections under GQA [17] contain dramatically fewer MP spectral outliers, suggesting a different signal/noise regime for value computation.
- 4) Cross-layer alignment analysis revealing persistent residual-stream channels in K and O.
- 5) Causal ablation experiments showing that spectral outliers encode a dominant functionally important component of attention weights, with per-component criticality following $K > Q \gg V$ in the tested setting.

II. METHOD

A. Marchenko Pastur Spectral Analysis

Given a weight matrix $W \in \mathbb{R}^{m \times n}$, we compute its singular values $\{s_i\}$ and analyze the nonzero spectrum through squared singular values. The MP threshold is used as a primary separation between a random-like bulk and unusually large spectral components. Let $\gamma = \max(m, n) / \min(m, n)$. We estimate the noise variance as $\hat{\sigma}^2 = \text{median}(s_i^2) / (1 + \gamma)$ and define the MP upper edge as:

Code is available at <https://github.com/Kasun-Dewage/spectral-outliers-attention>.

© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists (b) reuse of any copyrighted component of this work in other works.

Spectral Outliers in Transformer Attention via Random Matrix Theory

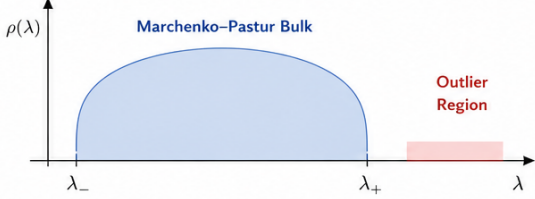
1. Random Matrix Theory: Marchenko–Pastur (MP) Law

For a large random matrix $X \in \mathbb{R}^{m \times n}$ with i.i.d. entries ($\mathbb{E}[X_{ij}] = 0$, $\mathbb{E}[X_{ij}^2] = \sigma^2$), the empirical spectral distribution of $\frac{1}{n} X X^T$ converges to the MP distribution when $m, n \rightarrow \infty$ with $\gamma = \frac{m}{n} \in (0, \infty)$.

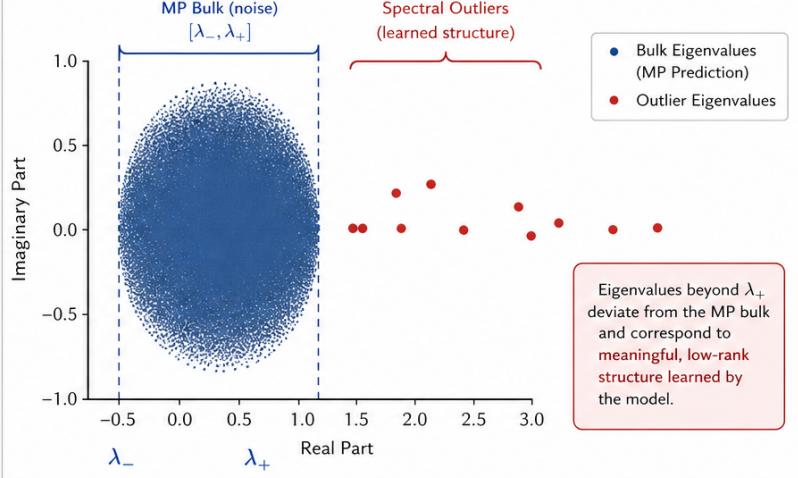
Marchenko–Pastur Density

$$\rho_{\text{MP}}(\lambda) = \frac{1}{2\pi\gamma\sigma^2} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \mathbf{1}_{[\lambda_-, \lambda_+]}(\lambda)$$

with support $[\lambda_-, \lambda_+]$, where $\lambda_{\pm} = \sigma^2(1 \pm \sqrt{\gamma})^2$.



2. Empirical Spectrum of Transformer Attention Weight Matrices



3. Interpreting Outliers: From Spectrum to Structure

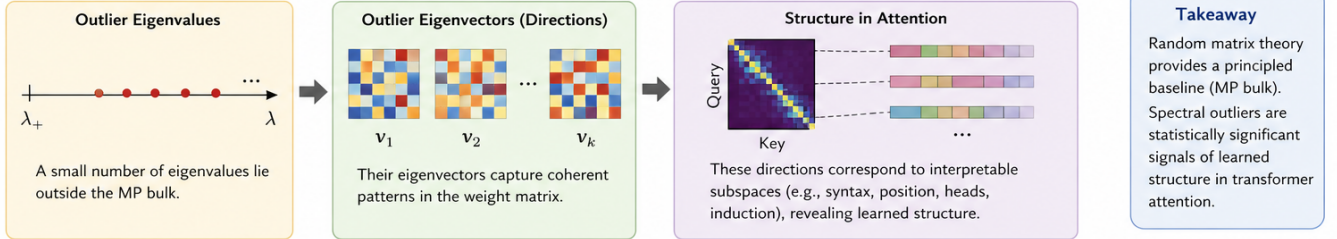


Fig. 1. Spectral density of transformer weight matrices illustrating the *Marchenko–Pastur* bulk and outlier eigenvalues (signal) corresponding to learned structure.

Singular values with $s_i^2 > \lambda_+$ are classified as **spectral outliers (signal)**; Fig. 1 illustrates this decomposition. We report the outlier count and the spectral energy ratio:

$$R = \frac{\sum_{i: s_i^2 > \lambda_+} s_i^2}{\sum_i s_i^2}. \quad (2)$$

This threshold is only approximate, because trained transformer weights are not random Gaussian matrices and may exhibit heavy-tailed spectra. We therefore treat the MP split as a diagnostic signal/noise proxy rather than as proof that all below-threshold components are noise.

B. Entry-Level Outlier Detection

For each weight matrix, we compute the standard deviation σ_W over all entries and flag positions (i, j) where $|W_{i,j}| > 4\sigma_W$. These are visualized as heatmaps, revealing spatial patterns that are invisible to spectral methods alone.

C. Cross-Layer Residual-Stream Alignment

For each layer ℓ and projection $P \in \{Q, K, V, O\}$ we compute Euclidean norms along the residual-stream axis of the weight matrix: column norms for Q, K and V, and row norms for O. Columns or rows whose norm exceeds $\mu + 4\sigma$ (where

μ and σ are the mean and standard deviation of the norms across all columns or rows) are flagged as **band outliers**. For Q, K, V matrices of shape $(n_h \cdot d_h) \times d$, outlier columns identify input residual-stream dimensions producing anomalously large projections; for O matrices of shape $d \times (n_h \cdot d_h)$, outlier rows identify output residual-stream dimensions receiving concentrated attention contributions.

A dimension is **persistent** if it appears as a band outlier in ≥ 3 layers. We further compute the **cross-component hit count** as the number of Q, K, V, and O components in which the same dimension appears as a band outlier in at least one layer. Dimensions included in the cross-component table are persistent in at least one component.

D. Ablation Methodology

Given $W = U\Sigma V^T$, we apply the MP threshold to classify each singular value as outlier (signal) or bulk. Two strategies:

- **spectral_outliers**: Zero singular values *above* λ_+ (remove signal, keep below-threshold components).
- **spectral_random_bulk**: Zero a randomly selected subset of singular values *below* λ_+ , with subset size $\lceil f \cdot N_{\text{outlier}} \rceil$, where N_{outlier} is the number of MP-identified outlier singular values in the same matrix. This

TABLE I
MODELS ANALYZED. GQA = GROUPED-QUERY ATTENTION [17], MHA = MULTI-HEAD ATTENTION [1].

Model	d	L	n_h	n_{kv}	Attn.
BERT-base [8]	768	12	12	12	MHA
RoBERTa-base [9]	768	12	12	12	MHA
OPT-125M [10]	768	12	12	12	MHA
LLaMA-1-7B [11]	4096	32	32	32	MHA
LLaMA-2-7B [12]	4096	32	32	32	MHA
LLaMA-3-8B [13]	4096	32	32	8	GQA
Mistral-7B [14]	4096	32	32	8	GQA
Qwen2.5-0.5B [15]	896	24	14	2	GQA
Qwen2.5-1.5B [15]	1536	28	12	2	GQA
Qwen2.5-7B [15]	3584	28	28	4	GQA
Phi-3-mini [16]	3072	32	32	32	MHA

provides a count-matched diagnostic comparison against outlier removal while keeping the MP outliers intact.

For per-component analysis, the `entry_outliers` strategy zeros all entries satisfying $|w_{ij}| > 4\sigma_W$. Ablations are applied before evaluation on HellaSwag [18], MMLU [19], and PIQA [20] under zero-shot conditions using `lm-evaluation-harness` [21]. The `spectral_random_bulk` results are mainly diagnostic comparisons against outlier removal rather than a full statistical characterization of all possible random bulk subsets.

E. Models Analyzed

Table I summarizes the 11 pretrained models, spanning six architecture families, used throughout this study.

III. OBSERVATIONAL RESULTS

A. Spectral Outlier Counts and Energy Ratios

Table II reports average spectral outliers (signal) per layer and energy ratios. Three broad patterns emerge across the analyzed models:

Pattern 1: Q dominates spectrally. Across most architectures, Q projections contain the most spectral outliers. In LLaMA-2-7B, Q averages 1518.7 outliers (87.8% energy) against 1392.3 (79.0%) for V.

Pattern 2: V is least structured under GQA. V projections consistently contain fewer outliers in GQA models. In LLaMA-3-8B, V averages only 219.1 outliers (43.1% energy)—a $6\times$ reduction relative to Q.

Pattern 3: Layer-wise decay. Energy ratios decrease from early to late layers, suggesting that early layers capture more structured, globally shared representations.

B. The GQA Effect on Value Projections

The spectral summary in Table II shows a dramatic reduction in V-projection structure under GQA: transitioning from MHA (LLaMA-1/2) to GQA (LLaMA-3, Mistral, Qwen) reduces V outlier counts by $6\text{--}27\times$ while Q counts remain high. Under GQA, shared key-value heads appear to operate in a different spectral regime, with representational capacity concentrating in Q and O. This suggests that the reduced

TABLE II
AVERAGE MP SPECTRAL OUTLIERS (SIGNAL) PER LAYER AND MEAN ENERGY RATIO (%).

Model	Q		K		V		O	
	avg	e%	avg	e%	avg	e%	avg	e%
<i>Encoder-Only (MHA)</i>								
BERT-base	275	86.5	276	86.7	268	82.5	261	79.7
RoBERTa-base	282	87.5	282	87.6	271	83.6	253	79.6
<i>Decoder-Only (MHA)</i>								
OPT-125M	299	91.8	296	90.5	271	83.5	267	82.7
LLaMA-1-7B	1509	88.7	1509	88.5	1391	79.9	1398	79.7
LLaMA-2-7B	1519	87.8	1517	87.5	1392	79.0	1393	79.1
Phi-3-mini	—	—	—	—	—	—	1051	79.8
<i>Decoder-Only (GQA)</i>								
LLaMA-3-8B	1535	90.1	341	75.4	219	43.1	1490	86.2
Mistral-7B	1511	87.5	341	74.7	212	43.6	1450	84.6
Qwen2.5-0.5B	350	93.6	44	71.1	13	15.6	349	91.9
Qwen2.5-1.5B	587	92.7	86	71.2	27	18.2	570	88.4
Qwen2.5-7B	1324	89.0	170	68.8	73	25.7	1290	85.6

TABLE III
TOTAL ENTRY-LEVEL OUTLIER COUNTS ($|W_{ij}| > 4\sigma_W$) SUMMED OVER ALL LAYERS, WITH THE PERCENTAGE SHARE OF EACH MODEL'S TOTAL. QWEN2.5-1.5B AND PHI-3-MINI ARE OMITTED.

Model	Q		K		V		O	
	Q	Q%	K	K%	V	V%	O	O%
<i>MHA models</i>								
BERT-base	3,305	25.5	3,306	25.5	3,213	24.8	3,127	24.1
RoBERTa-base	3,381	25.9	3,388	25.9	3,251	24.9	3,037	23.3
OPT-125M	3,593	26.4	3,547	26.1	3,257	23.9	3,209	23.6
LLaMA-1-7B	48,295	26.0	48,288	26.0	44,521	24.0	44,733	24.1
LLaMA-2-7B	48,599	26.1	48,542	26.1	44,553	23.9	44,583	23.9
<i>GQA models</i>								
LLaMA-3-8B	49,133	42.8	10,921	9.5	7,012	6.1	47,687	41.6
Mistral-7B	48,365	43.0	10,915	9.7	6,785	6.0	46,406	41.3
Qwen2.5-0.5B	8,400	46.3	1,059	5.8	315	1.7	8,366	46.1
Qwen2.5-7B	37,080	46.4	4,758	5.9	2,046	2.6	36,107	45.1

number of KV heads produces a regime in which the V signal is spectrally diffuse: learned structure is distributed across many singular values rather than concentrated in a few dominant ones.

C. Entry-Level Outlier Spatial Patterns

The heatmaps (Figs. 2–3) reveal coherent spatial structures: **row-bands in Q** (corresponding to specific attention heads with disproportionate learned structure), **column-bands in O** (indicating preferred output dimensions for attention contributions), and **sparse V under GQA** (consistent with low spectral outlier counts). Table III quantifies these counts: under GQA, Q and O together account for roughly 85–93% of all entry-level outliers, against a near-uniform 25% per projection under MHA. Outlier density in O peaks in middle layers (8–20), mirroring known importance of middle-layer representations [22]. We emphasize spatially repeated bands rather than individual high-magnitude entries, since isolated 4σ entries can occur by chance in very large matrices.



Fig. 2. Entry-level outlier locations for **LLaMA-2-7B (MHA, left)** and **LLaMA-3-8B (GQA, right)**. Under GQA, V becomes extremely sparse while Q retains row-bands and O shows column-bands.

D. Cross-Layer Residual-Stream Alignment

Table IV reports persistent dimensions per component. K dominates cross-layer persistence in MHA models (BERT: K=7 vs Q=3; LLaMA-2: K=52 vs Q=33), while V has near-zero persistence under MHA. GQA creates V persistence (LLaMA-3: 44 vs LLaMA-2: 1), as fewer KV heads may force stronger specialization.

Specific O dimensions persist across nearly all layers with remarkable consistency (e.g., LLaMA-1 dim 3840: 32/32 layers; LLaMA-2 dim 1512: 32/32 layers; RoBERTa dim 588:



Fig. 3. Mistral-7B (left) and Qwen2.5-7B (right). Both GQA models reproduce the Q-dense / V-sparse / O-banded pattern.

12/12 layers), representing candidate “write channels.”

Pattern 4: Persistent residual-stream channels. A small set of residual-stream dimensions acts as a band outlier in K and O across nearly all layers, forming candidate privileged communication channels.

Table V shows top cross-component persistent dimensions. In LLaMA-2-7B, dimension 2533 appears in all four components and is persistent in Q, K, and O, while appearing in V in one layer. It appears as a Q outlier in all 32 layers—a simultaneously dominant query direction, frequent key direction, and preferred output channel. Dimension 1512,

TABLE IV
PERSISTENT RESIDUAL-STREAM DIMENSIONS (≥ 3 LAYERS AS A BAND OUTLIER) PER COMPONENT. MISTRAL-7B, THE QWEN2.5 FAMILY AND PHI-3-MINI ARE OMITTED.

Model	Q	K	V	O	Union
BERT-base	3	7	0	2	11
RoBERTa-base	8	10	0	1	11
OPT-125M	15	17	0	12	24
LLaMA-1-7B	21	51	2	15	52
LLaMA-2-7B	33	52	1	17	54
LLaMA-3-8B	87	98	44	43	164

TABLE V
TOP CROSS-COMPONENT PERSISTENT DIMENSIONS. “HITS” IS THE NUMBER OF Q, K, V, AND O COMPONENTS IN WHICH THE DIMENSION APPEARS AT LEAST ONCE; INCLUDED DIMENSIONS ARE PERSISTENT IN AT LEAST ONE COMPONENT.

Model	Dim	Hits	Total	Q	K	V	O
BERT	381	3	16	7	8	0	1
	308	1	8	0	0	0	8
OPT	174	3	23	10	10	0	3
	638	1	7	0	0	0	7
LLaMA-2	2533	4	58	32	12	1	13
	3431	3	61	25	28	0	8
	1512	3	37	3	2	0	32
LLaMA-3	2977	3	54	24	22	0	8
	4055	1	30	0	0	0	30
	373	3	48	16	23	0	9

by contrast, is almost exclusively an O-channel (32/32 layers), serving as a dedicated write-heavy channel.

E. 3D Attention Stack Visualizations

Figs. 4 and 5 present 3D visualizations of cross-layer alignment for BERT-base and LLaMA-2-7B. BERT’s stack is sparse with persistence concentrated in K and O spanning only a fraction of layers. LLaMA-2’s stack is dense, with dozens of dimensions forming continuous lines across all 32 layers, reflecting deeper architecture and greater specialization.

IV. CAUSAL VALIDATION VIA ABLATION

A. Experimental Design

We conduct ablations on LLaMA-1-7B [11], LLaMA-2-7B [12] (MHA), LLaMA-3-8B [13], and Mistral-7B [14] (GQA). Table VI reports all configurations.

B. Spectral Outliers Carry a Dominant Learned Structure

Zeroing all 112,458 spectral outliers in Mistral-7B reduces HellaSwag from 0.814 to 0.256, MMLU from 0.627 to 0.269, and PIQA from 0.823 to 0.508, as reported in Table VI. These values are close to random-chance performance, providing causal evidence that MP-identified spectral outliers encode a dominant learned component in this model.

Pattern 5: Spectral outliers capture a dominant learned structure.

Conversely, zeroing count-matched below-threshold singular values preserves much of the model’s capability: in

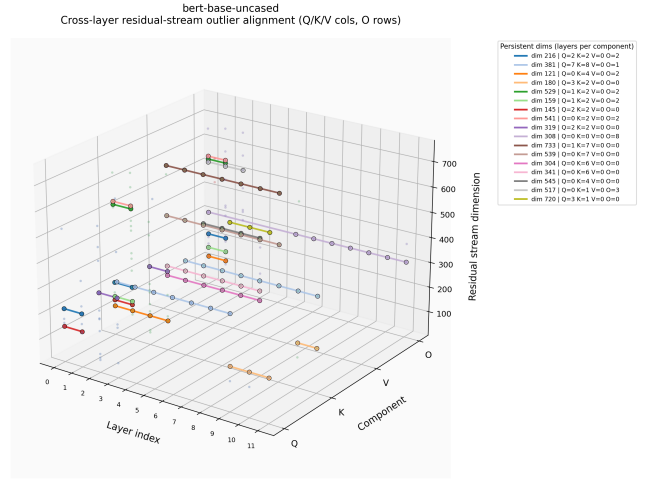


Fig. 4. 3D attention stack for BERT-base. Persistent dimensions concentrate in K and O; V shows none.

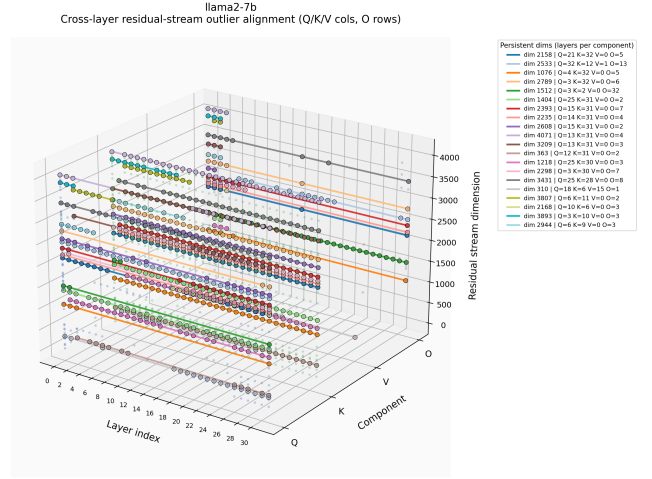


Fig. 5. 3D attention stack for LLaMA-2-7B. Dense persistent dimensions span nearly all 32 layers.

LLaMA-1-7B, zeroing 139,423 randomly selected below-threshold singular values matched to 75% of the MP-outlier count causes only a 0.7-point HellaSwag drop. However, MMLU shows greater sensitivity (6–13 point drops), suggesting knowledge-intensive tasks rely on information distributed more broadly across the spectrum, including within components classified as “bulk” by the MP threshold. This indicates that while the MP decomposition captures the primary signal, the bulk is not purely noise—it contains secondary learned structure that contributes to knowledge-intensive tasks.

C. Per-Component Criticality and V Bulk Catastrophe

K has highest per-parameter criticality in the LLaMA-3-8B entry-outlier ablations: zeroing 234,239 K entry outliers produces a 21.8-point MMLU drop, while zeroing 883,337 Q entry outliers ($3.8\times$ more) produces a 17.2-point drop. V entry outliers are least critical (4.8-point MMLU drop).

TABLE VI

ABLATION RESULTS. N_{ZERO} = SINGULAR VALUES/ENTRIES ZEROED. HS = HELLASWAG, PIQA = ACC_NORM. RANDOM CHANCE: HS $\approx$ 0.250, MMLU = 0.250, PIQA $\approx$ 0.500.

Model	Strategy	Comp.	f	N_{zero}	HS	Δ HS	MMLU	Δ MMLU	PIQA	Δ PIQA
<i>Baselines</i>										
LLaMA-1-7B	—	—	—	—	.761	—	.352	—	.793	—
LLaMA-2-7B	—	—	—	—	.760	—	.458	—	.790	—
LLaMA-3-8B	—	—	—	—	.821	—	.660	—	.812	—
Mistral-7B	—	—	—	—	.814	—	.627	—	.823	—
<i>Bulk (below-threshold) removal</i>										
LLaMA-1-7B	random_bulk	QKVO	0.75	139,423	.754	−.007	.292	−.060	.791	−.002
LLaMA-2-7B	random_bulk	QKVO	0.75	139,754	.737	−.023	.358	−.100	.773	−.017
LLaMA-2-7B	random_bulk	QKVO	1.00	186,273	.710	−.050	.329	−.129	.752	−.038
<i>Outlier (signal) removal</i>										
Mistral-7B	outliers	QKVO	1.00	112,458	.256	−.558	.269	−.358	.508	−.315
<i>Per-component ablations (LLaMA-3-8B)</i>										
LLaMA-3-8B	entry_outliers	Q	1.00	883,337	.757	−.064	.488	−.172	.792	−.020
LLaMA-3-8B	entry_outliers	K	1.00	234,239	.728	−.093	.442	−.218	.770	−.042
LLaMA-3-8B	entry_outliers	V	1.00	98,866	.793	−.028	.612	−.048	.799	−.013
LLaMA-3-8B	random_bulk	O	1.00	47,676	.789	−.032	.606	−.054	.797	−.015
LLaMA-3-8B	random_bulk	V	1.00	7,013	.649	−.172	.254	−.406	.779	−.033

TABLE VII

MMLU SUBCATEGORY ACCURACY UNDER SELECTED ABLATIONS.

Experiment	STEM	Hum	SocSci	Other
Mistral outliers QKVO	.286	.242	.311	.251
LL2 random-bulk QKVO $f=0.75$	.305	.335	.386	.420
LL3 entry K	.394	.375	.541	.494
LL3 entry Q	.439	.403	.599	.558
LL3 entry V	.536	.538	.717	.698
LL3 random-bulk V	.252	.246	.255	.265

The V bulk ablation reveals a regime difference under GQA: zeroing only 7,013 below-threshold V singular values causes catastrophic MMLU damage (to 0.254). Table VII breaks these results down by MMLU subcategory. The degradation is broadly uniform across subcategories for the two collapse cases (Mistral outlier removal and LLaMA-3 V bulk removal), whereas the milder per-component ablations preserve social-science and “other” accuracy noticeably better than STEM and humanities.

V. CONNECTIONS TO MECHANISTIC INTERPRETABILITY

Our analyses provide three complementary lenses. **Entry-level heatmaps** connect to the circuits framework [2]: row-bands in Q indicate head-specific structure potentially corresponding to specialized attention behavior, while column-bands in O suggest preferential writing to specific residual stream dimensions [23]. **Cross-layer alignment** reveals that residual-stream communication is not uniform but concentrated on privileged highway dimensions—K-persistent dimensions may encode features consistently attended to, while O-persistent dimensions carry features repeatedly updated by attention layers. **Causal ablations** close the loop by confirming that these structures are not merely visual artifacts:

the $K > Q \gg V$ hierarchy reflects K’s sparse but persistent outlier dimensions concentrating disproportionate functional importance in the tested LLaMA-3 setting.

VI. RELATED WORK

RMT for neural networks. Martin and Mahoney [6] applied RMT to study implicit regularization via heavy-tailed spectral analysis, demonstrating that well-trained DNNs exhibit heavy-tailed eigenvalue distributions that go beyond the clean MP bulk-plus-outlier picture. Our work operates within the simpler MP framework [5] as a first-order approximation, but validates its utility through causal ablations. Yang *et al.* [7] analyzed spectral norm scaling conditions for feature learning at large width. We extend these with spatial outlier analysis, cross-layer alignment, and causal validation.

Mechanistic interpretability. The circuits framework [2] and induction head analysis [3] identify functional components via activations. Our weight-space analysis complements these by revealing persistent residual-stream communication channels, validated by ablation.

Efficient adaptation and pruning. LoRA [4] exploits low-rank structure; our MP analysis provides a possible basis for heterogeneous rank allocation—adaptation may benefit from targeting outlier-rich subspaces, with K projections receiving proportionally more capacity despite their smaller dimensions under GQA [17]. The V-projection exception cautions against uniform singular-value thresholding across projections.

VII. LIMITATIONS

MP framework assumptions. Our framework uses the standard MP/spiked-model threshold as a first-order approximation, but Martin and Mahoney [6] showed that trained DNN weight matrices often exhibit heavy-tailed spectral distributions. This means the “bulk” identified by the MP threshold

is not purely random noise—it may contain structured information that is spectrally diffuse. The MMLU sensitivity to bulk removal (6–13 point drops) empirically confirms this. A heavy-tailed RMT framework could yield a more refined decomposition.

Causal claims. Our ablation demonstrates that MP-identified outliers are functionally important and that removing them can collapse model performance. However, it does not prove that no learned structure exists in the bulk—the bulk removal experiments show modest but non-trivial degradation, especially on knowledge-intensive benchmarks. We therefore characterize outliers as encoding a “dominant” rather than exclusive learned structure.

Random bulk ablations. The random-bulk experiments are intended as diagnostic contrasts against MP outlier removal. A more complete robustness study would average random bulk subsets over multiple seeds and report variance.

Entry-level thresholding. The $|W_{ij}| > 4\sigma_W$ threshold identifies interpretable spatial bands, but isolated extreme values can occur in large random matrices. Future work should include shuffled-weight and randomly initialized controls to separate true learned spatial organization from threshold artifacts.

Noise variance estimation. The median-based estimator $\hat{\sigma}^2 = \text{median}(s_i^2)/(1+\gamma)$ is approximate—the exact median of the MP distribution depends on γ and differs slightly from this formula. This could affect threshold placement for matrices with unusual aspect ratios.

Model coverage. For Phi-3-mini, only O-projection data is reported, because the fused QKV projection prevented separate Q, K, and V extraction. For the same reason, and to respect the page limit, Tables III and IV report a subset of the models in Table I. Patterns observed in models with excluded projection types should therefore be interpreted cautiously.

VIII. CONCLUSION

We presented a systematic MP analysis of attention weights across 11 transformers, augmented with entry-level outlier heatmaps, cross-layer residual-stream alignment, and causal ablation experiments. We identified five recurring patterns: Q dominance in spectral structure; V sparsification under GQA; structured spatial organization in Q and O; persistent residual-stream highways in K and O; and the causal finding that spectral outliers encode a dominant learned structure, while below-threshold singular values are typically less critical though not purely noise. These findings connect random matrix theory with mechanistic interpretability and offer practical guidance for pruning, compression, and low-rank adaptation. The observed cross-layer structure is complementary to CRAFT [24], which exploits correlations across attention layers through a frozen Tucker decomposition for parameter-efficient fine-tuning. Future directions—in particular spectrum-aware LoRA, MP-guided structured pruning, and per-projection low-rank compression—represent concrete pathways for translating these observations into practical efficiency gains for large language models.

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.
- [2] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah, “A mathematical framework for transformer circuits,” *Transformer Circuits Thread*, Dec. 2021. [Online]. Available: <https://transformer-circuits.pub/2021/framework/index.html>
- [3] C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah, “In-context learning and induction heads,” *arXiv preprint arXiv:2209.11895*, 2022.
- [4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.
- [5] V. A. Marchenko and L. A. Pastur, “Distribution of eigenvalues for some sets of random matrices,” *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, pp. 457–483, 1967, doi: 10.1070/SM1967v001n04ABEH001994.
- [6] C. H. Martin and M. W. Mahoney, “Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning,” *Journal of Machine Learning Research*, vol. 22, no. 165, pp. 1–73, 2021.
- [7] G. Yang, J. B. Simon, and J. Bernstein, “A spectral condition for feature learning,” *arXiv preprint arXiv:2310.17813*, 2023.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
- [9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized BERT pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [10] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, “OPT: Open pre-trained transformer language models,” *arXiv preprint arXiv:2205.01068*, 2022.
- [11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “LLaMA: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [12] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [13] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, “The Llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.
- [14] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed, “Mistral 7B,” *arXiv preprint arXiv:2310.06825*, 2023.
- [15] Qwen Team: A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, “Qwen2.5 technical report,” *arXiv preprint arXiv:2412.15115*, 2024.
- [16] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl *et al.*, “Phi-3 technical report: A highly capable language model locally on your phone,” *arXiv preprint arXiv:2404.14219*, 2024.
- [17] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai, “GQA: Training generalized multi-query transformer models from multi-head checkpoints,” in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 4895–4901.

- [18] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, “HellaSwag: Can a machine really finish your sentence?,” in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2019, pp. 4791–4800.
- [19] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, “Measuring massive multitask language understanding,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [20] Y. Bisk, R. Zellers, R. Le Bras, J. Gao, and Y. Choi, “PIQA: Reasoning about physical commonsense in natural language,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 05, 2020, pp. 7432–7439.
- [21] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac’h, H. Li, K. McDonell, N. Muenighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou, “A framework for few-shot language model evaluation,” Zenodo, ver. v0.4.0, Dec. 2023, doi: 10.5281/zenodo.10256836.
- [22] G. Jawahar, B. Sagot, and D. Seddah, “What does BERT learn about the structure of language?,” in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Florence, Italy, 2019, pp. 3651–3657.
- [23] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah, “Toy models of superposition,” *arXiv preprint arXiv:2209.10652*, 2022.
- [24] K. Dewage, M. Pensky, S. De Silva, and S. Mondal, “LORA-CRAFT: Cross-layer rank adaptation via frozen Tucker decomposition of pre-trained attention weights,” *arXiv preprint arXiv:2602.17510*, 2026.