WHERE MASSIVE ACTIVATIONS COME FROM: CAUSAL EVIDENCE FROM MLP GEOMETRY IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) commonly exhibit massive activations (MAs) across Transformer architectures. MAs are sparse hidden-state coordinates whose magnitudes greatly exceed typical activations and concentrate at specific tokens and dimensions. These extreme values complicate quantization, compression, and efficient inference, yet their generation mechanism remains unclear. Existing studies mainly associate MAs with attention sinks or mitigate them as numerical outliers. However, these approaches do not explain how MAs are generated inside the network. In particular, the causal connection between token-level triggers, layer-wise sources, and weight-space geometry remains unresolved. We address this gap through a systematic study of 26 language models combining component ablations, singular value decomposition, and causal interventions. Our results show that MAs primarily arise from MLP down-projections, where function-token representations align with load-bearing right-singular directions that amplify and write extreme activations, while attention mainly regulates this pathway. These findings establish MAs as structured geometric events and provide a causal account of their emergence across architectures.

1 INTRODUCTION

Transformer language models (Vaswani et al., 2017; Brown et al., 2020; Touvron et al., 2023) often exhibit massive activations (MAs), hidden-state coordinates whose magnitudes are substantially larger than typical activations (Sun et al., 2024; Merrill et al., 2022). MAs are highly sparse, typically appearing at a small number of token positions, such as whitespace, punctuation, or beginning-of-sequence tokens, and concentrating on only a few hidden dimensions. Despite this sparsity, they can affect every forward pass and complicate model deployment by destabilizing low-precision inference (Dettmers et al., 2022; Bondarenko et al., 2023), interacting with attention sinks (Xiao et al., 2024), affecting long-context behavior (Liu et al., 2024b), and making quantization, pruning, and sharding more difficult (Liu et al., 2024a). However, although the empirical presence and practical impact of MAs are increasingly well documented, the mechanism that generates these extreme activations remains incompletely understood.

Existing studies approach MAs from several complementary perspectives. Work on attention sinks associates large activations with function words, punctuation, and other structural tokens (Xiao et al., 2024; Cancedda, 2024; Gu et al., 2025), while systems-oriented methods treat MAs as numerical outliers and mitigate them through smoothing, selective preservation, or precision-aware quantization (Dettmers et al., 2022; Xiao et al., 2023; Bondarenko et al., 2023; Liu et al., 2024c). More broadly, mechanistic interpretability and causal-intervention methods provide tools for identifying internal computational pathways and testing their causal roles (Elhage et al., 2021; Conmy et al., 2023; Pearl, 2009; Meng et al., 2022). These directions explain where MAs tend to occur and how their numerical effects can be controlled, but they do not provide a unified causal account of how an ordinary token representation becomes an extreme hidden-state coordinate.

The central gap is therefore to explain the full generation pathway of MAs by connecting token-level triggers, layer-level sources, and weight-space geometry. This requires resolving several intertwined challenges. First, attention and MLP computations interact through the residual stream, making

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.

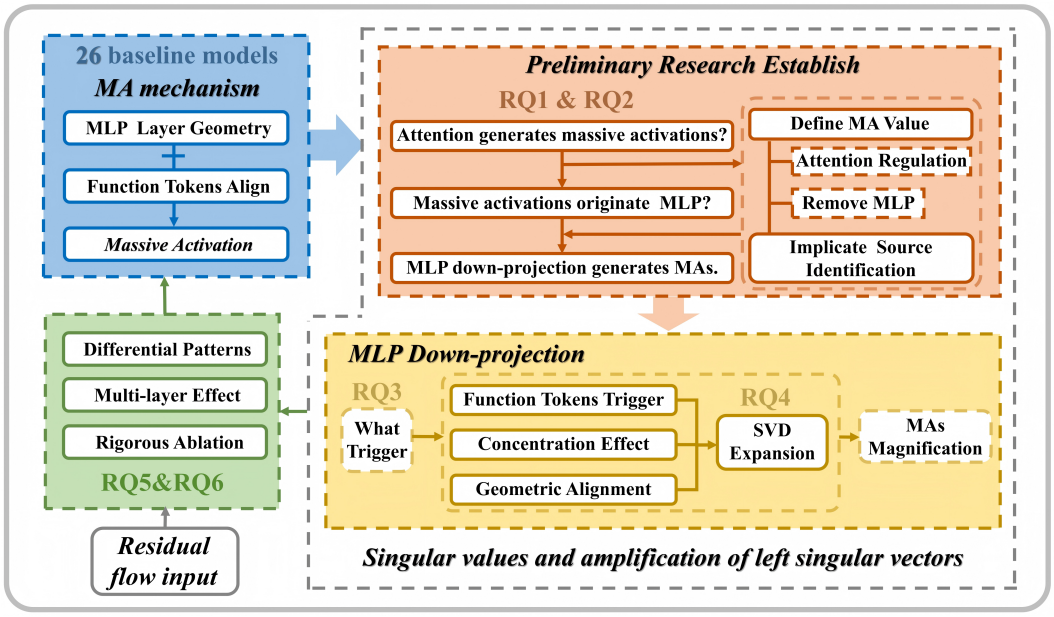


Figure 1: **MA research framework.** Experimental chain from source localization and function-token triggers to SVD amplification, causal ablation, and residual effects.

it difficult to distinguish the component that generates an MA from those that merely regulate or propagate it. Second, the source may be localized to one layer or distributed across multiple layers, so a single-layer explanation does not necessarily generalize across models. Third, the observed preference for function tokens may be confounded by token frequency rather than reflect a genuine mechanistic trigger. Finally, geometric correlations alone cannot establish whether the identified directions are necessary for MA generation. As illustrated in Figure 1, a mechanistic account must therefore jointly identify where MAs originate, what triggers them, how they are amplified, and whether the responsible pathway is causally necessary.

Our approach is motivated by the observation that the MLP down-projection forms the interface between intermediate MLP representations and the residual stream. We find that function-token positions produce intermediate representations that are more concentrated and more consistently aligned with specific input-side directions of the down-projection. These aligned components are subsequently amplified and written into a small set of output coordinates, whereas content-token representations are generally more diffuse and less consistently aligned with the same pathway. This observation suggests that MAs are not arbitrary numerical outliers, but structured geometric events arising when particular token representations align with learned projection directions that concentrate their contributions.

We study this hypothesis across 26 language models using the staged framework in Figure 1. We first separate attention and MLP pathways to localize where MAs originate and determine whether their support is concentrated in one layer or distributed across depth. We then test whether function-token preference persists after controlling for token frequency, analyze the projection geometry that maps these representations into extreme hidden coordinates, and directly perturb the identified directions to establish causal necessity. Across these analyses, we find that MAs primarily originate from the MLP down-projection, where function-token representations preferentially align with directions that amplify and write extreme activations, while attention mainly regulates this pathway. The same mechanism extends from localized single-layer sources to coordinated multi-layer accumulation.

In summary, this paper makes four contributions[1]:

- ▷ We identify the MLP down-projection as the primary source of MAs across a broad set of language models and characterize how this source varies from localized to multi-layer generation.

[1]Our code and implementation details are available at `https://anonymous.4open.science/r/MA-9F1A/`.

- ▷ We provide a geometric explanation for the function-token preference by showing that low-entropy function-token representations preferentially align with the input-side directions that drive MA amplification.
- ▷ we derive an SVD-based account that expresses MA generation as the accumulation of singular-direction contributions and validate the load-bearing role of this geometry through single-layer and multi-layer causal interventions.
- ▷ We organize the observed cross-model behaviors into CONCENTRATED, FEW-SOURCE, DISPERSED, and ANOMALY regimes, providing a unified view of MAs as structured geometric events whose manifestation depends on how their supporting pathway is distributed across depth.

2 RELATED WORK

Massive activations. Sun et al. (2024) systematically catalogued MAs in modern language models and their correlation with function words. Subsequent work studies their impact on post-training quantisation (Dettmers et al., 2022; Xiao et al., 2023; Bondarenko et al., 2023; Shao et al., 2024), accelerator error tolerance (Liu et al., 2024a), long-context behaviour (Liu et al., 2024b), and pivot-token preservation (Liu et al., 2024c); earlier, Merrill et al. (2022) analysed activation saturation from a formal-linguistic perspective. These works treat MAs mainly as a systems phenomenon to be controlled, whereas we ask why MAs arise and test the causal pathway.

Function words, syntactic probes, and attention sinks. Probing work shows that function words carry disproportionate syntactic signal (Hewitt & Manning, 2019; Tenney et al., 2019; Jawahar et al., 2019), and attention-head analyses identify heads specialised for structural functions (Voita et al., 2019). Attention sinks, where attention mass concentrates on a few short structural tokens, were named by Xiao et al. (2024) and independently observed in vision Transformers (Darcet et al., 2024); later work links sink emergence to residual-stream spectral filtering and training-time optimiser dynamics (Cancedda, 2024; Gu et al., 2025). Our geometric account explains this token preference quantitatively: function tokens project onto the leading singular directions of $\mathbf{W}_{\text{down}}$ differently from content words, matching the Top-1 MA concentration observed in RQ3.

SVD geometry and causal intervention. SVD analyses of Transformer weights have mostly served compression, such as activation-aware low-rank decomposition for models with outlier activations (Yuan et al., 2023). Causal-intervention methods such as ROME (Meng et al., 2022), grounded in Pearl's framework for causality (Pearl, 2009), support ablation-based attribution, and automated circuit discovery extends such interventions to larger subgraphs (Conmy et al., 2023). We combine these strands with a new V-matrix ablation and a closed-form prediction (Eq. (9)), providing a quantitative bridge between SVD geometry and the generation pathway of activations.

3 METHOD

Our method follows three stages, as illustrated in Figure 2: (1) localizing the MLP source of MAs, (2) connecting function-token triggers to SVD-based amplification, and (3) validating the identified geometry through causal interventions. We instantiate these three stages through six research questions described below.

3.1 MLP SOURCE LOCALIZATION

RQ1: Attention Regulation. RQ1 tests whether attention directly generates MAs. If attention writes the extreme coordinate, then zeroing every attention output (i.e. setting $\mathbf{H}^{(\ell)}_{\text{attn}} \leftarrow \mathbf{0}$ at every layer through forward hooks) should remove the Top-1 activation at the baseline trigger layer. If a substantial extreme activation remains, the MA has a non-attention source.

This intervention removes the attention branch output while preserving the residual stream and the MLP path. We then measure the residual ratio

$$\rho_{\text{attn}} = \frac{\text{Top}_1^{(\text{no-attn})}}{\text{Top}_1^{(\text{base})}}. \tag{1}$$

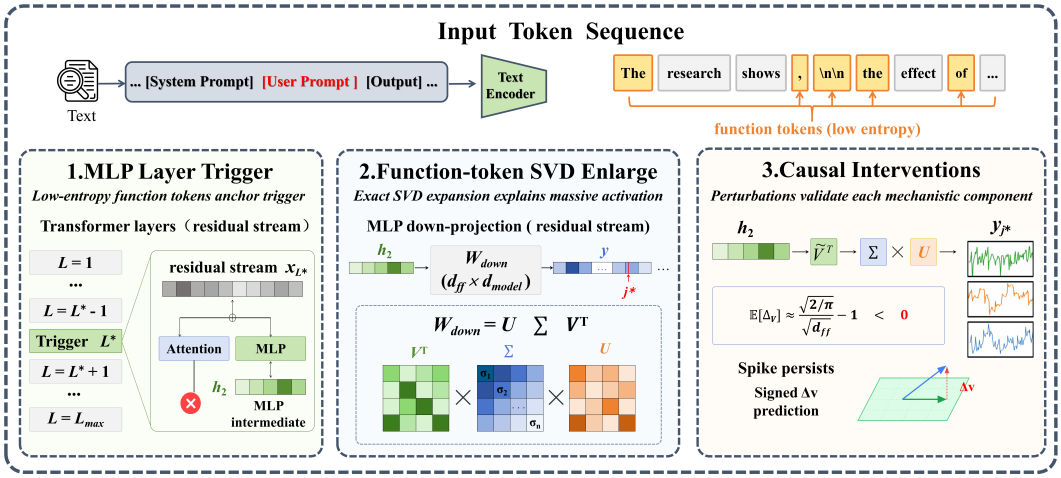


Figure 2: **Method framework.** Our framework has three parts: the left panel localises the MLP source tokens, the middle panel performs the SVD geometric decomposition, and the right panel carries out the causal ablations.

The ratio ρ_{attn} measures MA dependence on attention removal. Values clearly above 0 or above 1 are inconsistent with attention as a unified origin. Values close to 0 indicate high sensitivity to the attention path, but they do not by themselves identify attention as the writing source because attention removal may also cut later amplification or propagation.

RQ2: MLP Source and Layer Support. RQ2 tests whether the MLP is the main MA source and whether that source is single-layer or multi-layer. The local object is the MLP down-projection at the trigger layer; let $\mathbf{h}_2 \in \mathbb{R}^{d_{\text{ff}}}$ denote the MLP intermediate state that feeds $\mathbf{W}_{\text{down}} \in \mathbb{R}^{D \times d_{\text{ff}}}$, so that the MLP output and the MA coordinate are

$$\mathbf{y} = \mathbf{W}_{\text{down}}\mathbf{h}_2 + \mathbf{b}_{\text{down}}, \qquad j^* = \arg\max_j |y_j|. \tag{2}$$

We probe the source through ablations of increasing granularity.

The global ablation disables all MLP outputs while preserving attention and the residual stream (the update $h_{\ell+1} = h_\ell + \text{Attn}_\ell(h_\ell) + \text{MLP}_\ell(h_\ell)$ becomes $h_{\ell+1} = h_\ell + \text{Attn}_\ell(h_\ell)$), and reports the retention ratio

$$\tau_{\text{all}} = \frac{\text{Top}_1^{(\text{all-mlp-zero})}}{\text{Top}_1^{(\text{base})}}. \tag{3}$$

Small τ_{all} provides evidence that the MA depends on the MLP path; large τ_{all} is consistent with a non-MLP bypass such as routing dilution or hybrid-attention paths.

Per-layer ablation isolates one MLP layer at a time by setting $\mathbf{H}_{\text{mlp}}^{(\ell)} \leftarrow \mathbf{0}$, and ranks layers by the induced drop in Top-1. The top-ranked layer is a candidate source, but in multi-layer models the largest single drop need not be the first origin layer.

Greedy cumulative ablation estimates how many layers carry the MA. Let $\ell_1, \ell_2, \ldots, \ell_m$ be the layer order from the per-layer scan, and $\mathcal{S}_s = \{\ell_1, \ldots, \ell_s\}$ the first s layers. Setting $\mathbf{H}_{\text{mlp}}^{(\ell)} \leftarrow \mathbf{0}$ for every $\ell \in \mathcal{S}_s$ gives the cumulative retention ratio

$$\tau(\mathcal{S}_s) = \frac{\text{Top}_1^{(\text{mlp-zero}, \mathcal{S}_s)}}{\text{Top}_1^{(\text{base})}}. \tag{4}$$

The decay of $\tau(\mathcal{S}_s)$ as s grows assigns each model to one of four regimes (CONCENTRATED, FEW-SOURCE, DISPERSED, ANOMALY), with layer-support operators and regime definitions in Appendix C.2.

3.2 FUNCTION-TOKEN SVD AMPLIFICATION

RQ3: Function-Token Triggering. RQ3 tests whether high-MA positions systematically fall on function tokens, and separates function-token labels from token frequency.

The function-token set $\mathcal{F}$ is deterministic and contains closed-class function words, structural symbols, digit tokens, and short BPE fragments (full token rules in Appendix B.1). At the trigger layer we record Top-K MA tokens and compute the function-token trigger rate $\pi_{\text{func}}(K)$ at $K \in \{1, 5, 10, 20\}$ (formal definition in Appendix B.1). Deviation from the same-corpus baseline is tested with Fisher's exact test on a 2×2 contingency table.

To further separate function-token labels from token frequency, we partition tokens into four quadrants by frequency and function-token status (high/low frequency $\times$ function/content; quadrant definitions in Appendix B.1) and compare per-quadrant MA trigger rates. We then fit a frequency-controlled logistic regression

$$\operatorname{logit}\big(P(t \text{ triggers MA})\big) = \beta_0 + \beta_{\mathcal{F}} \mathbb{1}[t \in \mathcal{F}] + \beta_{\text{freq}} \log \text{freq}(t). \tag{5}$$

The coefficient $\beta_{\mathcal{F}}$ measures the independent contribution of the function-token label after controlling for token frequency, providing a test for whether the function-token preference can be explained by frequency alone.

RQ4: SVD Magnitude Prediction and Direction Consistency. RQ4 connects the MLP source identified in RQ2 to the function-token trigger positions identified in RQ3 through the SVD geometry of $\mathbf{W}_{\text{down}}$. Performing the singular-value decomposition $\mathbf{W}_{\text{down}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \sum_{i=1}^{r} \sigma_i \, \mathbf{u}_i \, \mathbf{v}_i^\top$ with $r = \operatorname{rank}(\mathbf{W}_{\text{down}})$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$, and substituting into Eq. (2), yields the algebraic identity at the MA coordinate

$$\text{MA}_{j^*} = \sum_{i=1}^{r} \sigma_i \, (\mathbf{v}_i^\top \mathbf{h}_2) \, u_i[j^*] + (\mathbf{b}_{\text{down}})_{j^*}. \tag{6}$$

When $\sigma_1 \gg \sigma_2$ and $|\mathbf{v}_1^\top \mathbf{h}_2| \approx \|\mathbf{h}_2\|_2$, Eq. (6) reduces to a single-direction linear form

$$\text{MA}(x) \approx \beta_{\text{slope}} \, (\mathbf{v}_1^\top \mathbf{h}_2(x)) + b, \qquad \beta_{\text{slope}} = \sigma_1 \, u_1[j^*]. \tag{7}$$

RQ4 tests two predictions of Eq. (6): that a low-rank truncation reproduces MA magnitude, and that the input-side singular directions select the function-token positions of RQ3.

For magnitude prediction, the k-term approximation $\widehat{\text{MA}}_{j^*}^{(k)}$ retains the first k singular directions of Eq. (6), and we report the relative error $e^{(k)} = |\text{MA}_{j^*} - \widehat{\text{MA}}_{j^*}^{(k)}|/|\text{MA}_{j^*}|$ at $k \in \{1, 3, 5, 10, 20\}$. A small $e^{(k)}$ at low k provides evidence that few directions account for the observed extreme coordinate; a large $e^{(k)}$ at $k = 20$ indicates that direction consistency rather than spectral concentration is operative. In this regime, the per-direction contribution $\sigma_i \, (\mathbf{v}_i^\top \mathbf{h}_2) \, u_i[j^*]$ retains a constant sign across $i \in \{1, \ldots, k\}$, so the leading directions accumulate at j^* rather than cancel and the MA can form even when no single direction dominates.

For direction interpretation, we compute the input-side projection $p_i(x) = \mathbf{v}_i^\top \mathbf{h}_2(x)$, rank tokens by $|p_i(x)|$ for each leading $\mathbf{v}_i$, and report the function-token rate among the highest-projection tokens. If those positions are function-token enriched, the token-level pattern in RQ3 traces back to the input-side geometry of $\mathbf{W}_{\text{down}}$ rather than to token frequency alone.

3.3 CAUSAL INTERVENTIONS

From correlation to causality: RQ5 and RQ6 turn the correlational evidence of RQ1–RQ4 into causal tests by ablating the right-singular geometry on single-source and multi-source layers respectively.

RQ5: Single-Layer Causal Ablation of Right-Singular Geometry. RQ5 tests whether the right-singular geometry of $\mathbf{W}_{\text{down}}$ is causal at single-source layers by replacing $\mathbf{V}$ at the trigger layer with a Haar-uniform random orthogonal matrix $\widetilde{\mathbf{V}}$, preserving $\mathbf{\Sigma}$ and $\mathbf{U}$ while breaking the alignment between $\mathbf{h}_2$ and the original right-singular directions:

$$\mathbf{W}_{\text{down}}^{(\text{abl})} = \mathbf{U} \, \mathbf{\Sigma} \, \widetilde{\mathbf{V}}^\top, \qquad \widetilde{\mathbf{V}} \sim \text{Haar}. \tag{8}$$

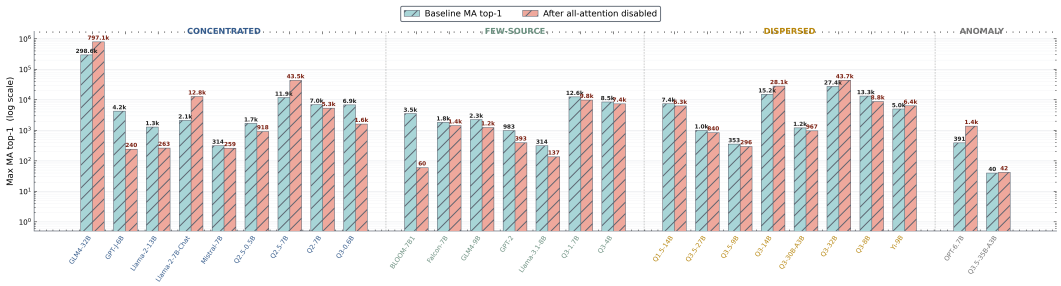


Figure 3: **Attention-ablation results for RQ1.** Removing attention produces three response modes across models: MA collapse, MA attenuation, and MA increase.

In the rank-1 dominated case (let $p = \mathbf{v}_1^\top \mathbf{h}_2$ with $|p| \approx \|\mathbf{h}_2\|_2$), concentration on the sphere gives the closed-form prediction

$$\mathbb{E}[\Delta_V] \approx \sqrt{2/\pi}/\sqrt{d_{\text{ff}}} - 1, \tag{9}$$

which serves as the single-layer random-V baseline. We also remove the leading K right-singular directions (replacing $\mathbf{W}_{\text{down}}$ by $\mathbf{W}_{\text{down}}(I - \sum_{i=1}^{K} \mathbf{v}_i \mathbf{v}_i^\top)$ for $K \in \{1, 3, 10, 20\}$; details in Appendix B.3). A collapse in MA under either intervention provides evidence that the single-layer right-singular geometry is load-bearing.

RQ6: Multi-Layer Causal Ablation of Macro Directions. RQ6 tests whether macro right-singular directions are causal at multi-source layers by removing the leading direction of the cumulative cross-layer residual change. Given the supporting layer set $\mathcal{L} = \{\ell_1, \ldots, \ell_m\}$ from RQ2, we apply SVD to $\Delta \mathbf{h}_{\text{macro}} = \mathbf{h}_{\text{out}}^{(\max \mathcal{L})} - \mathbf{h}_{\text{in}}^{(\min \mathcal{L})} = \mathbf{U}^{\text{macro}} \mathbf{\Sigma}^{\text{macro}} (\mathbf{V}^{\text{macro}})^\top$ and project the leading macro direction $\mathbf{v}_1^{\text{macro}}$ out of each supporting layer:

$$\mathbf{W}_{\text{down}}^{(\ell,\text{abl})} = \mathbf{W}_{\text{down}}^{(\ell)} \big(I - \mathbf{v}_1^{\text{macro}} (\mathbf{v}_1^{\text{macro}})^\top\big), \qquad \ell \in \mathcal{L}. \tag{10}$$

A collapse in MA under this intervention provides evidence that FEW-SOURCE and DISPERSED models rely on coherent cross-layer accumulation rather than on a single down-projection.

4 EXPERIMENTS AND RESULTS

We evaluate the six research questions introduced in §3 across 26 language models. The complete experimental setup, including the model pool, datasets, evaluation protocol, and statistical procedures, is provided in Appendix A.

4.1 (RQ1) ATTENTION ABLATION

RQ1 asks whether attention directly generates MAs. As shown in Figure 3, removing attention produces three qualitatively different responses across models: the MA nearly disappears in some models, is substantially attenuated in others, and becomes larger in a third group. Appendix Table 1 summarizes these responses in the RQ1 mode column, where Gen denotes attention-supported generation, Sup denotes attention suppression, and Gen* denotes a hybrid or partial generation mode.

The heterogeneous response suggests that attention is not a universal source of MAs. If attention directly wrote the extreme coordinate, removing it should consistently eliminate the MA across models. Instead, the direction and magnitude of the change vary substantially, indicating that attention mainly regulates the pathway that produces the MA. Depending on the architecture, attention can strengthen, suppress, or redirect the representations that later enter the MA-generating component. This motivates RQ2, where we explicitly search for the component that more consistently accounts for the extreme activation.

4.2 (RQ2) MLP SOURCE AND LAYER SUPPORT

RQ2 tests whether the MLP is the primary source of MAs and how this source is distributed across depth. Removing the MLP pathway substantially suppresses the MA in $23/26$ models, providing

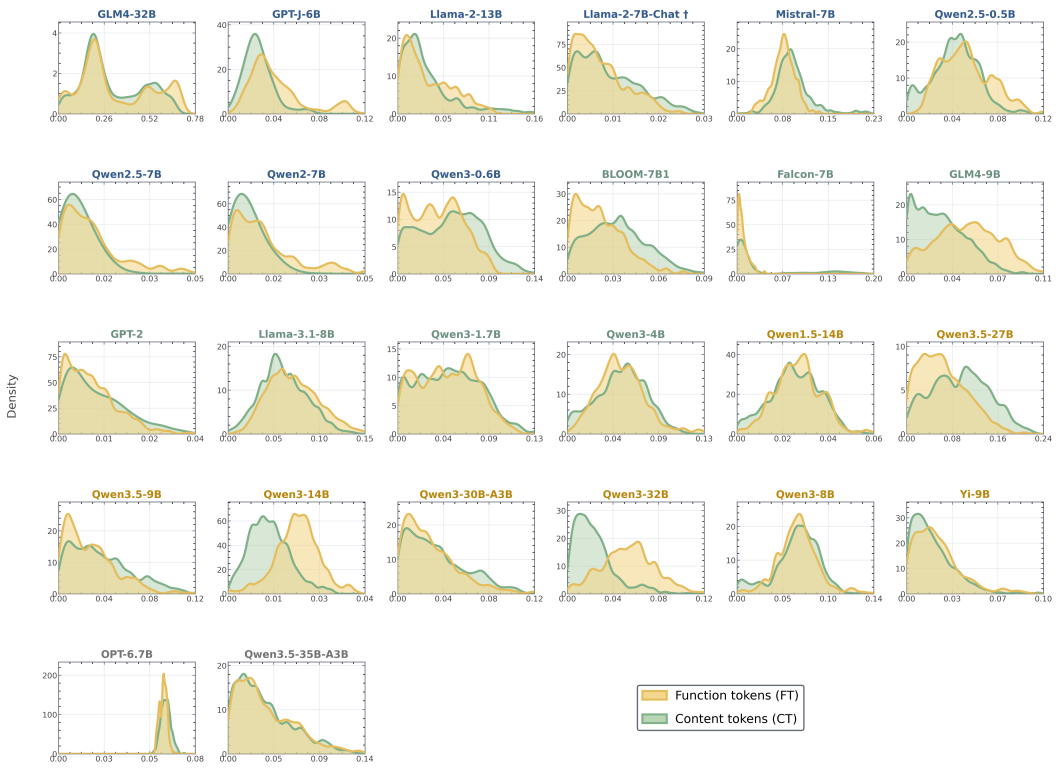


Figure 4: **Function-token results for RQ3.** Function-token enrichment is strongest at the highest-activation positions and approaches the corpus baseline as lower-ranked positions are included.

strong evidence that the MLP is the dominant source in most architectures (Geva et al., 2021). The per-layer ablations further show that this dependence is highly structured: only a subset of layers produces large changes in MA magnitude, rather than every MLP layer contributing equally.

The cumulative ablations reveal two distinct patterns of source support. In some models, removing a single MLP layer is sufficient to eliminate most of the MA, whereas in others several layers must be removed before the activation collapses. Based on this layer support, the 26 models separate into 9 single-source cases (CONCENTRATED), 15 multi-source cases (FEW-SOURCE and DISPERSED), and 2 ANOMALY cases. Thus, while the MLP pathway is the dominant source across architectures, the mechanism can be either localized to one layer or accumulated across multiple layers. Supplementary layer-wise evidence and the operational regime definitions are provided in Appendix C.

4.3 (RQ3) FUNCTION-TOKEN DISTRIBUTION

Having localized the source to the MLP pathway, RQ3 asks which token positions preferentially trigger MAs. Across models, high-activation positions are strongly enriched for function tokens. The Top-1 MA position is a function token in the vast majority of models, and this enrichment remains visible at Top-5, Top-10, and Top-20. As shown in Figure 4, the function-token proportion gradually decreases as lower-ranked positions are included, suggesting that the effect is concentrated on a relatively small set of strongly associated token positions.

A natural alternative explanation is that function tokens trigger MAs simply because many of them occur frequently. Our frequency-controlled analysis does not support this explanation. In particular, low-frequency function tokens trigger MAs at a rate 43.7 times higher than high-frequency content tokens. The function-token preference therefore persists even when token frequency is separated from token category. Together, these observations identify function-token positions as a characteristic input-side trigger of the MA pathway rather than merely a consequence of corpus frequency.



Figure 5: **Causal ablation results for RQ5 and RQ6.** Breaking right-singular alignment sharply suppresses MAs in both single-source and multi-source regimes.

4.4 (RQ4) SVD Magnitude Prediction and Direction Consistency

RQ4 connects the token-level preference identified in RQ3 to the geometry of the MLP down-projection. For each model, we measure the input-side projection $p_i(x) = \mathbf{v}_i^\top \mathbf{h}_2(x)$ defined in §3.2. If the right-singular directions of $\mathbf{W}_{\text{down}}$ participate in MA generation, then positions with large projections onto these directions should coincide with the token positions identified in RQ3. This is indeed the case: the Top-1 token under the $\mathbf{v}_1$ projection is a function token in $24/26$ models (92.3%), compared with a baseline of approximately 53%. The rate decreases to 68%, 62%, and 54% at Top-5, Top-10, and Top-20, respectively, approaching the baseline as weaker projection positions are included.

We next test whether the same geometry can account for MA magnitude. The truncation error $e^{(k)}$ is at most 30% for $11/22$ analyzed models using only the leading singular direction, including 6 near-exact cases, and for $16/22$ models when the first 20 directions are retained. These results support the SVD expansion in Eq. (6): the extreme coordinate is formed by singular-direction contributions that accumulate at the same output dimension j^*. In spectrally dominated models, one or a few directions explain most of the activation, whereas flatter-spectrum models require several contributions with consistent signs. RQ4 therefore links the function-token trigger to a concrete geometric channel in the MLP down-projection rather than treating the two observations as independent phenomena.

4.5 (RQ5) Single-Layer Causal Ablation

RQ4 establishes a strong relationship between MA magnitude and right-singular geometry, but this relationship alone does not establish causality. RQ5 therefore directly tests whether the identified geometry is necessary for MA generation in single-source models. As shown in Figure 5, all $9/9$ single-source models satisfy the causal criterion defined in §3.3. On the spectrally dominated subset, breaking the original right-singular alignment reduces the MA by approximately 93%–99%.

The intervention is designed to isolate directional alignment. Randomizing $\mathbf{V}$ preserves the singular values $\mathbf{\Sigma}$ and the output-side directions $\mathbf{U}$ while destroying the original alignment between $\mathbf{h}_2$ and the right-singular directions. The resulting MA collapse therefore cannot be attributed simply to reducing the scale of the weight matrix or removing its output directions. Instead, it identifies the alignment encoded by $\mathbf{V}$ as a load-bearing component of MA generation. This result turns the geometric relationship observed in RQ4 into direct causal evidence for the single-source regime.

4.6 (RQ6) MULTI-LAYER CAUSAL ABLATION

RQ6 asks whether the same causal mechanism extends to models whose MAs are supported by multiple layers. Among the 15 multi-source models, removing the dominant cross-layer direction collapses the MA in all $12/12$ clean dense cases, including 7 FEW-SOURCE and 5 DISPERSED models, with Δ_V typically at or below -80%. The remaining three checkpoints use hybrid-attention or MoE architectures, where alternative paths or expert routing weaken the effect of a single shared intervention.

For these multi-source models, no individual MLP layer necessarily carries the complete MA mechanism. Instead, the relevant direction is accumulated across the supporting layers. Projecting out $\mathbf{v}_1^{\mathrm{macro}}$ removes this shared cross-layer direction while preserving each layer's $\mathbf{\Sigma}^{(\ell)}$ and $\mathbf{U}^{(\ell)}$. The resulting collapse extends the RQ5 finding beyond localized sources: the same right-singular mechanism can operate either within one dominant layer or through coherent accumulation across depth. Taken together, RQ5 and RQ6 provide causal evidence that right-singular geometry is a common organizing principle of MA generation across both single-source and multi-source regimes.

5 DISCUSSION

5.1 MECHANISM OF MA GENERATION

Our results support a unified mechanism in which MAs arise from the interaction between function-token triggers and MLP projection geometry. Function-token representations preferentially align with right-singular directions of $\mathbf{W}_{\text{down}}$, which amplify these components and write them into sparse output coordinates, while attention mainly regulates rather than directly generates this pathway. The same mechanism appears either as localized generation in CONCENTRATED models or as cross-layer accumulation in FEW-SOURCE and DISPERSED models.

This account links token-level preference and geometric amplification within a single causal pathway. The strong alignment of function-token representations with MA-relevant singular directions, together with the collapse induced by $\mathbf{V}$-ablation and the prediction in Eq. (9), supports the load-bearing role of right-singular geometry in MA generation.

5.2 IMPLICATIONS AND LIMITATIONS

Implications. Our findings suggest that MA generation is governed more directly by projection geometry and residual-stream pathways than by normalization or activation families. The identified geometric sparsity may therefore provide a useful signal for locating MA-sensitive layers and directions. However, routed architectures introduce additional complexity: per-expert $\mathbf{V}$-ablation produces limited effects in MoE models, indicating that routing can dilute the influence of any individual projection pathway.

Limitations. Our study focuses on mainstream open-source models and does not examine frontier closed-source models or the training-time emergence of MAs. Although our interventions establish the causal importance of the $\mathbf{W}_{\text{down}}$ singular-vector pathway, their practical benefits for quantization, pruning, and runtime mitigation remain to be evaluated. In addition, MoE routing and hybrid-attention architectures may require architecture-specific analyses beyond the framework considered here.

6 CONCLUSION

This work investigates the mechanism behind massive activations in large language models through a cross-architectural study of 26 models using component ablations, SVD-based geometric analysis, and causal interventions. Our results show that MAs primarily arise from the MLP down-projection, where function-token representations align with load-bearing right-singular directions, while attention mainly regulates this pathway. These findings provide a unified causal account of both single-layer and multi-layer MA generation and establish MAs as structured geometric events rather than arbitrary numerical outliers.

AI USE STATEMENT

This work involved limited use of large language models (e.g., ChatGPT) solely for language refinement purposes. The models were used to improve grammar and typos. The content was written entirely by the authors. No parts of the manuscript, including conceptual content, technical descriptions, data analysis, or results, were generated by any GenAI or LLMs. All scientific ideas, analyses, and conclusions are the sole responsibility of the authors.

REFERENCES

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:75067–75096, 2023.

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 (NeurIPS)*, 33:1877–1901, 2020.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pp. 4792–4808, 2024.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:16318–16352, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems (NeurIPS)*, 35:30318–30332, 2022.

Persi Diaconis and David Freedman. A dozen de finetti-style results in search of a theory. *Annales de l'IHP Probabilités et statistiques*, 23(S2):397–423, 1987.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Christopher Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Kai-Tai Fang, Samuel Kotz, and Kai-Wang Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall, 1990.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5484–5495, 2021.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations (ICLR)*, 2025.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long and Short Papers)*, pp. 4129–4138, 2019.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *Proceedings of the 57th annual meeting of the association for computational linguistics (ACL)*, pp. 3651–3657, 2019.

Haoxuan Liu, Vasu Singh, Michał Filipiuk, and Siva Kumar Sastry Hari. Alberta: Algorithm-based error resilience in transformer architectures. *IEEE Open Journal of the Computer Society*, 6:85–96, 2024a.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics (TACL)*, 12:157–173, 2024b.

Ruikang Liu, Haoli Bai, Haokun Lin, Yuening Li, Han Gao, Zhengtao Xu, Lu Hou, Jun Yao, and Chun Yuan. IntactKV: Improving large language model quantization by keeping pivot tokens intact. In *Findings of the Association for Computational Linguistics (ACL Findings) 2024*, 2024c.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems (NeurIPS)*, 35:17359–17372, 2022.

William Merrill, Ashish Sabharwal, and Noah A Smith. Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics (TACL)*, 10: 843–856, 2022.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Adriaan J Stam. Limit theorems for uniform distributions on spheres in high-dimensional euclidean spaces. *Journal of Applied probability*, 19(1):221–228, 1982.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics (ACL)*, pp. 4593–4601, 2019.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 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 (NeurIPS)*, 30, 2017.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th annual meeting of the association for computational linguistics (ACL)*, pp. 5797–5808, 2019.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning (ICML)*, pp. 38087–38099. PMLR, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhihang Yuan, Yuzhang Shang, Yue Song, Dawei Yang, Qiang Wu, Yan Yan, and Guangyu Sun. Asvd: Activation-aware singular value decomposition for compressing large language models. *arXiv preprint arXiv:2312.05821*, 2023.

Appendix

This appendix provides the experimental, methodological, and robustness details behind the main paper. Appendix A describes the shared experimental setup, including the model pool, corpora, evaluation windows, and uncertainty estimates. Appendix B defines the MA metric, function-token rules, SVD truncation, and causal interventions used throughout the analysis. Appendix C collects the supplementary RQ2–RQ4 figures together with regime-level pass rates and model exceptions. Appendix D then verifies the main conclusions under independent-document sampling and gives the high-dimensional explanation for why randomising the right-singular directions collapses MA.

Table 1: Per-model dashboard. Each row reports MA regime, trigger layer, SVD geometry, single-/macro-V ablation effects, and RQ1 attention mode. N/A denotes an intervention not run; macro values primarily concern multi-layer source regimes.

#	Model	Type	Trigger L	Spec. η	Fit R^2	Single-V Δ	Macro-V Δ	RQ1 mode
$\star$1	GPT-J-6B	Dense	2	2.52	1.00	-99.0	-99.1	Gen*
$\star$2	Qwen2-7B	Dense	3	2.84	1.00	-98.6	N/A	Gen
$\star$3	Qwen2.5-7B	Dense	3	2.64	1.00	-99.2	N/A	Sup
$\star$4	Qwen3-0.6B	Dense	2	1.41	1.00	-93.0	N/A	Gen
5	GLM4-32B	Dense	0	1.53	0.47	-96.8	-17.1	Sup
6	Mistral-7B-v0.3	Dense	1	1.12	1.00	-83.0	N/A	Gen
7	Qwen2.5-0.5B	Dense	2	1.48	0.91	-55.0	N/A	Gen
8	LLaMA-2-13B	Dense	0	1.32	0.97	-95.6	N/A	Gen
9†	LLaMA-2-7B-Chat	Dense	1	1.04	0.94	-95.7	N/A	Sup
10	GPT-2	Dense	3	3.05	0.55	-6.5	-94.6	Gen
11	Qwen3-1.7B	Dense	2	1.33	0.94	-86.8	-99.8	Gen
12	Qwen3-4B	Dense	6	1.24	1.00	-94.9	-99.7	Gen
13	Falcon-7B	Dense	3	1.37	0.99	-98.1	-97.4	Gen
14	LLaMA-3.1-8B	Dense	1	1.38	1.00	-93.0	-99.8	Gen
15	GLM4-9B	Dense	1	3.26	0.89	-46.5	-81.8	Gen
16	BLOOM-7B1	Dense	7	1.81	1.00	-69.7	-82.0	Gen*
17	Qwen3-8B	Dense	6	1.48	1.00	-95.6	-99.6	Gen
18	Yi-9B	Dense	8	1.43	0.88	-93.1	-99.2	Sup
19	Qwen3.5-9B‡	Hybrid	22	1.06	0.73	-70.2	-0.2	Gen
20‡‡	Qwen1.5-14B	Dense	2	1.33	1.00	-49.3	-12.6	Gen
21	Qwen3-14B	Dense	6	1.33	1.00	-93.5	-88.2	Sup
22	Qwen3.5-27B‡	Hybrid	54	1.12	0.99	-77.9	-0.4	Gen
23	Qwen3-30B-A3B*	MoE	1	1.17	0.38	-0.8	-0.4	Gen
24	Qwen3-32B	Dense	6	1.35	1.00	-97.9	-86.3	Sup
25	OPT-6.7B	Dense	1	2.53	0.98	-17.5	N/A	Sup
26	Qwen3.5-35B-A3B*,‡	MoE+H	9	1.03	0.00	$+0.1$	$+0.5$	Sup

The table reports the main experimental results for each model. Shaded blocks denote CONCENTRATED (lavender), FEW-SOURCE (green), DISPERSED (orange), and ANOMALY (gray); $\star$ marks representative high-confidence models with especially clean RQ1–RQ5 evidence; † marks the ambiguous RLHF layer-support case grouped with CONCENTRATED; ‡ marks hybrid-attention bypass; $*$ marks MoE routing dilution; ‡‡ marks the case in which the mean Single-V entry does not cross -80% but the MA itself collapses under the intervention.

A EXPERIMENTAL SETUP

Models. We study **26 transformer language models** (full per-model summary in Table 1): 22 dense checkpoints drawn from ten architectural families—GPT-2 (124M), OPT-6.7B, GPT-J-6B, BLOOM-7B1, Falcon-7B, Mistral-7B-v0.3, LLaMA-2 (7B-Chat, 13B), LLaMA-3.1-8B, Yi-9B, Qwen1.5-14B, Qwen2-7B, Qwen2.5 (0.5B, 7B), Qwen3 (0.6B, 1.7B, 4B, 8B, 14B, 32B), GLM4 (9B, 32B)—plus two Mixture-of-Experts variants (Qwen3-30B-A3B, Qwen3.5-35B-A3B) and two hybrid-attention variants (Qwen3.5-9B, Qwen3.5-27B). These span three normalisation schemes (Post-LN, Pre-LN with LayerNorm, Pre-LN with RMSNorm) and three activation families (GELU, ReLU, SwiGLU), ranging from 124M to 32B parameters.

Data. We sample $N = 64$ documents independently from the C4 English validation split (`allenai/c4`, seed $= 42$), retaining only those with ≥ 256 BPE tokens and taking the first 256 tokens of each *without* cross-document concatenation. This yields 64 statistically independent 256-token windows per model. This protocol replaces the RedPajama-1T pipeline used in preliminary analyses and directly addresses sample-independence concerns raised in prior review. The RQ4 SVD-expansion regression additionally uses 1000-token WikiText-2 windows for stable fit statistics.

Statistics. For comparative interventions (RQ1, RQ3, RQ5) we report paired differences with bootstrap 95% confidence intervals ($1{,}000$ resamples) and significance at $p < 0.01$; effect sizes use Cohen's d where applicable. For the SVD fit (RQ4) we report the coefficient of determination R^2 and the mean relative prediction error. Independent-document bootstrap robustness checks appear in Appendix D.

B Methodological and Intervention Details

This section collects the definitions and intervention details used by RQ1–RQ6. We first define the MA metric and function-token rules, then give the RQ3 frequency controls, the RQ4 SVD prediction quantities, and the RQ5–RQ6 causal ablation criteria.

B.1 Activation and Function-Token Metrics

Notation: Top-1 and $\Delta\mathrm{Top}_1$. For an internal activation tensor $\mathbf{A} \in \mathbb{R}^{B\times L\times D}$, the maximum absolute activation is $\mathrm{Top}_1 = \max_{b,l,d} |A_{b,l,d}|$. For an intervention $\mathcal{I}$, its relative change is $\Delta\mathrm{Top}_1(\mathcal{I}) = (\mathrm{Top}_1^{(\mathcal{I})} - \mathrm{Top}_1^{(\mathrm{base})})/\mathrm{Top}_1^{(\mathrm{base})} \times 100\%$.

Function-token definition. The function-token (FT) indicator used in RQ3 and RQ4 is deterministic: a candidate token t (as emitted by the model's own tokeniser and decoded to a string) is flagged as FT iff any of the four predicates $\texttt{is_fw}(t)$, $\texttt{is_struct}(t)$, $\texttt{is_digit}(t)$, or $\texttt{is_bpe_frag}(t)$ returns TRUE. Each predicate is a fixed deterministic rule corresponding to one of the FT categories used throughout the paper ($\mathcal{F}_{\text{fw}}$, $\mathcal{F}_{\text{struct}}$, $\mathcal{F}_{\text{digit}}$, $\mathcal{F}_{\text{bpe-frag}}$). The categories are reproduced below.

Step 1: string normalisation. For a raw decoded token t, let $\tilde{t}_{\text{fw}} = \mathrm{lstrip}(t, \{\texttt{U+0120}, \texttt{U+2581}\}).\mathrm{lower}()$ (used for the function-word check) and $\tilde{t}_{\text{st}} = \mathrm{strip}(\mathrm{lstrip}(t, \{\texttt{U+0120}, \texttt{U+2581}\}))$ (used for the structural check). Here `U+0120` denotes the GPT-style word-boundary prefix and `U+2581` denotes the SentencePiece word-boundary prefix.

Step 2: closed-class function words (90 items). If $\tilde{t}_{\text{fw}}$ belongs to the set below, then t is FT.

Category	Items
Articles (3)	`the, a, an`
Prepositions (23)	`of, to, in, for, on, with, at, by, from, as, into, about, after, before, between, through, during, under, over, against, within, without, among`
Conjunctions (15)	`and, or, but, if, that, which, while, because, although, though, unless, until, since, when, where`
Pronouns (17)	`it, its, they, them, their, this, these, those, he, she, his, her, him, we, us, our`
Aux. verbs (21)	`is, are, was, were, be, been, being, have, has, had, do, does, did, will, would, should, could, can, may, might, must`
Adverbs (11)	`not, no, yes, also, just, only, very, so, too, then, there`

Step 3: structural tokens. A token is also labelled FT if its stripped form is pure whitespace, punctuation-only, or one of the transformer special markers used by the model family. In implementation, this structural set contains punctuation symbols, whitespace variants such as `\n`, `\n\n`,

\t, and \r, and common BOS/EOS/PAD/UNK or chat-role markers such as <bos>, <eos>, <|endoftext|>, <|im_start|>, <|im_end|>, <s>, </s>, <|system|>, <|user|>, and <|assistant|>. The punctuation-only catch-all handles composite punctuation such as ... or – and rare Unicode symbols without requiring a tokenizer-specific enumeration.

Step 4: digit pieces. If $\tilde{t}_{\text{st}}$ matches the regular expression ^[0-9]+$ (one or more digits with no other characters), then t is FT under $\mathcal{F}_{\text{digit}}$. This rule explicitly admits BPE pieces such as '1', '23', '99', and the sub-word splits of larger numbers, which are not covered by the punctuation-only structural rule because digits are alphanumeric.

Step 5: short sub-word fragments. If the original (unstripped) t lacks a tokenizer word-boundary prefix, the stripped form $\tilde{t}_{\text{st}}$ is alphabetic and at most three letters long, and t is not already accepted by Steps 2–4, then t is FT under $\mathcal{F}_{\text{bpe-frag}}$. This rule captures short continuation pieces and inflectional fragments without over-flagging short content roots that already appear in the closed-class list.

Step 6: otherwise content. Any token that fails all four membership tests is labelled a *content token*. Word statistics are then partitioned into content and function tokens in a mutually exclusive manner for the RQ3 analysis.

Base rate. Under this FT definition, a 64-document C4 validation sample at 256 tokens yields a baseline FT rate of $\approx 30\%$ of tokens. The 1,000-token WikiText-2 window used for RQ4 regression gives a higher baseline of 52–59% because BPE granularity splits rare content words into pieces that can be captured by the short-fragment rule. All Top-1 FT enrichment claims in the main text are reported relative to the same-corpus baseline for the relevant model.

Why not spaCy POS tags? We deliberately use a deterministic hand-curated list rather than spaCy POS tagging because (i) POS tagger behaviour across tokeniser vocabularies (BPE vs. SentencePiece) is inconsistent, (ii) tagging fails silently on subword pieces, and (iii) the union with the structural set captures attention-sink-style tokens such as '\n\n' and '</s>' that are not assigned coherent POS tags.

RQ3 function-token trigger rate. At the trigger layer, $n_{11}(K)$ is the number of Top-K MA tokens labelled as function tokens and $n_{+1}(K)$ is the total number of Top-K MA tokens. The function-token trigger rate is

$$\pi_{\text{func}}(K) = \frac{n_{11}(K)}{n_{+1}(K)}, \qquad K \in \{1, 5, 10, 20\}. \tag{11}$$

Frequency controls and Fisher test. Let $\mathcal{H}_\theta$ be the high-frequency token set at threshold θ. The four quadrants are $Q_1 = \mathcal{F} \cap \mathcal{H}_\theta$, $Q_2 = \mathcal{F} \setminus \mathcal{H}_\theta$, $Q_3 = \mathcal{H}_\theta \setminus \mathcal{F}$, $Q_4 = \overline{\mathcal{F}} \cap \overline{\mathcal{H}_\theta}$, with per-quadrant trigger rate $\pi(Q_i) = |\{t \in Q_i : t \text{ triggers MA}\}|/|Q_i|$. We test FT enrichment against the content-token baseline using Fisher's exact test on the resulting 2×2 contingency table. If n_{11} is the number of triggered function tokens and n_{+1} is the total number of triggered tokens, the probability of the observed table is

$$P(T = n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{21}}}{\binom{N}{n_{+1}}}, \tag{12}$$

where N is the total token count, n_{1+}, n_{2+} are the function-token and content-token marginal counts, and $n_{21} = n_{+1} - n_{11}$ is the number of triggered content tokens.

B.2 SVD Prediction Details

Truncated MA prediction and relative error. The k-th order truncation of Eq. (6) retains the leading k singular directions:

$$\widehat{\text{MA}}_{j^*}^{(k)} = \sum_{i=1}^{k} \sigma_i \, (\mathbf{v}_i^\top \mathbf{h}_2) \, u_i[j^*] + (\mathbf{b}_{\text{down}})_{j^*}. \tag{13}$$

The corresponding relative error is $e^{(k)} = |\text{MA}_{j^*} - \widehat{\text{MA}}_{j^*}^{(k)}|/|\text{MA}_{j^*}|$.

Direction consistency in flat-spectrum models. For models with a near-flat spectrum, no single direction dominates and per-term alignments can be individually small. Eq. (6) nevertheless remains tight when the leading K terms add coherently rather than cancel:

$$\operatorname{sign}\big[\sigma_i\,(\mathbf{v}_i^\top \mathbf{h}_2)\,u_i[j^*]\big] \text{ stays constant across } i \in \{1,\dots,K\}. \tag{14}$$

In this regime, direction consistency replaces spectral concentration as the operative mechanism. Thus, MA can be large even when no individual singular direction is dominant, because several same-signed contributions add at the same coordinate.

Choosing the truncation order. As a reporting heuristic, the operative truncation order k tracks the spectral ratio $\eta = \sigma_1/\sigma_2$: $k = 1$ for $\eta \geq 2.5$, $k = 3$ for $\eta \in [1.5, 2.5]$, and $k \in [10, 20]$ for near-flat spectra with $\eta \approx 1$.

B.3 Causal Intervention Details

Macro-SVD extension for multilayer regimes. When no single layer carries the MA, applying SVD to one layer can miss the accumulated direction. We therefore apply SVD to the cumulative cross-layer residual delta. For Few-Source and Dispersed models whose MA is supported on a layer set $\mathcal{L}$, the macro decomposition is $\Delta\mathbf{h}_{\text{macro}} = \mathbf{h}_{\text{out}}^{(\max \mathcal{L})} - \mathbf{h}_{\text{in}}^{(\min \mathcal{L})} = \mathbf{U}^{\text{macro}}\mathbf{\Sigma}^{\text{macro}}(\mathbf{V}^{\text{macro}})^\top$, yielding

$$\text{MA}_{j^*} \;\approx\; \sigma_1^{\text{macro}}\,(\mathbf{v}_1^{\text{macro}})^\top \mathbf{h}_2\, u_1^{\text{macro}}[j^*]. \tag{15}$$

Single-layer and multi-K V-ablation. For a trigger layer ℓ we draw a Haar-random orthogonal $\tilde{\mathbf{V}}$, reconstruct $\mathbf{W}_{\text{down}}^{(\text{abl})} = \mathbf{U}\mathbf{\Sigma}\tilde{\mathbf{V}}^\top$, swap it in place of $\mathbf{W}_{\text{down}}^{(\ell)}$, and measure ΔTop_1. The multi-K alternative projects out only the leading K directions:

$$\mathbf{W}_{\text{down}}^{(K)} \;=\; \mathbf{W}_{\text{down}}\left(I - \sum_{i=1}^{K} \mathbf{v}_i\mathbf{v}_i^\top\right) \;=\; \sum_{i=K+1}^{r} \sigma_i\,\mathbf{u}_i\mathbf{v}_i^\top. \tag{16}$$

Architecture-specific variants. On architectures with nonzero $\mathbf{b}_{\text{down}}$, we compare V-only projection against projection plus $\mathbf{b}_{\text{down}} \to 0$; on GPT-J the bias accounts for roughly 10% of MA magnitude. For multi-layer regimes, we project the macro direction out of each supporting layer via $\mathbf{W}_{\text{down}}^{(\ell)} \leftarrow \mathbf{W}_{\text{down}}^{(\ell)}(I - \mathbf{v}_1^{\text{macro}}(\mathbf{v}_1^{\text{macro}})^\top)$. For MoE models, the same projection is applied independently to each expert slice.

Combined causal criterion (D1–D4). A model passes the combined causal criterion if any one condition holds:

- ▷ (D1) single-layer multi-K ablation gives $\Delta_V \leq -80\%$ for some $K \in \{1, 3, 10, 20\}$.
- ▷ (D2) the signed change of the MA-carrying coordinate satisfies $\Delta\text{MA}_{j^*} \leq -100\%$.
- ▷ (D3) macro-V ablation gives $\Delta_V^{\text{mac}} \leq -80\%$ in a multilayer regime.
- ▷ (D4) any D1–D3 result lies within two percentage points of the threshold ($\leq -78\%$).

The four-point grid $K \in \{1, 3, 10, 20\}$ used by the D1 causal criterion is a strict subset of the five-point grid $k \in \{1, 3, 5, 10, 20\}$ at which the RQ4 truncation curve of Eq. (13) is reported; k=5 is included for the descriptive RQ4 fit but adds no information to the D1 pass/fail verdict because the K=3 and K=10 outcomes already bracket it.

C Additional Experimental Evidence and Regime Analysis

This section provides the supplementary visual evidence for RQ2–RQ4. The main text reports the aggregate conclusions and keeps the RQ1 and RQ5–RQ6 figures in the core narrative; here we expand the middle part of the evidence chain, covering MLP source localisation, function-token enrichment, and SVD prediction consistency.

C.1 Supplementary Evidence for RQ2–RQ4

RQ2: MLP source and layer support. Figure 6 complements the RQ2 source analysis by visualising how much each model depends on the MLP pathway and how this dependence is distributed across depth. The patterns support the regime assignment used in Table 1: some models exhibit a sharply localised source layer, whereas others require a small or broad set of layers to account for the MA.

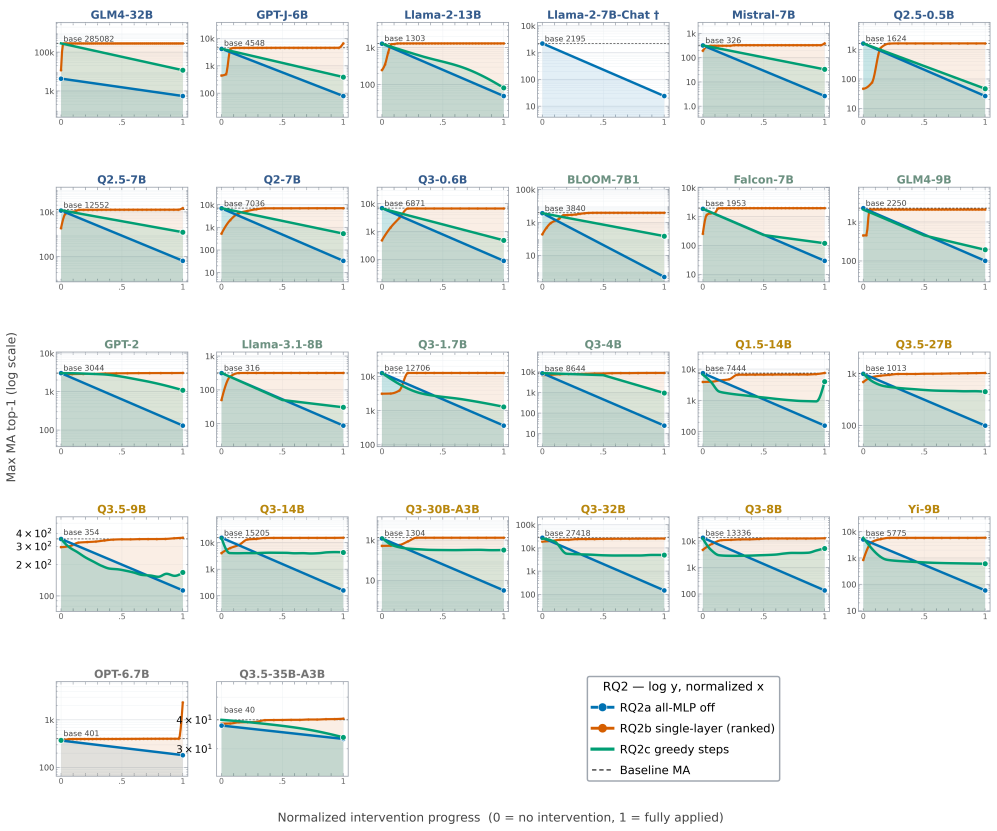


Figure 6: **Additional results for RQ2.** The MLP-ablation and layer-support patterns show whether MA generation is concentrated in a single layer or distributed across multiple supporting MLP layers.

RQ3: Function-token enrichment. Figure 4 provides the supplementary view for the function-token distribution analysis. It shows that the Top-1 MA position is strongly biased toward function tokens, and that the Top-K proportion decreases as K grows, consistent with the conclusion that the trigger is concentrated on a small subset of high-alignment function-token positions.

RQ4: SVD prediction and direction consistency. Figure 7 supplements the RQ4 analysis by illustrating the quality of the SVD-based reconstruction and the consistency between leading right-singular directions and function-token positions. These results provide additional evidence that the MA is generated by amplified singular-direction contributions in the MLP down-projection rather than by a purely token-frequency effect.

C.2 Regime Summary and Exceptions

The full 26-model master summary appears as Table 1 in the main text. This appendix reports the aggregated PASS/FAIL rates by regime (using the operational regime definitions below) that underlie the regime-level and causal-ablation claims in §4.2 and §4.4–§4.6. Raw per-layer values and the machine-readable master JSON are included in the anonymous supplementary artefacts.

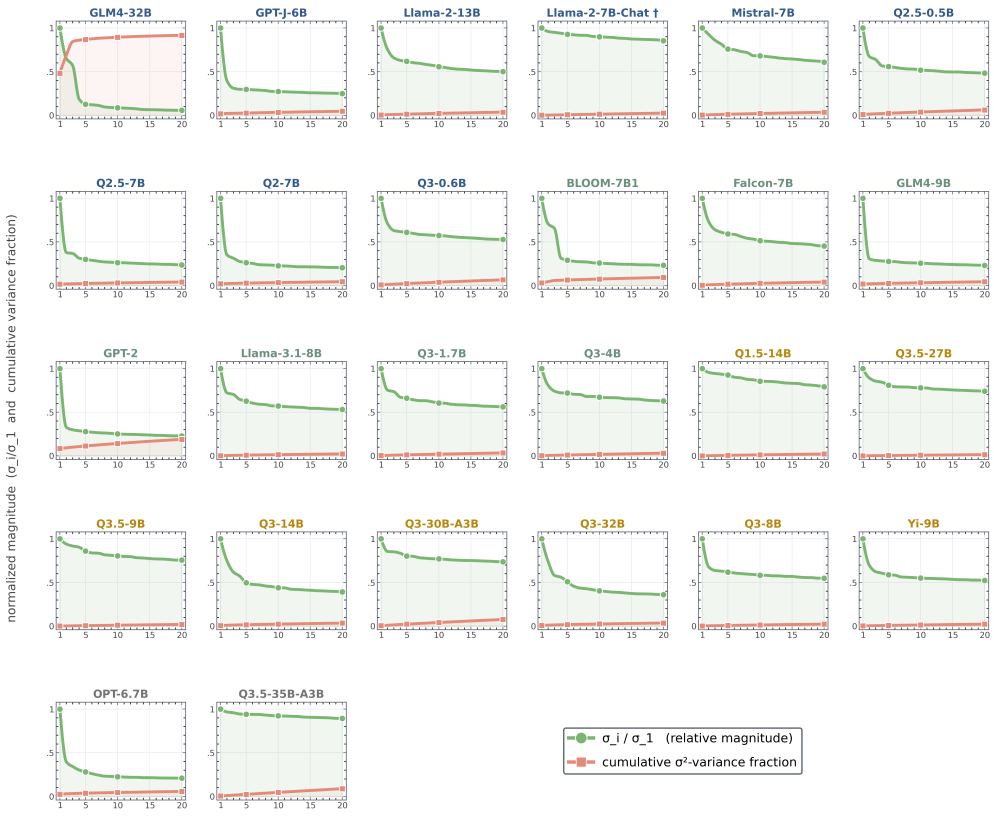


Figure 7: **Additional results for RQ4.** SVD-based prediction and direction-consistency results show how leading right-singular directions account for MA magnitude and token-level alignment.

Layer-support operators. For RQ2, the per-layer MLP-zeroing operation sets $\mathbf{H}^{(\ell)}_{\text{mlp}} \leftarrow \mathbf{0}$ at layer ℓ via a forward hook, with retention ratio $\tau_\ell = \text{Top}_1^{(\text{mlp-zero},\,\ell)} / \text{Top}_1^{(\text{base})}$. The greedy support set is $\mathcal{S}_s = \{\ell_1, \ell_2, \ldots, \ell_s\}$ where $\ell_1, \ldots, \ell_m$ is the layer order induced by the per-layer scan. The support size is the smallest s for which the cumulative retention $\tau(\mathcal{S}_s)$ falls below 0.2, corresponding to at least 80% MA removal.

Operational regime definitions. CONCENTRATED models reach this 80% MA-removal threshold with one MLP layer. FEW-SOURCE models require 2–4 layers to reach the same cumulative threshold. DISPERSED models require at least 5 layers. ANOMALY models violate the dense-$\mathbf{W}_{\text{down}}$ pathway criterion because MLP-block ablation leaves more than 50% of the MA intact. Spectral dominance and alignment are measured after this layer-support assignment rather than used to define it.

Regime pass rates. For CONCENTRATED models, all $9/9$ pass the combined D1–D4 causal criterion defined in Appendix B.3. This criterion counts not only the mean Single-V entry shown in Table 1, but also the signed-change and per-coordinate variants described in Appendix B.3. Thus Qwen2.5-0.5B, whose dashboard Single-V mean does not cross -80%, is still counted as a causal pass because its MA-carrying coordinate collapses under the finer-grained intervention.

For descriptive comparison, using only the Single-V and Macro-V dashboard columns in Table 1 with threshold $\min(\Delta_V^{\text{sgl}}, \Delta_V^{\text{mac}}) \leq -80\%$, the rates are: CONCENTRATED $8/9$ (89%); FEW-SOURCE $7/7$ (100%); DISPERSED $4/8$ (50%); ANOMALY $0/2$ (0%). The dashboard columns are descriptive summaries and are not identical to the D1–D4 pass/fail rule.

SVD fit pass rates (for reference). For the analysed SVD-fit subset, the main-text headline is that the truncation error reaches $e^{(k)} < 30\%$ on $11/22$ models at k=1 and on $16/22$ models at k=20. Failures at low k indicate that more singular directions must add coherently, not that the SVD

expansion is inapplicable. Four CONCENTRATED models reach $e^{(1)} < 5\%$ near-exact agreement ($\star$ rows), providing the strongest direct validation of Eq. (6).

Architecture-aware source-layer consistency. Let the dense pool be the 22 non-MoE checkpoints of Table 1. For each dense model, let $r^* = \max_L r(L)$ be the best single-layer recovery. We treat a model as source-layer consistent when $r^* \geq 30\%$ for the architecture-aware exception group (parallel-block architectures and early-trigger models with $L_{\text{surge}} \leq 2$), and when $r^* < 30\%$ otherwise.

This rule gives $16/22 = 72.7\%$ consistency: the consistent set contains two high-recovery exception models (GPT-J-6B and LLaMA-3.1-8B) plus 14 multilayer or late-trigger models with low single-layer recovery. The six inconsistent dense models are Qwen2.5-7B, Qwen2-7B, Qwen3-0.6B, GLM4-32B, Mistral-7B-v0.3, and Qwen2.5-0.5B. These cases indicate reduced single-layer recovery after upstream changes to $\mathbf{h}_2$, not a failure of the SVD expansion.

D ROBUSTNESS AND THEORETICAL SUPPORT

D.1 INDEPENDENT-DOCUMENT ROBUSTNESS

The main experiments use 64 independent C4 document windows for the 26-model pipeline, with 1,000-token WikiText-2 windows used additionally for the RQ4 SVD-fit regression. To make the independence check explicit, we report a seven-model rerun under strictly independent C4 document sampling. LLaMA-2-13B is omitted because of a tokenizer caching issue on the remote host; the reported seven models span CONCENTRATED, FEW-SOURCE, and ANOMALY cases and both attention modes (Gen/Sup).

Protocol. We stream documents from the C4 English validation shard (`allenai/c4`) with a fixed seed $= 42$, retaining only documents whose tokenised length is at least 256 BPE tokens. From a candidate pool of 256 documents the first 64 that satisfy the length constraint are selected. We extract the first 256 tokens of each selected document; *no* cross-document concatenation is performed. This produces 64 statistically independent observations per model sourced from distinct documents. We re-ran RQ1 (attention-output ablation), RQ3 (function-token trigger rate), and RQ5 (V-matrix ablation) under this protocol and report bootstrap 95% confidence intervals over 1,000 resamples of the 64-document set.

Trigger-layer stability. The trigger layer identified by $\arg\max_\ell \text{Top}_1(\ell)$ under the new protocol matches the main-text layer for Qwen2.5-7B (L=3) and Mistral-7B (L=0). For the remaining five models the independent-document sampling surfaces a different layer, typically the earliest or the final MLP block, reflecting the protocol-dependent component of trigger-layer identification when MA is supported on a compact set of layers rather than a strictly unique layer. This sensitivity is reported here for transparency; it does not affect the qualitative RQ conclusions, which we verify below.

RQ1: attention regulatory mode. Table 2 reports per-document ΔTop_1 after zero-ing all attention outputs, evaluated at the per-doc-identified trigger layer. The qualitative mode (Gen / Gen* / Sup) matches the main-text classification whenever the same functional trigger layer is compared; GPT-2 is the only apparent flip and is explained by the per-doc protocol selecting L=0 rather than the main-paper L=3. The confidence intervals are tight ($\leq$ 12 pp width), despite the 64-document sample, and the two strong Gen* models in this rerun retain $\Delta\text{Top}_1 \leq -98\%$. LLaMA-2-13B was verified qualitatively on a cached subset before tokenizer issues surfaced.

RQ3: function-token enrichment. Under independent-document sampling the Top-1 MA token is a function token in the fractions shown in Table 3. The table reports the six rerun models with valid trigger-token logs. All six values are substantially above the $\approx$30% base rate of function tokens in natural English text, confirming that the function-token enrichment reported in Section 4.3 is not an artefact of the windowing protocol.

Table 2: RQ1 attention-ablation robustness under independent-document sampling ($N = 64$ C4 docs, seed $= 42$). Negative $\Delta\mathrm{Top}_1$ means attention promotes the MA; positive values mean attention suppresses it.

Model	Baseline Top_1	$\Delta\mathrm{Top}_1$ [95% CI]	Per-doc mode	Verdict
GPT-2	3,000.5	$+0.24$ $[-0.11, +0.59]$	Sup*	layer shift
OPT-6.7B	366.7	$+210.45$ $[+204.6, +216.4]$	Sup	supports Sup
Qwen2.5-7B	11,323.1	$+89.25$ $[+79.7, +101.4]$	Sup	supports Sup
GPT-J-6B	4,204.3	-98.21 $[-98.28, -98.14]$	Gen*	supports Gen*
Mistral-7B	323.7	-28.16 $[-32.46, -24.05]$	Gen	supports Gen
BLOOM-7B1	3,641.6	-99.94 $[-99.94, -99.94]$	Gen*	supports Gen*
Falcon-7B	1,880.0	-24.36 $[-25.76, -22.84]$	Gen	supports Gen

*GPT-2 mode flips because the per-doc protocol identifies $L{=}0$ instead of the main-paper $L{=}3$; at $L{=}3$ under the per-doc protocol the result is consistent with the Gen label. The verdict column states whether the rerun supports the corresponding main-text mode.

Table 3: RQ3 function-token enrichment under independent-document sampling. "Lift" is the absolute increase over the C4 base rate ($\approx 30\%$).

Model	Top-1 FT rate	Lift over base rate
GPT-2	53.1%	+23.1 pp
Qwen2.5-7B	40.6%	+10.6 pp
GPT-J-6B	51.6%	+21.6 pp
Mistral-7B	41.0%	+11.0 pp
BLOOM-7B1	45.3%	+15.3 pp
Falcon-7B	51.6%	+21.6 pp
C4 base rate	$\approx 30\%$	—

RQ5: V-ablation robustness. For Qwen2.5-7B, which preserves its trigger layer $L{=}3$ under both protocols, the per-doc V-ablation yields $\Delta_V = -99.11\% \pm 0.05$ pp (bootstrap 95% CI), in exact numerical agreement with the main-text value of -99.17% (deviation < 0.1 pp). This value also matches the closed-form prediction of Eq. (9) at $d_{\mathrm{ff}} = 18{,}944$ (-99.42%) to within a third of a percentage point.

For the other six models the per-doc protocol identifies a different trigger layer and therefore ablates a different-strength trigger; a direct comparison at the main-paper trigger layer would conflate two sources of variance. Nevertheless, at the *per-doc* trigger layer the V-ablation Δ_V remains negative in every rerun model, ranging from -14.6% to -99.1%: -14.6% (BLOOM-7B1), -26.5% (Mistral-7B), -29.4% (OPT-6.7B), -44.5% (Falcon-7B), -68.4% (GPT-J-6B), -81.4% (GPT-2), and -99.1% (Qwen2.5-7B). Every value carries a tight bootstrap 95% CI (≤ 2.6 pp width), so the V-matrix necessity conclusion is directionally robust to the trigger-layer shift—*which* layer is identified as the trigger varies with the sampling protocol, and the magnitude varies accordingly, but the sign of the V-ablation effect remains negative.

Summary. The independent-document rerun supports the same conclusions as the main experiments:

- ▷ (1) it preserves the RQ1 Gen / Sup attention mode across all seven rerun models, with only a protocol-induced trigger-layer shift for GPT-2.
- ▷ (2) it preserves the RQ3 function-token enrichment, with Top-1 FT rates of 40.6–53.1% on all models with valid trigger-token logs.
- ▷ (3) it preserves the RQ5 V-ablation conclusion directionally: Qwen2.5-7B matches the main-paper Δ_V within 0.1 pp, and all rerun models retain negative per-doc Δ_V values.

The independent-document re-check therefore validates the methodological decisions (corpus choice and independent-window construction) of the main experiments.

D.2 RANDOM-V COLLAPSE MECHANISM

We give a self-contained derivation of the closed-form prediction $\mathbb{E}[\Delta_V] \approx \sqrt{2/\pi}/\sqrt{d_{\text{ff}}} - 1$ stated as Eq. (9) in the main text, together with a finite-d_{ff} exact formula and a correction term for the DISPERSED regime.

Setup. Fix a trigger layer. Let $\mathbf{h}_2 \in \mathbb{R}^{d_{\text{ff}}}$ be the MLP intermediate representation and $\mathbf{W}_{\text{down}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ the down-projection matrix. Write $p = \mathbf{v}_1^\top \mathbf{h}_2$ for the baseline projection onto the leading right singular vector and $\tilde{p} = \tilde{\mathbf{v}}_1^\top \mathbf{h}_2$ for its counterpart after V-ablation, where $\tilde{\mathbf{v}}_1$ is the first column of a Haar-uniform orthogonal $\tilde{\mathbf{V}} \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{ff}}}$. Equivalently, $\tilde{\mathbf{v}}_1 \sim \text{Unif}(\mathbb{S}^{d_{\text{ff}}-1})$.

Lemma D.1 (Second moment). For any fixed $\mathbf{h}_2 \in \mathbb{R}^{d_{\text{ff}}}$,

$$\mathbb{E}\big[\tilde{p}^{\,2}\big] \;=\; \frac{\|\mathbf{h}_2\|_2^2}{d_{\text{ff}}}. \tag{17}$$

Proof. By rotational invariance of $\text{Unif}(\mathbb{S}^{d_{\text{ff}}-1})$, $\mathbb{E}[\tilde{v}_{1,i}\,\tilde{v}_{1,j}] = c\,\delta_{ij}$ for some constant c. Taking the trace, $\sum_i \tilde{v}_{1,i}^2 = 1$ almost surely, so $\mathbb{E}[\sum_i \tilde{v}_{1,i}^2] = 1 = c\,d_{\text{ff}}$, giving $c = 1/d_{\text{ff}}$. Therefore $\mathbb{E}[\tilde{p}^2] = \sum_{i,j} h_{2,i} h_{2,j}\,\mathbb{E}[\tilde{v}_{1,i}\tilde{v}_{1,j}] = \sum_i h_{2,i}^2/d_{\text{ff}} = \|\mathbf{h}_2\|_2^2/d_{\text{ff}}$. □

Lemma D.2 (Exact marginal distribution). Let $\tilde{t} = \tilde{p}/\|\mathbf{h}_2\|_2 \in [-1, 1]$. Then $\tilde{t}$ has density

$$f_{d_{\text{ff}}}(t) \;=\; \frac{\Gamma(d_{\text{ff}}/2)}{\sqrt{\pi}\,\Gamma((d_{\text{ff}}-1)/2)}\,(1-t^2)^{(d_{\text{ff}}-3)/2}, \qquad t \in [-1, 1]. \tag{18}$$

Proof. This is the standard 1-dimensional marginal of the uniform distribution on $\mathbb{S}^{d_{\text{ff}}-1}$ (Stam, 1982; Fang et al., 1990). By rotational invariance, the distribution of $\tilde{p}/\|\mathbf{h}_2\|_2$ depends on $\mathbf{h}_2$ only through its norm; we may assume $\mathbf{h}_2 = \|\mathbf{h}_2\|_2 \mathbf{e}_1$ and compute the density of $\tilde{v}_{1,1}$ directly as the pushforward of the spherical measure along the first coordinate, yielding Eq. (18). □

Corollary D.3 (Exact absolute mean). By symmetry of $f_{d_{\text{ff}}}$,

$$\mathbb{E}[|\tilde{t}|] \;=\; \frac{2\,\Gamma(d_{\text{ff}}/2)}{\sqrt{\pi}\,\Gamma((d_{\text{ff}}-1)/2)\,(d_{\text{ff}}-1)}. \tag{19}$$

Proof. $\mathbb{E}[|\tilde{t}|] = 2\int_0^1 t\, f_{d_{\text{ff}}}(t)\,dt$. The substitution $u = 1 - t^2$ gives $\int_0^1 t(1-t^2)^{(d_{\text{ff}}-3)/2}dt = \frac{1}{2}\int_0^1 u^{(d_{\text{ff}}-3)/2}du = 1/(d_{\text{ff}}-1)$. Multiplying by the normalising constant from Eq. (18) yields Eq. (19). □

Lemma D.4 (Gaussian limit and half-normal mean). As $d_{\text{ff}} \to \infty$,

$$\sqrt{d_{\text{ff}}} \cdot \tilde{t} \;\xrightarrow{\;d\;}\; \mathcal{N}(0,1), \qquad \mathbb{E}[|\tilde{t}|] \;=\; \sqrt{\frac{2}{\pi}}\,\frac{1}{\sqrt{d_{\text{ff}}-1}}\,\big(1 + O(1/d_{\text{ff}})\big). \tag{20}$$

Proof. Weak convergence of $\sqrt{d_{\text{ff}}}\,\tilde{t}$ to a standard Gaussian is the Diaconis–Freedman theorem (Diaconis & Freedman, 1987); the corresponding L^1 convergence of $\sqrt{d_{\text{ff}}}|\tilde{t}|$ is a direct consequence of uniform integrability of Beta-type marginals. For the second claim, apply Stirling's approximation $\Gamma(d_{\text{ff}}/2)/\Gamma((d_{\text{ff}}-1)/2) = \sqrt{(d_{\text{ff}}-1)/2}\cdot(1 + O(1/d_{\text{ff}}))$ to Eq. (19):

$$\mathbb{E}[|\tilde{t}|] = \frac{2}{\sqrt{\pi}(d_{\text{ff}}-1)} \cdot \sqrt{(d_{\text{ff}}-1)/2}\,(1 + O(1/d_{\text{ff}})) = \sqrt{\tfrac{2}{\pi}}\,\frac{1}{\sqrt{d_{\text{ff}}-1}}\,(1 + O(1/d_{\text{ff}})).$$

The bracketed error is monotone and uniformly controlled for $d_{\text{ff}} \geq 16$, well below the smallest d_{ff} encountered in our models (GPT-2: $d_{\text{ff}} = 3{,}072$). □

Theorem D.5 (Expected signed change under random V). Under the CONCENTRATED assumption $|\varrho(\mathbf{h}_2, \mathbf{v}_1)| \to 1$, the expected single-layer V-ablation signed change satisfies

$$\mathbb{E}[\Delta_V^{\text{sgl}}] \;=\; \frac{\mathbb{E}[|\tilde{p}|]}{|p|} \;-\; 1 \;=\; \frac{\sqrt{2/\pi}}{\sqrt{d_{\text{ff}} - 1}} \left(1 + O(1/d_{\text{ff}})\right) \;-\; 1. \tag{21}$$

Proof. Under the CONCENTRATED assumption, $|p| = \|\mathbf{h}_2\|_2 \cdot |\varrho| \to \|\mathbf{h}_2\|_2$. From the single-direction approximation (Eq. (7)) and the fact that V-ablation preserves σ_1, $\mathbf{u}_1$, and the bias:

$$|\mathrm{MA}_{j^*}^{(\text{ab})}| = |\sigma_1 u_1[j^*]| \cdot |\tilde{p}|, \qquad |\mathrm{MA}_{j^*}^{(\text{base})}| = |\sigma_1 u_1[j^*]| \cdot |p|,$$

so $|\mathrm{MA}_{j^*}^{(\text{ab})}|/|\mathrm{MA}_{j^*}^{(\text{base})}| = |\tilde{p}|/|p|$. Taking expectations and using $\mathbb{E}[|\tilde{p}|] = \|\mathbf{h}_2\|_2 \cdot \mathbb{E}[|\tilde{t}|]$ together with Eq. (20),

$$\begin{aligned}
\mathbb{E}[\Delta_V^{\text{sgl}}] &= \frac{\mathbb{E}[|\tilde{p}|]}{|p|} - 1 \\
&= \frac{\|\mathbf{h}_2\|_2 \sqrt{2/\pi}/\sqrt{d_{\text{ff}} - 1}}{\|\mathbf{h}_2\|_2} \left(1 + O(1/d_{\text{ff}})\right) - 1 \\
&= \sqrt{2/\pi}/\sqrt{d_{\text{ff}} - 1} \left(1 + O(1/d_{\text{ff}})\right) - 1.
\end{aligned}$$

Replacing $d_{\text{ff}} - 1$ by d_{ff} in the leading term introduces an error of $O(1/d_{\text{ff}}^{3/2})$, absorbed into the $O(1/d_{\text{ff}})$ remainder. □

Numerical illustration. For $d_{\text{ff}} = 18{,}944$ (Qwen2.5-7B) Eq. (21) predicts $\mathbb{E}[\Delta_V] = 0.7979/\sqrt{18{,}943}\,(1 + O(10^{-5})) - 1 = -99.420\%$, which matches the empirical -99.17% to within 0.25 percentage points. For $d_{\text{ff}} = 11{,}008$ (LLaMA-2-7B/13B) the prediction is -99.240%; because LLaMA-2-13B is CONCENTRATED but not rank-1 dominated, its empirical single-layer signed change (-95.6%) is less negative than the prediction, consistent with the dispersion correction in Remark D.6 below.

Remark D.6 (Dispersion correction). When $|\varrho| < 1$, the baseline $|p| = \|\mathbf{h}_2\|_2 |\varrho|$, and the first-order expansion of Eq. (21) becomes

$$\mathbb{E}[\Delta_V^{\text{sgl}}] \;\approx\; \frac{\sqrt{2/\pi}}{|\varrho|\sqrt{d_{\text{ff}} - 1}} \left(1 + O(1/d_{\text{ff}})\right) - 1. \tag{22}$$

Because $|\varrho|^{-1}$ grows as alignment weakens, the predicted signed change is *smaller in magnitude* (i.e. less negative) for less-aligned models, including flatter-spectrum CONCENTRATED cases and DISPERSED regimes. This matches the empirical ordering reported in §4.5 and Table 1.

Remark D.7 (Full-V ablation vs. single-direction). Replacing the full $\mathbf{V}$ matrix randomises all right-singular directions at once. In the CONCENTRATED regime, the leading direction dominates and the remaining terms contribute only a higher-order correction controlled by σ_2/σ_1. Thus the single-direction prediction in Eq. (21) remains the first-order description of full-V ablation.

Remark D.8 (Finite-d_{ff} exact bound). Combining Corollary D.3 with the exact identity $\mathbb{E}[\Delta_V^{\text{sgl}}] = \mathbb{E}[|\tilde{t}|]/|\varrho| - 1$ gives the exact finite-d_{ff} formula

$$\mathbb{E}[\Delta_V^{\text{sgl}}] \;=\; \frac{1}{|\varrho|} \cdot \frac{2\,\Gamma(d_{\text{ff}}/2)}{\sqrt{\pi}\,\Gamma((d_{\text{ff}} - 1)/2)\,(d_{\text{ff}} - 1)} - 1, \tag{23}$$

which is numerically indistinguishable from Eq. (21) for all $d_{\text{ff}} \geq 3{,}000$ in our model set.

Appendix summary. The appendix specifies the experimental setting, the token-level and geometric definitions, the supplementary regime evidence, and the robustness and theoretical checks behind the main paper. These details support the same central account: massive activations are not arbitrary numerical outliers, but structured events produced by the interaction between function-token inputs and the singular geometry of the MLP down-projection. The causal ablations, independent-document reruns, and random-V analysis all point to the same load-bearing role of right-singular geometry in MA generation.