

Zipf Made Me Do It: A Causal Account of Outlier Dimensions

Kartik Sharma^{*1} Shreshth Rai^{*1} Nikunj Gupta^{*1} Dinesh Kumar Vishwakarma¹

¹Artificial Intelligence and Machine Learning Society

Delhi Technological University

{kartiksharma_ep24b07_049, shreshthrai_mc24b06_039, nikunjgupta_ee24a17_055, dinesh}@dtu.ac.in

Abstract

Outlier dimensions are a recurring geometric property of transformer representations that have been linked to token frequency in pre-trained language models. However, studies on pretrained models cannot disentangle token frequency from semantic and optimization confounds, preventing a causal understanding of whether frequency itself drives outlier formation. These distortions also degrade probing accuracy and hinder model compression, motivating a controlled study of their origin. We address this issue by isolating frequency skew as a controlled variable in a synthetic transformer trained on data with a tunable Zipfian token distribution. We find that outlier magnitude remains stable under low skew but increases sharply beyond a moderate threshold. A frozen frequency-based output bias fails to reproduce this behavior, model capacity exhibits a non-monotonic U-shaped effect, and an L_2 -normalized cosine softmax suppresses outlier geometry only under a large temperature not reached by standard training. These results identify frequency skew as a causal driver of outlier dimensions and show that removing their geometry requires an intervention beyond standard training.

1 Introduction

Transformer language models (Vaswani et al., 2017) frequently develop *outlier dimensions* (ODs): a small number of hidden-state coordinates whose activations are anomalously large relative to every other coordinate, consistently across inputs (Koval-eva et al., 2021; Timkey and van Schijndel, 2021). ODs are not a benign curiosity; they distort probing and similarity measures computed on raw hidden states (Timkey and van Schijndel, 2021), and they are a primary obstacle to low-bit post-training quantization (Bondarenko et al., 2023; Sun et al., 2024). Prior work links ODs to token frequency,

but only correlationally, in pretrained models where frequency cannot be isolated from other confounds (Section 2). This leaves open whether skew *causes* outlier formation, and whether a specific mechanism - the *softmax bottleneck* (Yang et al., 2018) explains why.

We test this account on small transformers trained on synthetic, Zipfian-skewed data (α -controlled) with a context-conditional structure that preserves the marginal distribution while inducing genuine higher-order frequency dependencies (Section 3), following precedent for controlled synthetic corpora (Chan et al., 2022). Unlike deployment-scale pretrained models, this setup lets us set α exactly and keeps systematic ablations tractable across multiple seeds. We evaluate six output heads, including unnormalized, frequency-biased, and cosine-softmax variants, across six hidden dimensions d . Two candidate fixes are tested: a static, frequency-based output bias in the spirit of logit adjustment (Menon et al., 2021), and an L_2 -normalized cosine softmax with similarity temperature γ that removes the ability to encode information via magnitude. γ is tested as learned and fixed, since the model may shift magnitude into γ .

Our contributions are as follows:

- Causal isolation linking outlier magnitude to Zipfian skew, with vocabulary, model size, and configuration fixed.
- A hidden-dimension sweep linking d to outlier magnitude via a U-shaped relation from fixed-norm activation normalization.
- Demonstrating high- γ cosine softmax removes representational anisotropy; frequency bias and learned γ fail to do so.
- An explanation for absent INT8 gains from outlier suppression at this scale, attributed to activation magnitude ratios far below large pretrained models.

^{*}Equal contribution.

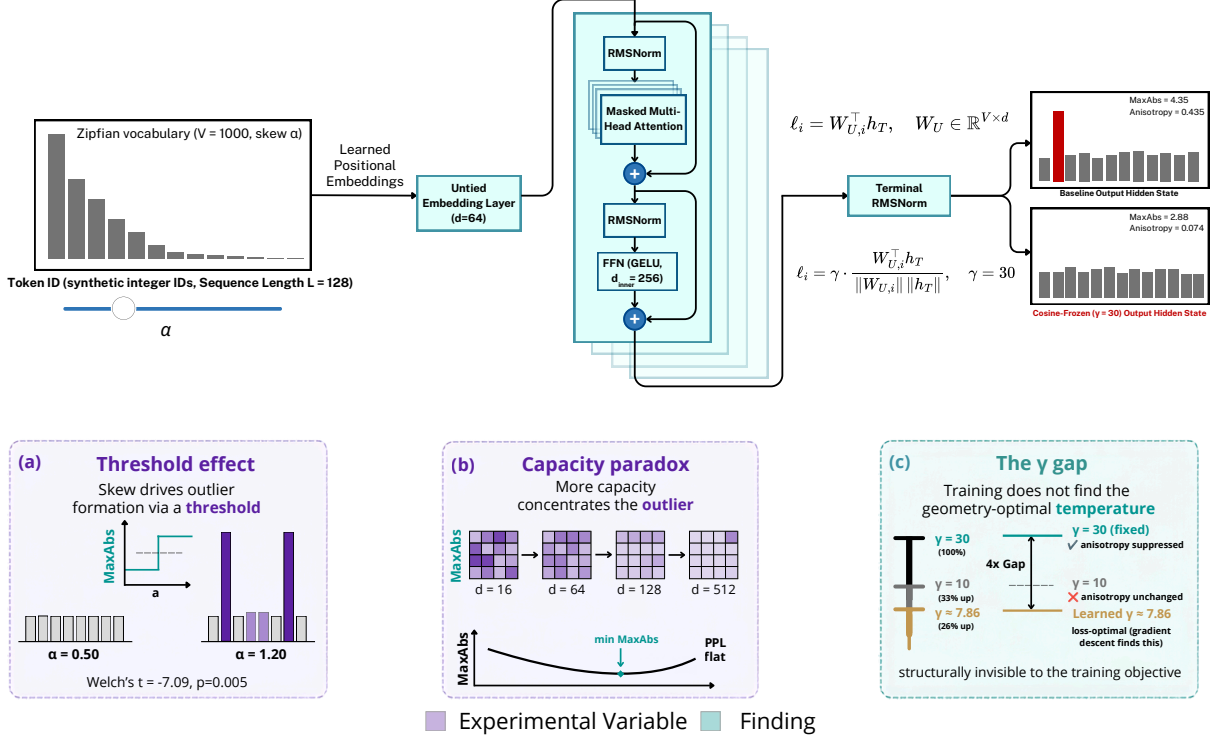


Figure 1: Overview of the causal mechanism and its three findings. A synthetic transformer is trained on Zipfian data with controllable skew α ; two independently trained output heads differ only in projection design, producing a hidden state h_T with an anomalous dimension (BASELINE) or none (COSINE-FROZEN, $\gamma = 30$). **(a)** Outlier formation is a threshold effect in α , not a gradual increase (Section 4.1). **(b)** Outlier magnitude is non-monotonic in hidden dimension d , falling then concentrating once capacity exceeds task requirements (Section 4.2). **(c)** A large, fixed temperature suppresses anisotropy, but the value gradient descent finds on its own does not (Section 4.3).

2 Related Work

Outlier dimensions and frequency. ODs were first reported as a disruptive artifact of BERT-style models (Kovaleva et al., 2021; Timkey and van Schijndel, 2021), and are one symptom of a broader anisotropy problem in learned representations: word and contextual embeddings alike tend to occupy a narrow cone of the embedding space rather than spreading isotropically (Gao et al., 2019; Ethayarajh, 2019). Puccetti et al. (2022) link ODs to token frequency directly: retraining BERT-medium with word frequency shuffled almost removes the effect, the closest prior work to a causal claim, though still on a real pretrained model where frequency cannot be manipulated independently of everything else it correlates with. Macocco et al. (2025) show last-layer ODs recur across eight modern pretrained models (7B–14B parameters) and conclude they are a specialized, useful mechanism for a frequent-word prediction heuristic, not simply a bug. Both studies are correlational or model-specific; neither isolates skew as an independent variable, and neither ties the phenomenon to a spe-

cific rank-based mechanism. Causal-intervention methodology is well established elsewhere in NLP interpretability (Vig et al., 2020; Geiger et al., 2021), but has not previously been applied to outlier dimensions. Our synthetic causal isolation is a controlled complement to these findings, not a competing claim: assuming frequency and ODs are linked in real models, we ask what causes that link.

Theoretical grounding. The *softmax bottleneck* (Yang et al., 2018) shows a standard softmax output layer is a low-rank bilinear form that cannot represent arbitrarily complex context-conditional word distributions (Chang and McCallum, 2022). No prior OD study connects this theory to outlier formation specifically; we use it as our candidate mechanism.

Suppressing outliers geometrically. L2-normalized, temperature-scaled softmax variants are established in representation learning for enforcing angular rather than magnitude-based decision boundaries, most prominently in face recognition, where a constant or adaptively-scaled

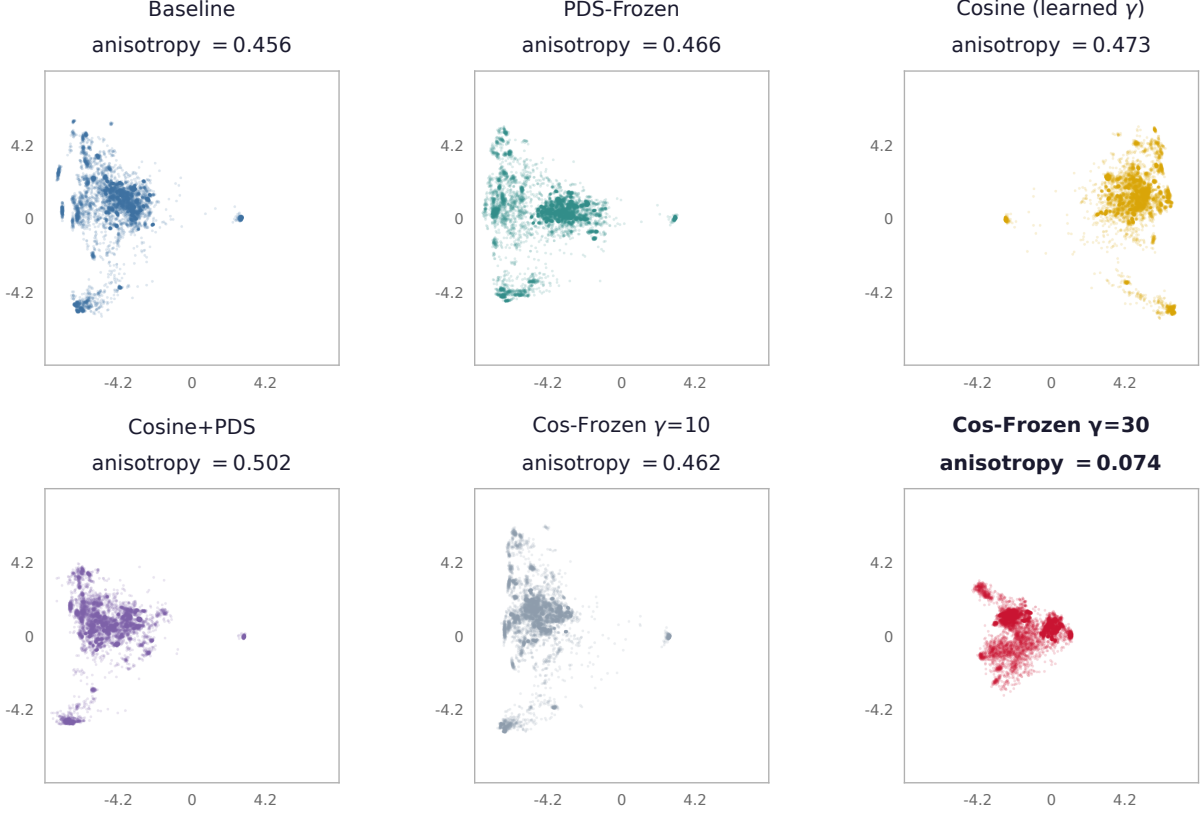


Figure 2: 2D PCA projections of the final hidden state across all six output heads ($\alpha = 1.2$, $d = 64$), annotated with mean anisotropy. Five of six configurations, including the learned- γ head, show a dense core stretched along one dominant direction, which is the outlier dimension known to distort probing and quantization. Only COSINE-FROZEN ($\gamma = 30$) produces a contained, rounder cloud, indicating an isotropic geometry free of this distortion. As Section 4.3 establishes, achieving this requires fixing γ ; gradient descent alone does not find it.

cosine margin replaces the standard inner-product softmax (Wang et al., 2018; Deng et al., 2019; Kobayashi, 2025), and in nGPT (Loshchilov et al., 2025), which normalizes the entire residual stream onto a hypersphere for training-stability reasons. Logit adjustment (Menon et al., 2021) addresses long-tailed label frequency directly at the output layer instead. None of these lines of work tests whether a constant vs. learned temperature changes the geometric outcome, which is the falsification test at the center of our cosine softmax result.

Downstream effects. Outlier dimensions are known to make INT8 post-training quantization lossy, since quantizer resolution is calibrated to the largest activation present (Bondarenko et al., 2023), and to resemble the “massive activations” documented across large models (Sun et al., 2024). We test whether suppressing ODs recovers quantization robustness, and find the effect is scale-dependent (Section 4).

3 Methodology

3.1 Synthetic Zipfian corpus

We define a vocabulary of 1000 abstract tokens whose marginal frequencies follow a Zipfian distribution, governed by a single skew parameter $\alpha \in [0, 1.95]$.

$$p_{\text{zipf}}(w_i) \propto (i + 1)^{-\alpha}, \quad i \in [1, V - 1], \quad (1)$$

normalized over the sampling range. At $\alpha = 0$ this reduces to a uniform distribution; at $\alpha = 1$ it approximates natural-language token frequency.

Context-conditional structure. To create genuine higher-order frequency dependencies without disturbing the marginal, we generate token sequences from a second-order Markov chain whose transitions are structured as follows. We split the vocabulary into a frequent head F of the 20 most common tokens and a rare tail R containing the remaining 979. For every bigram context with both tokens in F (400 such contexts), we sample the head-restricted transition distribution

$\mathbf{h}_c \in \Delta^{|F|-1}$ from a Dirichlet and allocate the full context-conditional transition as

$$\mathbf{h}_c \sim \text{Dirichlet}(\beta \cdot \mathbf{q}_F) \quad (2)$$

$$\begin{aligned} \mathbf{p}_c[F] &= M_F \cdot \mathbf{h}_c, \\ \mathbf{p}_c[R] &= \mathbf{p}_{\text{zipf}}[R], \end{aligned} \quad (3)$$

where $\beta = 0.1$, $\mathbf{q}_F$ is the Zipfian head distribution renormalized to sum to one, and M_F is the total probability mass assigned to the head. For any bigram involving a rare token, the marginal Zipfian is used directly in place of $\mathbf{p}_c$. This construction ensures every frequent bigram context receives a distinct local frequency skew while the corpus-wide marginal remains exactly Zipfian: because the Dirichlet mean is $\mathbf{q}_F$ and the tail is deterministic, the expected transition over all contexts recovers p_{zipf} . We verify this empirically by sampling 10^6 tokens and confirming that the resulting empirical distribution matches the target (KL divergence below 10^{-2}). Sequences are generated autoregressively from this chain, warm-started with two tokens drawn from p_{zipf} , with fixed logged seeds.

3.2 Model

We use a 4-layer decoder-only transformer with hidden dimension d (default 64; swept in Section 4.2), 4 attention heads, a GELU feed-forward inner dimension of $4d$, pre-normalization with RMSNorm, and learned positional embeddings up to length $L = 128$. Token and unembedding matrices are untied to isolate W_U behavior. At $d = 64$ the model has approximately 2.1M parameters. Weight tying, biases in FFN and attention, and dropout are all disabled.

3.3 Output-head configurations

Let $h_T \in \mathbb{R}^d$ be the final-layer post-norm hidden state at the last position, $W_U \in \mathbb{R}^{V \times d}$ the unembedding matrix, and $b \in \mathbb{R}^V$ an optional bias. We compare six output heads:

BASELINE. $\ell_i = W_{U,i}^\top h_T$. A standard inner-product softmax head with no normalization, bias, or margin.

PDS-FROZEN. $\ell_i = W_{U,i}^\top h_T + b_i$, with $b_i = \log \hat{p}_{\text{emp}}(w_i)$ computed on the empirical corpus (with Laplace smoothing) and *frozen* throughout training. Instantiates the logit-adjustment intervention of Menon et al. (2021) as a static marginal correction.

COSINE (LEARNED γ). $\ell_i = \gamma \cdot \frac{W_{U,i}^\top h_T}{\|W_{U,i}\|_2 \|h_T\|_2}$, with $\gamma \in \mathbb{R}$ a learnable scalar initialized at $\gamma_0 = \log V \approx 6.91$.

COSINE + PDS-FROZEN. The cosine head above with the additive frozen marginal bias from PDS-FROZEN.

COSINE-FROZEN ($\gamma = 10$). The cosine head with γ at 10 throughout training. Isolates whether the geometric constraint alone (without temperature learning) produces the OD-suppression effect.

COSINE-FROZEN ($\gamma = 30$). Same, with $\gamma = 30$. This value is the standard scale in representation-learning cosine-softmax variants (Deng et al., 2019).

The Cosine-Frozen heads are the falsification probe for the cosine intervention: if learned γ suppresses ODs by finding a large value, its behavior should match the constant-large- γ heads; if it instead settles at a value that does not suppress ODs, that reveals a gap between the loss-minimizing γ and the geometry-minimizing γ .

3.4 Training and evaluation

All configurations are trained for 10,000 steps with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$), using a linear warmup (Goyal et al., 2017) over the first 500 steps followed by cosine decay from 3×10^{-4} to 3×10^{-5} . We use a batch size of 256 sequences of length 128, weight decay of 0.01 (excluded from normalization parameters, biases, and γ), and gradient clipping (Pascanu et al., 2013) at global norm 1.0. For the capacity sweep (Section 4.2), we additionally calibrate the learning rate per hidden-dimension size through short pilot runs to ensure the outlier-capacity relationship is not confounded by optimization differences across widths. All experiments use fixed random seeds.

Metrics. We extract the final-layer hidden state h_T at the last position of each test sequence and collect these into a matrix H of size $10,000 \times d$. From H we compute three geometric metrics. *Max-Abs* captures the single largest activation across all dimensions and all test examples. Per-dimension excess *kurtosis* measures how peaked each dimension’s activation distribution is relative to a normal distribution; we report the maximum across dimensions. *Anisotropy* quantifies the redundancy of the

full representation by measuring the average cosine similarity between randomly paired hidden states, where higher values indicate a more collapsed, less isotropic geometry. We also report standard test perplexity. To assess downstream quantization effects, we apply symmetric per-channel INT8 quantization to the unembedding matrix W_U and per-tensor quantization to h_T , following Jacob et al. (2018), and report the relative perplexity change $\Delta\text{PPL}\%$.

Statistical protocol. Every configuration is run for 3 independent seeds; the headline Sweep 3 configurations BASELINE and COSINE (LEARNED γ) are run for 5 seeds. Aggregate statistics are reported as mean $\pm$ standard deviation over seeds. Pairwise comparisons in Sweep 3 use Welch’s t -test with Holm–Bonferroni correction over the five head-vs-BASELINE pairs; effect sizes are reported as Cohen’s d .

4 Results

Table 1 and Figure 3a summarize the skew sweep (Section 4.1); Table 2 and Figure 4 summarize the capacity sweep (Section 4.2); Table 3 and Figures 2, 5, 6 summarize the output-head interventions (Section 4.3).

4.1 Skew Drives OD Formation

Across 22 skew values spanning $\alpha \in [0, 1.95]$, the results were as follows. For $\alpha \leq 0.85$, MaxAbs remains indistinguishable from the uniform-vocabulary baseline. A sharp jump occurs between $\alpha = 0.85$ and $\alpha = 1.00$, after which the effect largely saturates, rising only gradually to its maximum at $\alpha = 1.95$. Anisotropy follows a different trajectory: it stays near zero before the transition and peaks later, at $\alpha = 1.70$, well after MaxAbs has already plateaued, indicating that outlier magnitude and representational collapse are not the same signal. Key points are summarized in Table 1.

Paralleling Section 4.3, anisotropy dissociates from MaxAbs: it remains near zero pre-transition before peaking at $\alpha = 1.70$ (anisotropy = 0.751). Conversely, per-dimension kurtosis peaks exclusively during the transition ($\alpha = 0.85$ – 1.00 ; kurtosis 25–28) and drops to single digits elsewhere, indicating it captures the structural discontinuity rather than steady-state outlier severity. The full 22-point grid, including per- α seed counts, is given in Appendix A.

α	MaxAbs	Anisotropy	MaxKurt
0.00	2.150 ± 0.143	0.0002	1.07
0.75	1.888 ± 0.071	0.0010	9.65
1.00	3.669 ± 0.267	0.1129	28.49
1.20	4.319 ± 0.256	0.4558	7.81
1.70	4.141 ± 0.165	0.7513	4.11
1.95	4.557 ± 0.190	0.6381	14.10

Table 1: Landmark points from the α sweep. Full 22-point grid in Appendix A.

4.2 Capacity Affects Outlier Magnitude Non-Monotonically

Table 2 and Figure 4 show a clear non-monotonic relationship between hidden dimension and outlier magnitude, contrary to the monotonic decrease predicted by a simple rank-bottleneck account. MaxAbs is minimized at $d = 128$ and rises $2.4\times$ by $d = 512$, while test perplexity remains essentially flat, indicating little task benefit from additional capacity beyond $d \approx 64$ – 128 . We interpret this as evidence that once capacity exceeds the task’s requirements, additional hidden dimensions no longer reduce outlier pressure. Instead, under fixed-norm (RMSNorm) activation normalization, *unused* capacity appears to concentrate rather than dilute the outlier, consistent with a fixed representational “budget” being redistributed across more dimensions.

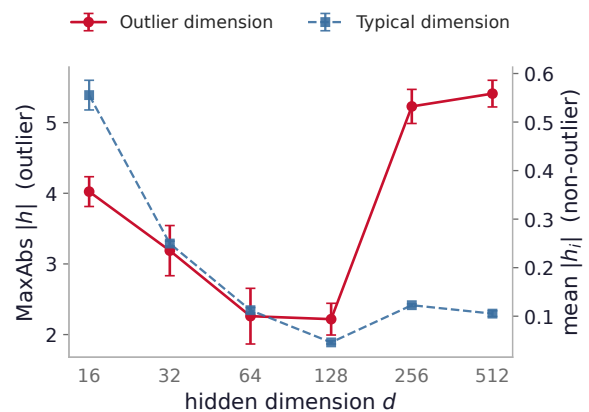
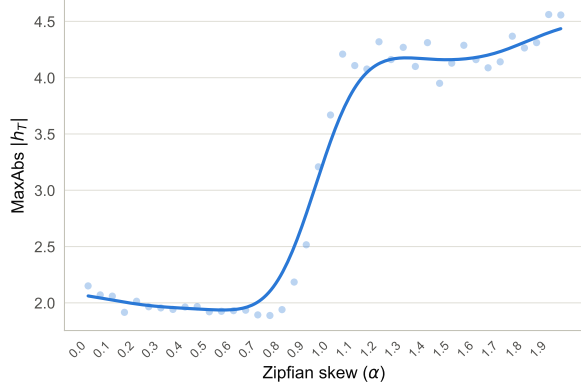
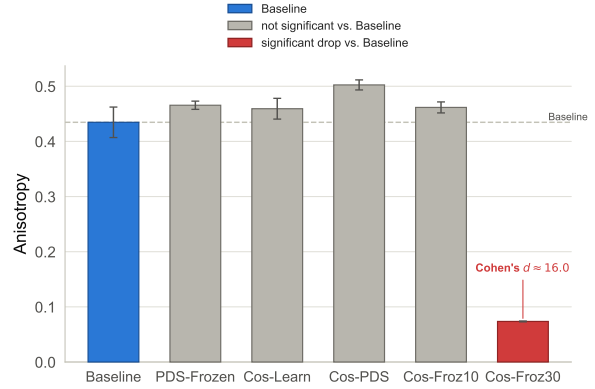


Figure 4: Outlier magnitude (left axis) and typical non-outlier magnitude (right axis) vs. hidden dimension d , at $\alpha = 1.0$, BASELINE head, with per-width LR calibration (Section 3.4). Both decrease as d increases from 16 to 128, then outlier magnitude rises sharply for $d \geq 256$ while typical-dimension magnitude remains flat.



(a) MaxAbs $|h_T|$ vs. Zipfian skew α (BASELINE, $d = 64$). Outlier magnitude is statistically indistinguishable from $\alpha = 0$ up to $\alpha \approx 0.85$ and rises sharply between $\alpha = 0.85$ and $\alpha = 1.00$ (+1.48); by $\alpha = 1.00$, MaxAbs already differs significantly from $\alpha = 0$ (Welch’s $t = -7.09$, $p = 0.005$), and the effect largely plateaus for $\alpha \gtrsim 1.2$.



(b) Anisotropy across the six output-head configurations at $\alpha = 1.2$, $d = 64$. Only COSINE-FROZEN ($\gamma = 30$) differs significantly from BASELINE in the direction of reduced anisotropy (Holm–Bonferroni corrected $p < 0.001$, Cohen’s $d \approx 16.0$).

Figure 3: (a) Zipfian skew causally drives outlier formation, with a sharp transition rather than a smooth increase. (b) Of five candidate interventions, only a large, constant cosine-softmax temperature meaningfully flattens representational geometry.

d	MaxAbs	Typical mag.	PPL	LR
16	4.024 ± 0.211	0.556 ± 0.031	216.73	3×10^{-3}
32	3.188 ± 0.355	0.250 ± 0.005	214.38	3×10^{-3}
64	2.261 ± 0.394	0.112 ± 0.006	214.40	3×10^{-3}
128	2.218 ± 0.224	0.046 ± 0.003	214.68	3×10^{-3}
256	5.229 ± 0.241	0.123 ± 0.003	217.01	1×10^{-3}
512	5.411 ± 0.189	0.105 ± 0.006	217.91	1×10^{-3}

Table 2: Capacity sweep at $\alpha = 1.0$, BASELINE head, 3 seeds. MaxAbs is minimized at $d = 128$ and rises $2.4\times$ by $d = 512$.

4.3 Cosine Softmax Removes Anisotropy at Larger Temperatures

Table 3 gives the full six-head comparison pre-viewed in Figure 2 and Figure 3b. Three results stand out.

The frozen marginal bias does not work. PDS-FROZEN is statistically indistinguishable from BASELINE on both MaxAbs and anisotropy: a static, context-independent frequency correction fails to remove the outlier, consistent with the mechanism being driven by context-conditional, not marginal, frequency pressure.

Only a large, constant similarity temperature γ flattens geometry. COSINE-FROZEN ($\gamma = 30$) is the sole configuration with a significant anisotropy reduction, consistent with the visibly contained, rounder projections in Figure 2. Critically, the *learned* γ head, which could in principle discover this same value through gradient descent,

instead converges to $\gamma \approx 7.86$ with no significant anisotropy change from BASELINE. The trajectory is smooth and decelerating, so this is a genuine optimum of the training objective, not an artifact of insufficient steps: gradient descent finds the loss-optimal value, not the geometry-optimal one, and the two are far apart. This gap matters because nothing in the training signal penalizes it; perplexity is essentially identical whether γ sits near 8 or near 30, so the model has no incentive to move further even though the geometric consequence of doing so is large. COSINE-FROZEN ($\gamma = 10$) confirms this is a threshold effect rather than a smooth function of γ , essentially matching the learned- γ result.

One configuration moves in the wrong direction: COSINE+PDS-FROZEN shows a significant *increase* in anisotropy relative to BASELINE despite also showing a significant MaxAbs decrease. MaxAbs and anisotropy are not interchangeable proxies for the same underlying quantity; an intervention can shrink the single largest activation while making the overall representation *more*, not less, redundant.

No configuration shows a significant change in per-dimension kurtosis relative to BASELINE, including COSINE-FROZEN ($\gamma = 30$) itself. Anisotropy and kurtosis capture distinct geometric properties: the former reflects whole-representation redundancy, the latter single-dimension peakedness, and our strongest interven-

Head	MaxAbs	Anisotropy	MaxKurt	PPL	Δ PPL% (INT8)	γ
BASELINE	4.351 ± 0.347	0.435 ± 0.028	10.26 ± 3.74	78.23	0.008	—
PDS-FROZEN	4.316 ± 0.350	0.466 ± 0.008	7.83 ± 1.13	78.20	0.011	—
COSINE (LEARNED γ)	3.906 ± 0.199	0.459 ± 0.019	11.64 ± 1.18	78.42	0.005	7.86
COSINE+PDS-FROZEN	$3.585 \pm 0.116^*$	0.502 ± 0.009	11.25 ± 3.73	78.70	0.005	7.78
COSINE-FROZEN ($\gamma=10$)	3.768 ± 0.146	0.462 ± 0.010	7.82 ± 0.50	78.35	0.005	10
COSINE-FROZEN ($\gamma=30$)	$2.883 \pm 0.116^{**}$	$0.074 \pm 0.001^{**}$	9.69 ± 2.74	80.43	0.049	30

Table 3: Sweep 3 at $\alpha = 1.2$, $d = 64$ (BASELINE and COSINE (LEARNED γ): 5 seeds; others: 3 seeds). Stars denote significance vs. BASELINE, Welch’s t -test with Holm–Bonferroni correction over 5 comparisons per metric ($^*p < 0.05$, $^{**}p < 0.001$). COSINE+PDS-FROZEN’s anisotropy is significantly *higher* than BASELINE (Cohen’s $d = -2.92$), not lower.

tion affects only the former.

4.4 Robustness Checks

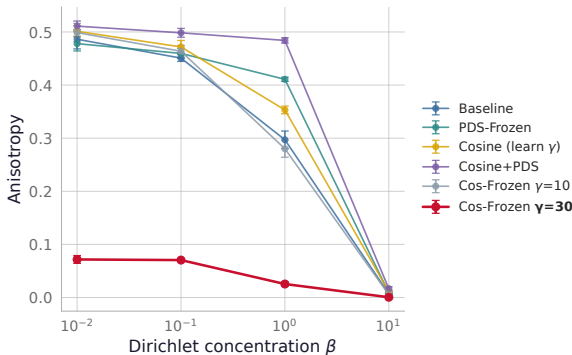


Figure 5: Anisotropy vs. Dirichlet head concentration β (Section 3.1), one line per head, $\alpha = 1.2$, $d = 64$, 3 seeds. COSINE-FROZEN ($\gamma = 30$) remains near-isotropic across three orders of magnitude of β ; the gap to the other five heads narrows only once $\beta = 10$ drives *every* configuration toward isotropy (context-conditional structure is essentially absent at this β ; see text).

Training dynamics. Figure 5 shows the $\gamma = 30$ result is robust across three orders of magnitude of β . At low and moderate concentrations, COSINE-FROZEN ($\gamma = 30$) stays near-isotropic while the other five heads cluster far higher (Appendix B). At $\beta = 10$, per-context Dirichlet draws become near-uniform and context-conditional structure essentially vanishes, so all six heads converge toward zero anisotropy, as there is little left for any head to distort. The $\gamma = 30$ advantage is thus specific to settings with a real phenomenon to suppress, not a blanket property of the cosine head.

Figure 6 shows outlier formation is non-monotonic in training time as well: MaxAbs and anisotropy both collapse within the first 500 steps before recovering to their converged values over the remaining run. This recovery is not synchronized

with loss convergence; perplexity reaches within 5% of its final value well before anisotropy has closed most of the distance to its own converged value, continuing to rise for thousands of additional steps after prediction quality has effectively converged. Representational geometry reorganizes on a slower timescale than the loss itself.

INT8 quantization. Consistent with Section 4.3’s magnitude comparisons, no configuration shows meaningful INT8 post-training quantization degradation at this scale (Δ PPL% $\in [0.005, 0.049]$ across all six heads, Table 3), including COSINE-FROZEN ($\gamma = 30$) despite its substantially lower MaxAbs. We return to why this null result is expected at this scale in Section 5.

5 Discussion

Our results establish Zipfian skew as a causal, sufficient driver of outlier-dimension formation, and that the relationship is a **threshold effect rather than a gradual increase**. Below $\alpha \approx 0.85$, outlier magnitude is indistinguishable from the uniform-vocabulary baseline; the mechanism activates sharply in a narrow window and largely saturates thereafter, consistent with a phase-transition-like emergence (Power et al., 2022) rather than a graded response to frequency imbalance.

The most striking finding is not that cosine-softmax with a constant large temperature removes anisotropy, but that the same choice with γ learned does not. Gradient descent converges consistently to $\gamma \approx 7.86$, where anisotropy remains indistinguishable from BASELINE; the trajectory is smooth and stable across seeds, and COSINE-FROZEN ($\gamma = 10$) reproduces the same null, ruling out a training-length artifact. **The loss-optimal and geometry-optimal γ are separated by a factor of four**, and standard training reaches only the first

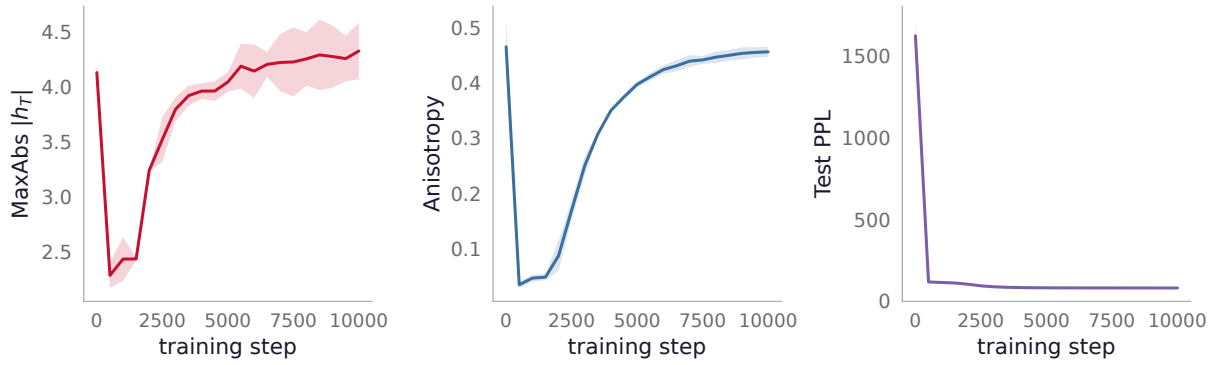


Figure 6: MaxAbs, anisotropy, and test perplexity over training steps, BASELINE head, $\alpha = 1.2$, mean over 3 seeds. Both MaxAbs and anisotropy collapse sharply within the first 500 steps before recovering non-monotonically.

(Papayan et al., 2020). This gap persists because perplexity is essentially flat across it, so nothing in the objective rewards moving further even though the geometric consequence of doing so is large. Properties a downstream user might value, such as isotropy, quantization robustness, or interpretability, need not lie on the training objective’s trajectory, and optimization gives no guarantee of finding them even when reachable in principle (Soudry et al., 2018). Interventions requiring externally-set hyperparameters are load-bearing on choices gradient descent will not make.

The capacity sweep produces a non-monotonic, U-shaped relationship between hidden dimension and outlier magnitude, minimal near $d = 128$ with a $2.4\times$ rebound at $d = 512$, while perplexity stays flat throughout. Once d exceeds task requirements, fixed-norm normalization (Zhang and Sennrich, 2019) appears to concentrate rather than dilute the outlier: we hypothesize that the representational budget redistributes across dimensions in a way that raises the largest activation. MaxAbs, anisotropy, and kurtosis dissociate throughout both sweeps (Godey et al., 2024) and are not interchangeable proxies for one quantity, as COSINE+PDS-FROZEN shrinks MaxAbs while anisotropy rises. Macocco et al. (2025) established the frequency-outlier link correlationally across eight pretrained models; our design confirms it causally and reveals three findings inaccessible to that framework: the threshold onset, the non-monotonic capacity curve, and the γ dose-response.

6 Conclusion

We isolated Zipfian frequency skew as a controlled variable in a synthetic transformer and showed that

it is a causal driver of outlier-dimension formation, that the effect is a threshold rather than a slope, and that model capacity has a non-monotonic influence at fixed skew. A frozen frequency-based output bias does not remove the phenomenon; an L_2 -normalized cosine softmax removes representational anisotropy, but only under a large, constant similarity temperature that gradient descent does not discover on its own. Standard training on the language-modeling objective converges to a temperature that leaves the outlier geometry intact. Outlier dimensions are, in this sense, **both causally driven by the data and structurally invisible to the objective that produces them**.

7 Limitations

Single architecture. Our experiments use a decoder-only transformer; outlier dimensions also occur in encoder-only and encoder-decoder models (Kovaleva et al., 2021; Timkey and van Schijndel, 2021), and extending our causal findings there is left to future work. The mechanism we identify is architecture-general in principle, though thresholds may vary across families.

Output-head intervention only. Cosine softmax modifies only the output projection, leaving the residual stream, attention, and feed-forward layers untouched. Broader interventions, such as normalized residual streams (Loshchilov et al., 2025), may yield complementary effects beyond our scope.

No downstream evaluation. We evaluate perplexity and representational geometry, not downstream tasks. The $\gamma = 30$ head’s small ($\sim 2.6\%$) perplexity increase over BASELINE (Table 3) may or may not be recovered by its geometric benefit in probing or transfer.

References

- Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. 2023. [Quantizable transformers: Removing outliers by helping attention heads do nothing](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 75067–75096. Curran Associates, Inc.
- Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. *Advances in neural information processing systems*, 35:18878–18891.
- Haw-Shiuan Chang and Andrew McCallum. 2022. [Softmax bottleneck makes language models unable to represent multi-mode word distributions](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8048–8073, Dublin, Ireland. Association for Computational Linguistics.
- Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. [ArcFace: Additive angular margin loss for deep face recognition](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699.
- Kawin Ethayarajh. 2019. [How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.
- Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. [Representation degeneration problem in training natural language generation models](#). In *International Conference on Learning Representations*.
- Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. [Causal abstractions of neural networks](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc.
- Nathan Godey, Éric Clergerie, and Benoît Sagot. 2024. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35–48.
- Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. [Accurate, large minibatch SGD: training imagenet in 1 hour](#). *CoRR*, abs/1706.02677.
- Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713.
- Takumi Kobayashi. 2025. Temperature in cosine-based softmax loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22199–22208.
- Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. [BERT busters: Outlier dimensions that disrupt transformers](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3392–3405, Online. Association for Computational Linguistics.
- Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. 2025. ngpt: Normalized transformer with representation learning on the hypersphere. In *International Conference on Learning Representations*, volume 2025, pages 74014–74038.
- Iuri Macocco, Nora Graichen, Gemma Boleda, and Marco Baroni. 2025. [Not a nuisance but a useful heuristic: Outlier dimensions favor frequent tokens in language models](#). In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 109–136, Suzhou, China. Association for Computational Linguistics.
- Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2021. [Long-tail learning via logit adjustment](#). In *International Conference on Learning Representations*.
- Vardan Papyan, XY Han, and David L Donoho. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663.
- Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. [On the difficulty of training recurrent neural networks](#). In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1310–1318, Atlanta, Georgia, USA. PMLR.
- Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.
- Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell’Orletta. 2022. [Outlier dimensions that disrupt transformers are driven by frequency](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1286–1304, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. 2018. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. 2024. [Massive activations in large language models](#). In *First Conference on Language Modeling*.

William Timkey and Marten van Schijndel. 2021. [All bark and no bite: Rogue dimensions in transformer language models obscure representational quality](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. [Investigating gender bias in language models using causal mediation analysis](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. 2018. [Breaking the softmax bottleneck: A high-rank RNN language model](#). In *International Conference on Learning Representations*.

Biao Zhang and Rico Sennrich. 2019. [Root mean square layer normalization](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

A Full skew sweep

Table 4 gives the complete 22-point α sweep referenced in Section 4.1. Four points ($\alpha \in \{0.25, 1.30, 1.80, 1.95\}$) were run with $n = 2$ seeds due to a partial rerun; all others are $n = 3$.

B Full β ablation table

Table 5 gives the numerical values underlying Figure 5 in Section 4.3. COSINE-FROZEN ($\gamma = 30$) achieves anisotropy below 0.075 at every β tested. At $\beta = 10$, all six heads converge toward isotropy: the per-context Dirichlet draws are close to uniform at this concentration, so context-conditional structure is essentially absent from the training signal, and there is little anisotropy for any head to have. PPL also rises uniformly from ~ 76 at $\beta = 0.01$ to

α	n	MaxAbs	Anisotropy	MaxKurt
0.00	3	2.150 ± 0.143	0.0002	1.07
0.10	3	2.098 ± 0.180	0.0002	1.31
0.25	2	2.061 ± 0.043	0.0006	1.85
0.35	3	2.001 ± 0.174	0.0004	2.31
0.50	3	1.918 ± 0.192	0.0007	6.29
0.65	3	1.885 ± 0.087	0.0006	8.44
0.75	3	1.888 ± 0.071	0.0010	9.65
0.85	3	2.185 ± 0.141	0.0018	25.13
0.90	3	2.744 ± 0.242	0.0324	27.11
1.00	3	3.669 ± 0.267	0.1129	28.49
1.05	3	3.913 ± 0.194	0.2103	22.87
1.10	3	4.096 ± 0.223	0.3308	16.05
1.20	3	4.319 ± 0.256	0.4558	7.81
1.30	2	4.194 ± 0.187	0.5641	6.02
1.40	3	4.098 ± 0.201	0.6602	5.09
1.50	3	4.126 ± 0.214	0.7211	4.53
1.60	3	4.135 ± 0.178	0.7418	4.24
1.70	3	4.141 ± 0.165	0.7513	4.11
1.80	2	4.269 ± 0.155	0.7295	6.83
1.85	3	4.340 ± 0.176	0.7128	8.24
1.90	3	4.481 ± 0.202	0.6759	11.66
1.95	2	4.557 ± 0.190	0.6381	14.10

Table 4: Full 22-point α sweep, BASELINE head, $d = 64$. Mean $\pm$ std across seeds.

~ 118 at $\beta = 10$, confirming the task itself becomes flatter as β grows.

B.1 Vocabulary permutation control

To confirm that outlier formation is driven by *frequency structure* rather than by which specific token identities occupy the frequent head, we retrained the BASELINE head at $\alpha = 1.2$, $d = 64$ with the vocabulary permuted before training: token identities were reassigned by a random permutation over $\{1, \dots, V - 1\}$ so that the Zipfian rank of each token no longer corresponds to its original identity, while the corpus statistics (marginal distribution shape, context-conditional structure) are preserved. This is the analogue in our controlled setting of the token-swap retraining performed by Puccetti et al. (2022) on encoder-only models.

Table 6 gives the results. MaxAbs (4.184 ± 0.302) is statistically indistinguishable from the unpermuted BASELINE reported in Table 3 (4.351 ± 0.347), as is anisotropy (0.456 ± 0.017 vs. 0.435 ± 0.028) and PPL (78.22 vs. 78.23). Outlier formation depends on the frequency distribution over the vocabulary, not on the specific token identities that happen to occupy its head.

C Hyperparameters and training details

Model. 4-layer decoder-only transformer. Hidden dimension $d = 64$ throughout except

β	Head	MaxAbs	Anisotropy	PPL	γ_{final}
0.01	BASLINE	4.757 ± 0.487	0.4862 ± 0.0182	76.38	—
0.01	PDS-FROZEN	4.441 ± 0.298	0.4780 ± 0.0138	76.38	—
0.01	COSINE (LEARNED γ)	3.744 ± 0.142	0.5011 ± 0.0105	76.47	7.88
0.01	COSINE+PDS-FROZEN	3.520 ± 0.263	0.5110 ± 0.0096	76.86	7.79
0.01	COSINE-FROZEN ($\gamma=10$)	3.714 ± 0.218	0.4988 ± 0.0173	76.46	—
0.01	COSINE-FROZEN ($\gamma=30$)	2.647 ± 0.051	0.0717 ± 0.0073	78.33	—
0.10	BASLINE	4.527 ± 0.603	0.4508 ± 0.0062	78.24	—
0.10	PDS-FROZEN	4.336 ± 0.185	0.4596 ± 0.0096	78.28	—
0.10	COSINE (LEARNED γ)	4.154 ± 0.319	0.4718 ± 0.0122	78.42	7.88
0.10	COSINE+PDS-FROZEN	3.847 ± 0.081	0.4984 ± 0.0083	78.80	7.76
0.10	COSINE-FROZEN ($\gamma=10$)	4.000 ± 0.189	0.4636 ± 0.0109	78.39	—
0.10	COSINE-FROZEN ($\gamma=30$)	3.123 ± 0.170	0.0705 ± 0.0013	80.61	—
1.00	BASLINE	3.905 ± 0.376	0.2970 ± 0.0165	90.35	—
1.00	PDS-FROZEN	4.224 ± 0.190	0.4108 ± 0.0039	90.31	—
1.00	COSINE (LEARNED γ)	3.799 ± 0.076	0.3532 ± 0.0071	90.83	7.67
1.00	COSINE+PDS-FROZEN	3.862 ± 0.103	0.4841 ± 0.0051	90.82	7.30
1.00	COSINE-FROZEN ($\gamma=10$)	3.476 ± 0.147	0.2802 ± 0.0161	91.24	—
1.00	COSINE-FROZEN ($\gamma=30$)	2.554 ± 0.079	0.0254 ± 0.0036	96.13	—
10.0	BASLINE	2.276 ± 0.039	0.0069 ± 0.0013	118.05	—
10.0	PDS-FROZEN	2.834 ± 0.386	0.0087 ± 0.0018	118.17	—
10.0	COSINE (LEARNED γ)	2.259 ± 0.111	0.0112 ± 0.0024	118.12	6.81
10.0	COSINE+PDS-FROZEN	3.300 ± 0.486	0.0167 ± 0.0036	118.09	5.81
10.0	COSINE-FROZEN ($\gamma=10$)	2.347 ± 0.111	0.0054 ± 0.0013	118.30	—
10.0	COSINE-FROZEN ($\gamma=30$)	2.314 ± 0.122	0.0005 ± 0.0001	118.77	—

Table 5: Anisotropy, MaxAbs, and PPL by (β , head), 3 seeds each, $\alpha = 1.2$, $d = 64$. Mean $\pm$ std.

Seed	MaxAbs	MaxKurt	Anisotropy	PPL
0	4.129	11.05	0.480	78.17
1	3.843	11.36	0.446	78.20
2	4.580	7.60	0.442	78.28
mean $\pm$ std	4.184 ± 0.302	10.00 ± 1.68	0.456 ± 0.017	78.22

Table 6: Vocabulary permutation control, BASELINE head, $\alpha = 1.2$, $d = 64$, 3 seeds.

the capacity sweep (Section 4.2), where $d \in \{16, 32, 64, 128, 256, 512\}$. Number of attention heads = 4. Feedforward hidden dimension = $4d$ with GELU nonlinearity. Pre-LayerNorm using RMSNorm. Learned positional embeddings. No dropout. Embedding and unembedding matrices are not tied.

Training. AdamW, learning rate 3×10^{-4} (for the capacity sweep, per-width calibrated from $\{10^{-4}, 3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$ using 200-step pilots), cosine schedule with 500-step linear warmup, weight decay 0.01, gradient clipping at global norm 1.0, batch size 256, sequence length 128, 10,000 optimizer steps.

Data. Vocabulary size $V = 1000$. Frequent head size $|F| = 20$. Context-conditional structure via Dirichlet-Multinomial sampling with concentration $\beta = 0.1$ over the head at each of the $|F|^2 = 400$ distinct head-bigram contexts (Section 3.1). Pre-

flight marginal-preservation KL threshold 10^{-2} .

Reproducibility. All experiments run on a single NVIDIA L4 GPU (24GB VRAM). Three seeds per configuration except the two headline configurations of Sweep 3 (BASELINE and COSINE (LEARNED γ)), which use five seeds. Random seed values $\{0, 1, 2\}$ for three-seed runs and $\{0, 1, 2, 3, 4\}$ for five-seed runs.

D Learned γ trajectory

Figure 7 shows the trajectory of the learned γ scalar during training. The stable, decelerating convergence rules out the possibility that the observed final value ($\gamma \approx 7.86$) is an artifact of insufficient training: at 10,000 steps the parameter has effectively converged, and further training would not close the gap to the $\gamma = 30$ value at which COSINE-FROZEN suppresses anisotropy.

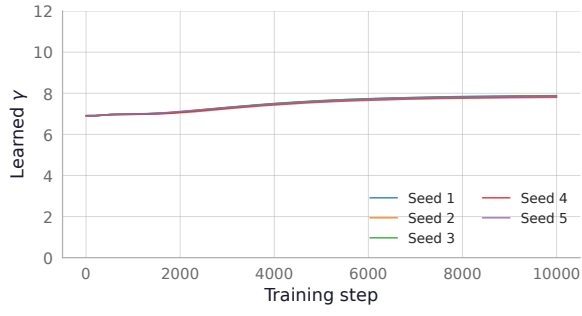


Figure 7: The figure shows the learned γ trajectory over training for all 5 seeds at $\alpha = 1.2$, $d = 64$. From initialization at $\gamma_0 = \log V \approx 6.91$, the value rises smoothly and converges to $\gamma \approx 7.86$ by step 8,000, moving less than 0.05 across the final 2,000 steps for all seeds.