2026 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEP. 28– OCT. 1, 2026, ATLANTA, USA

COMMUTATOR-BASED SYMMETRY ANALYSIS IDENTIFIES REPLACEABLE ATTENTION HEADS IN TRANSFORMERS

Mitchell A. Thornton

Darwin Deason Institute for Cyber Security
Southern Methodist University, Dallas, TX, USA
mitch@smu.edu

ABSTRACT

We propose a group-theoretic diagnostic for transformer attention based on commutator norms with finite permutation groups. The permutation-baseline-normalized commutator residual δ_{norm} measures whether a head's average attention matrix commutes with a candidate group more than a random matrix would, computed from attention statistics alone with no labels and no probing classifier. On the GPT-2 family, heads partition cleanly into a positional class ($\delta_{\text{norm}} < 0.5$, approximately circulant) and a content class ($\delta_{\text{norm}} > 0.9$). The classification is actionable and its value is structural: replacing positional heads with a fixed causal circulant pattern is functionally exact away from the class boundary (perplexity ratios 0.997 to 1.006 for the strongest heads on every GPT-2 size, improving on GPT-2 Large; boundary heads at $\delta_{\text{norm}} > 0.44$ cost up to 2.5% on the two smallest models) at 3.2% attention-FLOP savings and zero added parameters, while uniform substitution of content heads degrades perplexity by up to 89%. Because the commutator ranking is orthogonal to ablation-based importance (0% top-20 overlap), it composes with deletion-based pruning: the combination affects 19% of heads at a 0.980 perplexity ratio and 17% attention-FLOP savings, with the improvement attributable entirely to the deletion component. The classification is stable under fine-tuning and extends qualitatively across architectures: learned-absolute (OPT), rotary (Pythia, Phi-3, Gemma-2), and bidirectional (BERT, ModernBERT) models all exhibit the positional class. Sensitivity analyses show the positional census depends on the calibration corpus and on fine-tuning only at the class boundary.

***Index Terms*—** attention head compression, transformer interpretability, group-theoretic methods, structured pruning, language models

1. INTRODUCTION

Transformer [1] compression has matured along three axes: head pruning by gradient sensitivity or ablation loss [3, 4, 5, 6, 7], quantization of weights and activations [8, 9], and structured removal of layers or blocks [10]. These methods are effective but agnostic to *why* a head is compressible: an empirical property discovered by search.

We propose a complementary axis: *symmetry-based head replacement*. The premise is that some attention heads compute, on average, a fixed pattern that depends only on relative position, and these heads can be replaced exactly by their structural approximation rather than removed. The structural approximation costs no parameters and no computation beyond a single fixed convolution. Identifying which heads qualify reduces to a measurement problem: which heads' average attention matrices commute with a finite permutation group?

Concretely, we develop a commutator-residual metric δ_{norm} that classifies the attention heads of GPT-2 [2] and related models into a positional class and a content class. The classification is binary and clean: $\delta_{\text{norm}} < 0.5$ versus $\delta_{\text{norm}} > 0.9$, with few intermediate heads. Replacing positional-class heads with fixed causal circulant patterns is functionally exact away from the class boundary: the strongest heads replace at perplexity ratios 0.997 to 1.006 on all four GPT-2 sizes, the full canonical sets at 0.997 to 1.025, with the excess attributable to boundary heads at $\delta_{\text{norm}} > 0.44$. Uniform substitution of content heads is catastrophic (up to $1.89\times$). We state the standalone contribution precisely: circulant replacement alone yields a 1.022 perplexity ratio (1.003 for the six core heads) with 3.2% attention-FLOP savings at zero added parameters. Improvements arise only in combination with ablation-based deletion [3], whose head choices are disjoint from ours (0% top-20 overlap): the combined strategy affects 19% of heads at a 0.980 perplexity ratio and 17% attention-FLOP savings, with the improvement attributable entirely to the deletion component (0.953 alone). The structural account also yields a falsifiable empirical trend within the GPT-2 family: the number of structured blocks grows monotonically with depth, approximately doubling at each model-size step (1, 2, 4, 8 structured blocks at depths 6, 12, 24, 36). Additionally, a sensitivity analysis is provided in Sec. 4, cross-architecture results including modern compact models in Sec. 8, and measured deployment metrics in Sec. 9.

2. METHOD: COMMUTATOR-RESIDUAL DIAGNOSTICS

Definitions. Let $G \subset S_M$ be a finite group acting on $\mathbb{R}^M$ by permutation matrices $\{\pi_g\}_{g\in G}$. For $R \in \mathbb{R}^{M\times M}$, the commutator $[\pi_g, R] = \pi_g R - R\pi_g$ vanishes iff R commutes with π_g. We define the commutator residual and its permutation-baseline normalization

$$\bar{\delta}(G,R) = \frac{1}{|G|}\sum_{g\in G}\frac{\|[\pi_g,R]\|_F}{\|\pi_g\|_F\,\|R\|_F}, \quad \delta_{\text{norm}} = \frac{\bar{\delta}(G,R)}{\bar{\delta}(G,R_{\text{rand}})}, \quad (1)$$

where R_{rand} is a Wishart matrix of the same dimension. Normalization is essential: without it, small block-size groups produce mechanically lower $\bar{\delta}$ on any input because the typical commutator norm scales with the support of π_g; the baseline exposes such artifacts ($\delta_{\text{norm}} > 1$). Values $\delta_{\text{norm}} < 1$ indicate genuine algebraic structure; $\delta_{\text{norm}} > 1$ indicates anti-structure, a non-trivial possibility

Code and per-head diagnostic data: https://github.com/mitch-thornton/transformer-head-symmetry.

for trained representations. The construction sits within equivariant deep learning [11, 12] but inverts its direction: rather than building equivariance in, we measure the equivariance training left behind, and exploit what is found.

Role of Lemma 1. The diagnostic threshold acquires an estimation-theoretic meaning through a Cramér–Rao argument. The zero-mean Gaussian model is the standard setting in which the Fisher information of a covariance parameter is available in closed form; it is not an assumption about attention statistics but the canonical vehicle for the bound. Let $\hat{R}_G(x) = |G|^{-1} \sum_g (\pi_g x)(\pi_g x)^\top$ denote the group-averaged outer-product estimator.

Lemma 1 (CRB for G-invariant covariance estimation)**.** *Let $G \subset S_M$ act on $\mathbb{R}^M$ by permutation matrices and let $\Sigma \succ 0$ commute with every π_g. For a single observation $x \sim \mathcal{N}(0, \Sigma)$, the estimator $\hat{R}_G(x)$ is unbiased for Σ and attains the Cramér–Rao lower bound with equality among unbiased estimators of the G-invariant covariance; it is moreover the unique uniformly minimum-variance unbiased estimator.*

Proof. Equip $\mathrm{Sym}(M)$ with the Frobenius inner product $\langle X, Y \rangle = \mathrm{Tr}(XY)$, let $\mathcal{S}_G = \{S \in \mathrm{Sym}(M) : \pi_g S = S \pi_g \ \forall g\}$, and fix an orthonormal basis $\{B_a\}_{a=1}^m$ of $\mathcal{S}_G$. (i) The Reynolds operator $P_G(A) = |G|^{-1} \sum_g \pi_g A \pi_g^\top$ is idempotent and self-adjoint with range $\mathcal{S}_G$, hence the orthogonal projector onto $\mathcal{S}_G$; and $\hat{R}_G(x) = |G|^{-1} \sum_g \pi_g (x x^\top) \pi_g^\top = P_G(x x^\top)$. (ii) $\mathbb{E}[x x^\top] = \Sigma \in \mathcal{S}_G$, so $\mathbb{E}[\hat{R}_G] = P_G(\Sigma) = \Sigma$. (iii) $\mathcal{S}_G$ is the symmetric part of the commutant algebra of $\{\pi_g\}$, which is closed under products and inverses; hence $\Sigma \in \mathcal{S}_G$ iff $\Sigma^{-1} \in \mathcal{S}_G$. Writing $\Sigma^{-1} = \sum_a \eta_a B_a$ gives $\log p_\Sigma(x) = -\frac{1}{2} \sum_a \eta_a \langle B_a, x x^\top \rangle - \frac{1}{2} \log \det(2\pi\Sigma)$: the constrained model is a full linear exponential family with natural parameter η ranging over an open subset of $\mathbb{R}^m$ and complete sufficient statistic $t(x) = (\langle B_a, x x^\top \rangle)_a$ [30]. Since $\langle B_a, x x^\top \rangle = \langle B_a, P_G(x x^\top) \rangle$, we have $\hat{R}_G = \sum_a t_a B_a$: a linear bijection of t, unbiased for $\theta_a = \langle B_a, \Sigma \rangle$ coordinatewise. (iv) The Fisher information of the zero-mean Gaussian in the coordinates θ is $I_{ab} = \frac{1}{2} \mathrm{Tr}(\Sigma^{-1} B_a \Sigma^{-1} B_b) = \langle \frac{1}{2} \Sigma^{-1} B_a \Sigma^{-1}, B_b \rangle$, while Wick's theorem gives $\mathrm{Cov}(t_a, t_b) = \langle 2\, \Sigma B_a \Sigma, B_b \rangle$. The self-adjoint maps $A \mapsto \frac{1}{2} \Sigma^{-1} A \Sigma^{-1}$ and $A \mapsto 2\, \Sigma A \Sigma$ preserve $\mathcal{S}_G$ (products within the commutant algebra) and are mutually inverse on it, so their Gram matrices I and $\mathrm{Cov}(t)$ in the basis $\{B_a\}$ are mutually inverse: $\mathrm{Cov}(\hat{\theta}) = I^{-1}$, i.e., the bound holds with equality. Uniqueness follows from completeness of t and the Lehmann–Scheffé theorem [30]. □

Remark (single-sample attainment). Equality from one observation is not an artifact of averaging over $|G|$ terms: the constrained model remains a full exponential family, $\hat{R}_G$ is a bijection of its sufficient statistic, and in the mean-value parametrization the bound coincides with that statistic's covariance; group averaging is merely the geometric realization of P_G. If instead $\Sigma \notin \mathcal{S}_G$, the same estimator is unbiased for $P_G(\Sigma)$, and δ_{norm} measures what that projection discards.

The consequence used throughout the paper is the replacement rule: if $\delta_{\mathrm{norm}}(G, R) \ll 1$ then R lies near the G-invariant subspace, and the projection $P_G(R)$ onto that subspace is the natural minimal-parameter substitute for R. For the cyclic group this projection is exactly the best-fit circulant. Lemma 1 states that, on the invariant subspace, no unbiased use of the same observation does better; the diagnostic therefore identifies precisely the heads for which the substitution loses nothing that the head was computing.

Candidate groups. The cyclic group $\mathbb{Z}_M$ detects shift invariance (circulant structure); block-cyclic groups $\mathbb{Z}_b^{M/b}$, $b \in \{2, 4, 8, 16, 32\}$, detect block-circulant structure. The group with smallest δ_{norm} is reported as the best fit.

Computation and the causal-asymmetry question. For each block we compute the post-softmax attention matrix $A \in \mathbb{R}^{T \times T}$ averaged over the calibration corpus. Because Eq. (1) requires a square symmetric argument for the spectral interpretation of Sec. 10, the *diagnostic* evaluates δ_{norm} on $R_{\mathrm{attn}} = (A + A^\top)/2$. Symmetrization enters nowhere else: the *replacement* operation projects the raw causal A onto the nearest causal circulant (row-normalized averaging along the valid sub-diagonals), so the substituted pattern respects the causal mask exactly. The validity of using a symmetrized diagnostic to guide a causal substitution is established empirically rather than assumed: positional heads selected by the symmetrized δ_{norm} tolerate causal-circulant replacement at perplexity ratios 0.997 to 1.025 with a clean δ_{norm}-ordered dose-response, while uniform controls degrade by up to 89% (Table 3). For positional-structure measurements on hidden states we use $R_{\mathrm{pos}} = HH^\top/d$. Diagnostics run in float32; the compression experiments use a float64 manual forward verified to match the native model's perplexity exactly before modification (GPT-2 Medium's native float32 forward overflows to NaN on the evaluation set, so its baseline is the float64 value).

Canonical classification. All experiments in this paper use one classification, produced by the diagnostic script `compute_commutator_residual.py`, in the companion repository, on a fixed 4-text calibration corpus at $M = 64$; per-head values are also published in the repository. As a robustness check, we also recompute the classification inside each experiment script with its own calibration corpus, which produces superficially inconsistent positional-head counts; Sec. 4 reports that recomputation as a corpus-sensitivity analysis.

3. DIAGNOSTIC FINDINGS

We analyze four GPT-2 family models: DistilGPT-2 [13] (6 blocks, 82M), GPT-2 Small (12, 124M), Medium (24, 355M), and Large (36, 774M), all with the same calibration corpus.

Learned positional embeddings have strong Toeplitz structure. The positional embedding Gram matrix $R_{\mathrm{pos}} = W_{\mathrm{pos}} W_{\mathrm{pos}}^\top / d$ has $\delta_{\mathrm{norm}} = 0.056$ on $\mathbb{Z}_{64}$ for GPT-2 Small (94% below random), likewise across all GPT-2 variants and OPT-125m ($\delta_{\mathrm{norm}} = 0.285$); the Toeplitz diagonal's coefficient of variation is 0.064, confirming approximate translation equivariance.

The U-shape across the block stack. Fig. 1 plots δ_{norm} versus relative depth. All four sizes show the same shape: structured first blocks ($\delta_{\mathrm{norm}} < 0.5$), a strongly anti-structured middle plateau invariant at $\delta_{\mathrm{norm}} = 1.40 \pm 0.01$, and a return to structure in the final blocks. The input-side structured shoulder widens with depth (0, 1, 3, and 6 blocks across the four sizes) while the output-side shoulder stays at one to two blocks; the total structured-block count (1, 2, 4, 8) approximately doubles at each model-size step. We characterize this as an empirical within-family trend rather than a scaling law; whether the trend extends beyond the GPT-2 family is addressed in Sec. 8.

Attention heads partition into positional and content classes. Table 1 gives the census from the canonical diagnostic run; the histogram of head-level δ_{norm} is bimodal with few heads in $0.5 \leq \delta_{\mathrm{norm}} \leq 0.9$. Positional heads are a small fraction (1.0% to 6.9%) that concentrates in the early and early-middle layers (layers 0 to 4

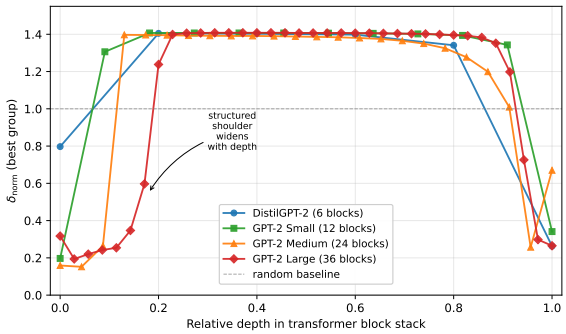

Fig. 1. Commutator residual δ_{norm} versus relative depth across the GPT-2 family. First and last blocks preserve cyclic structure; the middle plateau is anti-structured at $\delta_{\text{norm}} \approx 1.4$, invariant across model sizes.

Table 1. Attention head classification from the canonical diagnostic run (matches the released per-head CSVs). "Pos." is $\delta_{\text{norm}} < 0.5$; "Mixed" is $0.5 \leq \delta_{\text{norm}} \leq 0.9$; "Content" is $\delta_{\text{norm}} > 0.9$.

Model	Total	Pos.	Mixed	Content
DistilGPT-2	72	5	8	59
GPT-2 Small	144	7	13	124
GPT-2 Medium	384	4	23	357
GPT-2 Large	720	11	53	656

in Small) and spreads deeper in larger models (through layer 14 in Large), consistent with the depth trend above.

Distillation preserves the classification. DistilGPT-2 is initialized from alternate layers of GPT-2 Small [13]. Its positional heads are L0:{1,3,5} and L2:{7,11}; under the alternate-layer correspondence (student L2 ↔ teacher L4), these match the teacher's positional sets L0:{1,3,4,5} and L4:{7,11} exactly, except for teacher head L0H4, which is borderline in the teacher ($\delta_{\text{norm}} = 0.467$) and lands just above threshold in the student. The head-level classification, though not the precise δ_{norm} values, transfers through the distillation loss without any explicit structural constraint.

4. CALIBRATION SENSITIVITY OF THE CENSUS

The census of Table 1 could in principle depend on the calibration corpus, and we ran this experiment where each script recomputed the model classification on its own corpus. Table 2 reports the outcome for GPT-2 Small. The two heads adjacent to the threshold ($\delta_{\text{norm}} = 0.459$ and 0.467 canonically) cross it under corpus changes, and shorter effective sequence lengths ($M = 32$) admit one to two additional borderline heads; the core of the class is invariant. Practically, the classification is corpus-stable away from the boundary. Boundary heads should be validated individually as done in Table 3(b).

5. SYMMETRY-GUIDED COMPRESSION

All experiments below use the canonical classification and report the perplexity ratio against the unmodified model on a six-text held-out evaluation corpus (see the fine-tuning discussion in Sec. 7 that uses WikiText-2 [14] proper). Table 3 contains the four compression

Table 2. Corpus sensitivity of the GPT-2 Small positional census ($\delta_{\text{norm}} < 0.5$). Rows are alternative calibration settings, one per experiment script. Head L0H1 is positional under every setting (δ_{norm} 0.13 to 0.16).

Calibration corpus	Texts	M	Pos. heads
Canonical (diagnostic)	4 long	64	7
Replacement corpus	6 long	64	3
Strategy corpus	8 short	32	5
Fine-tuning analysis	4 short	32	5
SST-2 in-domain	50 sent.	32	4

experiments presented as a sectioned table.

(a) Selective replacement with controls. Each canonical positional head's average attention is projected onto the nearest causal circulant, which replaces the head's Q/K/softmax computation at inference. Two controls replace the highest-δ_{norm} content heads with a uniform causal pattern and with the head's own *average* attention matrix, the fairer comparison since it applies the identical averaging procedure to heads the diagnostic rejects. The controls separate sharply: content-head averages are far from uniform (mean per-row total-variation distance 0.85 to 0.91), uniform substitution is catastrophic on the smaller models (up to 1.889), and average-attention substitution is tolerated (0.998 to 1.015). Content heads are therefore not irreplaceable per se; their fixed substitute requires the full T^2-parameter average, whereas $\delta_{\text{norm}} < 0.5$ identifies exactly the heads whose average compresses further to a T-parameter causal circulant. A δ_{norm}-ordered progressive replacement exhibits the corresponding dose-response: on GPT-2 Small the six strongest heads cost 0.3%, and the seventh, the boundary head L0H4 that also crosses the threshold under fine-tuning (Sec. 7), accounts for the remaining 2.2%.

(b) Extended replacement. All heads with $\delta_{\text{norm}} < 0.9$ (positional plus mixed) are replaced, probing whether the mixed class is partially structural. The larger models tolerate and even improve under extended replacement (0.979 on Large with 64 heads); the smaller models degrade (1.141, 1.156).

(c) Quantization, matched control. Simulated INT8 and INT4 quantization of Q/K/V activations. Positional heads tolerate INT4 at zero cost on every model, but the matched-size content-head control tolerates it equally (row c, control); sensitivity appears only when all heads are quantized at once (up to 27.6% on Small). The matched control shows the effect is subset size and is not attributable to head class.

(d) Mixed-precision strategy. Circulant replacement for positional heads plus INT8 for all remaining heads touches every attention head using the classification as the sole guide, with no gradient information and no labeled data.

6. COMPARISON WITH ABLATION-BASED IMPORTANCE

To establish that δ_{norm} captures a property not covered by ablation-based head importance, we implement the Michel et al. [3] experiment on GPT-2 Small: zero each of the 144 heads and record the perplexity increase. The two rankings identify disjoint head sets at top-20 (zero overlap; random-chance expectation 20.8%), with Spearman $\rho = -0.402$ ($p < 10^{-4}$): heads with smaller commutator residual tend to have higher ablation importance, not lower. Fig. 2 visualizes the relationship.

Table 3. Symmetry-guided compression with the canonical classification (perplexity ratio on the held-out evaluation corpus; 1.000 is no degradation; lower is better; boldface marks ≤ 1.005). N is heads modified.

		DistilGPT-2		GPT-2 Small		GPT-2 Medium		GPT-2 Large	
Section	Experiment	N	Ratio	N	Ratio	N	Ratio	N	Ratio
(a)	Positional $\rightarrow$ causal circulant	5	1.017	7	1.025	4	**1.000**	11	**0.997**
(a)	Content $\rightarrow$ uniform (control)	5	1.444	7	1.889	4	1.015	11	1.054
(a)	Content $\rightarrow$ average attn. (control)	5	1.015	7	**0.998**	4	**1.001**	11	**1.001**
(b)	All $\delta_{\text{norm}} < 0.9 \rightarrow$ circulant	13	1.141	20	1.156	27	**0.995**	64	**0.979**
(c)	Positional INT4	5	**0.995**	7	**1.004**	4	**0.999**	11	**1.002**
(c)	Content INT4 (control)	5	**0.991**	7	**1.003**	4	**1.000**	11	**0.997**
(c)	All heads INT4	72	1.066	144	1.276	384	1.032	720	1.023
(d)	Circulant pos. + INT8 rest	72	1.018	144	1.021	384	**0.999**	720	**0.998**

Table 4. Deletion plus replacement on GPT-2 Small (canonical seven-head set; eight-text held-out corpus). Random deletion is the control.

Strategy	PPL	Ratio	Heads	FLOPs
Baseline	43.52	1.000	0	—
δ_{norm} only (7 circ.)	44.48	1.022	7	3.2%
Michel only (20 del.)	41.46	**0.953**	20	13.9%
Comb. (20 del., 7 circ.)	42.65	**0.980**	27	17.1%
Random delete 20 (ctrl)	152.10	3.495	20	—

Mean Michel importance is nearly identical for the two classes (2.75% versus 2.89%); the distinction is what happens under *modification*, not removal. Positional heads survive replacement by a fixed pattern because the pattern is their function; ablation measures removal cost and is blind to this.

Complementarity, and attribution of the combined gains. Because the head sets are disjoint, the methods compose: Michel deletion removes heads whose contribution is dispensable; circulant replacement compresses heads whose contribution is a fixed pattern. Table 4 makes the attribution explicit: circulant replacement alone gives ratio 1.022 with 3.2% attention-FLOP savings and zero parameters; the improvement of the combined strategy (0.980) comes from the 20 Michel deletions (0.953 alone), and the gap between the two is the boundary-head circulant cost of Sec. 5. The contribution of this paper is the structural axis and its zero-cost exactness, not the raw size of the combination.

Downstream preservation. On SST-2 [15] with the compressed model as a frozen feature extractor and a logistic-regression probe, circulant replacement of the canonical positional heads gives 78.56% validation accuracy versus 79.13% baseline, within 0.6 points with a seven-head set that includes both boundary heads; the purpose of this experiment is to verify non-degradation beyond perplexity, not to claim a downstream gain. The Michel-deletion row uses the true ablation ranking of Sec. 6. Additional tasks extend the check: on AG-News, circulant replacement gives 84.70% versus 85.30% baseline, and on MRPC 67.16% versus 66.91%. Across all three tasks the replacement stays within 0.6 points of baseline while the 20 Michel deletions cost 0.5 to 3.2 points, consistent with replacement preserving function that deletion removes.

Table 5. Cross-architecture per-head classification with the identical diagnostic. "Best" is the smallest head-level δ_{norm}.

Model	Encoding	Heads	Pos.	Best δ_{norm}
OPT-125m	learned, causal	144	2	0.36
Pythia-70m	RoPE, causal	48	2	0.38
Pythia-160m	RoPE, causal	144	3	0.10
BERT-base	learned, bidir.	144	11	0.32
Phi-3-mini	RoPE, causal	1024	2	0.43
Gemma-2-2B	RoPE, causal	208	10	0.05
ModernBERT	RoPE, bidir.	264	37	0.10

7. STABILITY UNDER FINE-TUNING

We fine-tune GPT-2 Small on WikiText-2 for eight epochs (2000 sequences of 128 tokens, AdamW, lr 5×10^{-5}) and track the canonical classification at each epoch, using the canonical calibration corpus at $M = 64$. The pre-trained checkpoint reproduces Table 1 exactly (seven positional heads, identical δ_{norm} values), verifying the tracking setup. Across all nine checkpoints the U-shape is preserved and the middle-layer plateau is locked at $\delta_{\text{norm}} = 1.406$ to 1.407: eight epochs of gradient updates move it by less than 0.001. Six of the seven canonical positional heads retain their identity at every checkpoint with maximum drift 0.078 (L4H7); the strongest head L0H1 moves only from 0.163 to 0.182. The single exception is precisely the boundary head: L0H4, canonically at $\delta_{\text{norm}} = 0.467$, crosses the 0.5 threshold at epoch 2 and ends at 0.541. The output-side layer gradually weakens ($0.249 \rightarrow 0.389$) as the model acquires content-specific structure. Fine-tuning therefore neither creates nor destroys positional structure away from the class boundary, consistent with Sec. 4: classify once on the pre-trained model, compress once, fine-tune freely, revalidating only boundary heads.

8. CROSS-ARCHITECTURE GENERALITY

To demonstrate that the findings are not merely GPT-2 idiosyncrasies, Table 5 applies the identical canonical diagnostic across other positional-encoding schemes and attention directions.

We make three observations regarding the non-GPT-2 models. First, the positional/content bimodality is not specific to learned absolute positional embeddings: every RoPE model exhibits positional heads, and Gemma-2-2B contains the strongest structured head we have measured (L5H4, $\delta_{\text{norm}} = 0.049$), with ten positional heads distributed through the middle of the stack rather than concentrated

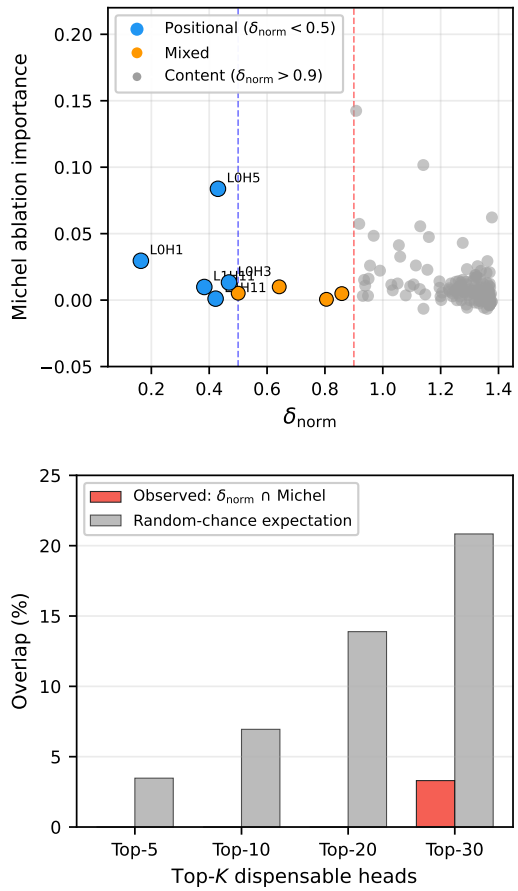

Fig. 2. Top: δ_{norm} versus Michel ablation importance for all 144 GPT-2 Small heads; the properties are uncorrelated in importance but disjoint in structure (two high-importance content heads, L0H0 and L0H10, are clipped by the axis for readability). Bottom: top-K overlap between the two rankings versus chance.

at the input. The census size is model-specific rather than encoding-specific: Phi-3-mini reproduces the U-shape with an essentially identical plateau ($\delta_{\text{norm}} = 1.409$ across 27 middle blocks) yet carries only two boundary-grade positional heads ($\delta_{\text{norm}} = 0.43$, 0.50), while Gemma-2-2B's hidden-state profile lacks the input shoulder entirely. The U-shape and depth trend are GPT-2-family regularities; head-level bimodality and the anti-structured middle persist across architectures. Second, attention direction controls the anti-structure: the two bidirectional models carry by far the largest censuses (BERT-base 11 of 144; ModernBERT 37 of 264, with layer-level $\delta_{\text{norm}} < 0.19$ at 16 of its 23 stack positions), and ModernBERT is RoPE-based, so the effect follows the mask and not the encoding. Third, replacement transfers: causal-circulant substitution of all ten identified Gemma-2-2B positional heads leaves perplexity unchanged (ratio 0.998), a progressive sweep in δ_{norm} order stays within 0.992 to 1.002 at every prefix, and the identical substitution applied to the ten highest-δ_{norm} content heads degrades perplexity by 12.8%. The selectivity of the diagnostic, not merely its tolerance, extends to a modern rotary-encoded, grouped-query-attention model.

Table 6. Measured greedy-decode metrics, NVIDIA GB10, float16, prefill 64 plus 960 decode steps; mean of 5 runs (SD $\leq$ 0.5 tok/s). Small combined = 20 Michel deletions plus 7 circulant heads; Large has no deletion component (circulant-only, 11 heads).

Configuration	Tokens/s	ms/tok	Mem. (GB)
Small baseline, b=1	124.8	8.01	0.91
Small combined, b=1	119.5	8.37	0.91
Small baseline, b=8	930.7	8.60	1.17
Small combined, b=8	910.0	8.79	1.14
Large baseline, b=1	37.0	27.01	5.32
Large circ.-only, b=1	35.1	28.53	5.33
Large baseline, b=8	223.3	35.83	6.62
Large circ.-only, b=8	212.1	37.72	6.65

9. DEPLOYMENT METRICS

Attention-FLOP savings do not automatically translate to wall-clock gains. Table 6 gives greedy-decode throughput, per-token latency, and peak memory for GPT-2 Small under the combined strategy (20 deletions plus 7 circulant heads) and GPT-2 Large under circulant-only replacement (11 heads; the Michel ranking exists only for Small), on an NVIDIA DGX Spark (GB10, float16, prefill 64 plus 960 decode steps, mean of 5 runs). The implementation realizes the savings structurally (deleted heads leave the projections and cache entirely; circulant heads carry no Q/K projection, scores, or K cache); its unmodified form reproduces native logits to 1.5×10^{-4}. Wall-clock does not improve: the modified models run between 2.2% and 5.1% slower across both scales and batch sizes (Table 6); repeated benchmark invocations shift these figures by one to two points while the within-invocation SD stays below 0.5 tok/s. The dispatch overhead of unfused per-head replacement is therefore not a small-model artifact, and realizing the removed FLOPs in wall-clock terms would require fused kernels. The measured deployment benefit is memory, and it belongs chiefly to the deletion component: the combined strategy shrinks the KV cache by 16.3% (75.5 to 63.2 MB at $T = 2048$), while circulant-only replacement on Large saves 0.8%; the saving grows with sequence length and batch size; deployment measurement on modern architectures such as Gemma-2 is future work.

10. DISCUSSION

Why the causal mask destroys cyclic symmetry. A circulant matrix depends only on relative position $i - j$; the causal mask zeros entries with $j > i$, breaking shift invariance and propagating anti-structure through the stack. The bidirectional results of Sec. 8 are the architectural contrast that confirms the mechanism: remove the mask and the anti-structure recedes, in a learned-absolute model (BERT-base) and in a RoPE model (ModernBERT) alike.

The anti-structure plateau. Middle-layer $\delta_{\text{norm}} \approx 1.4$ means trained representations are less commutative than random matrices. The identity $\|[\pi_\sigma, R]\|_F^2 = \sum_k (\lambda_k - \lambda_{\sigma(k)})^2$ explains this: training concentrates eigenvalue spectra, and large spectral gaps produce large commutators with any non-identity permutation. Invariance of the plateau at 1.40 ± 0.01 across model sizes, and its recurrence at 1.409 in Phi-3-mini, suggests a fixed point of the training dynamics for causal decoders.

Beyond the fixed circulant. Regarding the use of the fixed circulant target and its potential optimality, the fixed circulant is the

projection dictated by the cyclic group and Lemma 1, hence optimal among $\mathbb{Z}_M$-invariant substitutes, but the framework is not tied to it: every candidate group induces its own projection target (block-circulant substitutes for block-cyclic groups), and the mixed class invites *partial* substitution, a fixed average component plus a low-rank content-dependent residual. The diagnostic tells us which invariant subspace is nearest; richer residual parameterizations are extensions rather than competitors.

Complementarity, not subsumption. Probing classifiers [23, 24] report whether a representation contains a property, attention visualization [25] suggests positional behavior qualitatively, and mechanistic interpretability [26, 27] reverse-engineers circuits head by head; δ_{norm} instead reports whether the matrix's structure permits a low-parameter algebraic substitution, with the substitution rule attached and no per-circuit analysis. Fixed and synthetic attention patterns have been trained into models from scratch [20, 21], and circulant weight matrices have long been used for compression by construction [22]; the present work differs in direction, discovering post hoc which trained heads already are fixed patterns.

Scope and limitations. The quantitative compression results concern the GPT-2 family; the cross-architecture section establishes qualitative transfer and exact replacement transfer on Gemma-2-2B, but modern-model compression at scale, Mamba-class state-space models, and RoPE-specific diagnostics (rotation groups rather than permutation groups) remain open. Whole-model INT4 sensitivity decreases with size within the GPT-2 family and should be re-examined on modern compact architectures.

Conclusion. The commutator residual δ_{norm} is a structural, label-free diagnostic that identifies precisely the attention heads whose function is a fixed causal circulant, composes with ablation-based pruning because it targets a disjoint head set, survives fine-tuning, and transfers across positional encodings and attention directions. It provides a structural explanation for a class of compressibility that importance scores can only discover by search.

11. REFERENCES

[1] A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI Technical Report, 2019.

[3] P. Michel, O. Levy, and G. Neubig, "Are sixteen heads really better than one?" in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[4] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," in *Proc. ACL*, 2019.

[5] X. Ma, G. Fang, and X. Wang, "LLM-Pruner: On the structural pruning of large language models," in *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023.

[6] E. Frantar and D. Alistarh, "SparseGPT: Massive language models can be accurately pruned in one-shot," in *Proc. ICML*, 2023.

[7] M. Sun, Z. Liu, A. Bair, and J. Z. Kolter, "A simple and effective pruning approach for large language models," in *Proc. ICLR*, 2024.

[8] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "GPT3.int8(): 8-bit matrix multiplication for transformers at scale," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022.

[9] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "GPTQ: Accurate post-training quantization for generative pre-trained transformers," in *Proc. ICLR*, 2023.

[10] M. Xia, Z. Zhong, and D. Chen, "Structured pruning learns compact and accurate models," in *Proc. ACL*, 2022.

[11] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proc. ICML*, 2016.

[12] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *arXiv:2104.13478*, 2021.

[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv:1910.01108*, 2019.

[14] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," in *Proc. ICLR*, 2017.

[15] R. Socher *et al.*, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. EMNLP*, 2013.

[16] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, 2024.

[17] S. Biderman *et al.*, "Pythia: A suite for analyzing large language models across training and scaling," in *Proc. ICML*, 2023.

[18] S. Zhang *et al.*, "OPT: Open pre-trained transformer language models," *arXiv:2205.01068*, 2022.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019.

[20] Y. Tay, D. Bahri, D. Metzler, D.-C. Juan, Z. Zhao, and C. Zheng, "Synthesizer: Rethinking self-attention for transformer models," in *Proc. ICML*, 2021.

[21] W. You, S. Sun, and M. Iyyer, "Hard-coded Gaussian attention for neural machine translation," in *Proc. ACL*, 2020.

[22] C. Ding *et al.*, "CirCNN: Accelerating and compressing deep neural networks using block-circulant weight matrices," in *Proc. IEEE/ACM MICRO*, 2017.

[23] J. Hewitt and C. D. Manning, "A structural probe for finding syntax in word representations," in *Proc. NAACL-HLT*, 2019.

[24] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni, "What you can cram into a single vector: Probing sentence embeddings for linguistic properties," in *Proc. ACL*, 2018.

[25] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? An analysis of BERT's attention," in *Proc. ACL Workshop BlackboxNLP*, 2019.

[26] N. Elhage *et al.*, "A mathematical framework for transformer circuits," *Transformer Circuits Thread*, 2021.

[27] C. Olsson *et al.*, "In-context learning and induction heads," *Transformer Circuits Thread*, 2022.

[28] M. Abdin *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv:2404.14219*, 2024.

[29] Gemma Team, "Gemma 2: Improving open language models at a practical size," *arXiv:2408.00118*, 2024.

[30] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York: Springer, 1998.

[31] B. Warner *et al.*, "Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference," *arXiv:2412.13663*, 2024.