SPECTRAL SHAPE RESIDUAL: A LIGHTWEIGHT SPECTRAL DIAGNOSTIC FOR ATTENTION WEIGHT MATRICES

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce the **Spectral Shape Residual (SSR)**, a lightweight metric computed directly from the query and key weight matrices of attention layers—no forward pass, no training data needed. SSR measures how similar the singular value distributions of W_Q and W_K are: when they match perfectly, SSR is zero. We study SSR across four settings organized as a 2×2 analytical framework (toy vs. production LLM; static snapshot vs. training dynamics)—a design that systematically covers the conditions under which spectral weight-space signals can arise. In toy transformers trained on modular arithmetic, SSR falls sharply during training, dropping below 50% of its peak value 200–1,900 steps *before* the model generalizes—making SSR an early warning signal—and converges to near-machine precision once grokking is complete. In production LLMs (LLaMA-3-8B , Gemma-4-31B-IT), SSR decreases systematically with layer depth, and RL-distilled models (DeepSeek-R1-Distill-Qwen-14B) show lower SSR than their base model (Qwen2.5-14B-Instruct) in 45 out of 48 layers. In OLMoE-1B-7B pretraining, 18.8% of attention heads show persistent monotonic decline in SSR before grokking completes, concentrated in deep layers (L10–L13); the median SSR of these heads falls 40% below its peak 582B tokens before grokking completes. SSR requires only one SVD per attention head and no evaluation data, making it deployable as a zero-overhead structural health monitor directly inside training pipelines. For both pre-training (does the model grok?) and post-training (does RL alignment shift spectral structure?), SSR provides a continuous signal without running a single forward pass.

1 INTRODUCTION

Training a large language model is expensive and structurally opaque: practitioners watch loss curves and benchmarks, but these give little insight into whether attention heads are learning well or whether query and key projections are developing compatible representations— questions that matter for generalization but usually require a forward pass to answer—which costs time and compute.

We ask a simpler question: *can the weight matrices themselves tell us something useful, without any forward pass?*

We propose the **Spectral Shape Residual (SSR)**: the mean absolute difference between the ℓ_2-normalized singular value spectra of W_Q and W_K in each attention head (formal definition in Section 3, Eq. 1). SSR costs one SVD per head—negligible compared to any forward pass.

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.

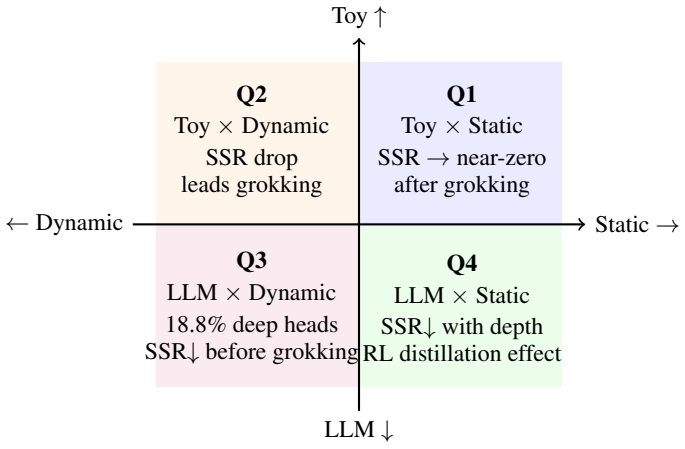

Figure 1: The 2×2 experimental framework. **Horizontal axis**: Dynamic (left) vs. Static (right). **Vertical axis**: LLM (bottom) vs. Toy (top). Q1 (top-right) is the simplest setting; Q3 (bottom-left) the most complex. Paper presents results in order Q1→Q2→Q4→Q3.

What we find. Studying SSR across four experimental settings, we establish four findings:

1. **Q1 (Toy × Static).** After grokking on modular arithmetic, SSR converges to near-machine precision ($\leq 10^{-5}$) across real and complex domains.
2. **Q2 (Toy × Dynamic).** SSR drop precedes grokking by 200–1,900 training steps across three random seeds, providing a leading indicator of generalization.
3. **Q3 (LLM × Dynamic).** In OLMoE-1B-7B pretraining, 18.8% of attention heads (concentrated in deep layers L10–L13) show a persistent decline in SSR before grokking completes; the median SSR of these heads falls 40% below its peak 582B tokens (139k steps) before grokking completes.
4. **Q4 (LLM × Static).** SSR decreases systematically with layer depth in LLaMA-3-8B and Gemma-4-31B-IT— consistent with the view that deeper layers handle higher-order semantic processing (Geva et al., 2021). RL distillation (DeepSeek-R1-Distill-Qwen-14B vs. Qwen2.5-14B-Instruct) further reduces SSR in 45/48 layers, suggesting RL training aligns Q/K spectra.

We present Q1→Q2→Q4→Q3 (simplest to most complex; Figure 1), and find that SSR is a *structural fingerprint* of Q/K coordination—observable from weights alone.

2 Background

Transformer self-attention (Vaswani et al., 2017) computes $\mathrm{softmax}(QK^\top/\sqrt{d_h})V$, where $W_Q, W_K \in \mathbb{R}^{d_h \times d}$ project inputs into query and key spaces; for the attention pattern to be meaningful, Q and K must live in compatible spaces. *Grokking* (Power et al., 2022) refers to delayed generalization: a model memorizes the training set, then—sometimes thousands of steps later—suddenly generalizes. Recent work confirms grokking in LLM pretraining (Li et al., 2026; Lv et al., 2025); Saxe et al. (2019) provide theoretical grounding via singular value dynamics in linear networks.

3 Spectral Shape Residual

3.1 Definition

For a single attention head with query weight $W_Q \in \mathbb{R}^{d_h \times d}$ and key weight $W_K \in \mathbb{R}^{d_h \times d}$, let $s_q = (s_{q,1}, \ldots, s_{q,d_h})$ and $s_k = (s_{k,1}, \ldots, s_{k,d_h})$ be their singular values in descending order.

Define the normalized spectra:

$$\tilde{s}_{q,i} = \frac{s_{q,i}}{\|s_q\|_2}, \qquad \tilde{s}_{k,i} = \frac{s_{k,i}}{\|s_k\|_2}.$$

The **Spectral Shape Residual** is:

$$\mathrm{SSR}(W_Q, W_K) \;=\; \frac{1}{d_h} \sum_{i=1}^{d_h} \left| \tilde{s}_{q,i} - \tilde{s}_{k,i} \right|. \tag{1}$$

Normalization by $\|s\|_2$ (the Euclidean norm: $\|s\|_2 = \sqrt{\sum_i s_i^2}$) removes the overall scale of W_Q and W_K, so SSR measures *shape* differences only—not magnitude. When W_Q and W_K have identical spectral shapes, $\mathrm{SSR} = 0$.

3.2 Why Spearman ρ for Trend Detection

When tracking SSR over training, we use Spearman's rank correlation ρ between SSR values and training step indices to quantify monotonicity. We deliberately avoid Pearson r because SSR curves are often monotone but nonlinear (fast early, slow late), and Pearson r would underestimate the trend strength. Spearman ρ measures only the direction of change, not the rate, making it robust to the log-shaped decay common in training curves.

3.3 GQA Pseudo-Bulk Aggregation

Modern LLMs use Grouped Query Attention (GQA) (Ainslie et al., 2023), where multiple query heads share a single key head. For example, LLaMA-3-8B has 32 query heads and 8 key heads per layer: every 4 consecutive query heads share one key head, forming 8 groups.

Naively averaging SSR across all 32 heads would be misleading, because the 4 query heads in each group all compare against the *same* key matrix. We therefore use a two-step **pseudo-bulk** aggregation:

1. **Within each group**: for each query head, compute its SSR against the shared key head, then take the *median* across the group. This gives one representative SSR value per KV group.
2. **Across groups**: take the *median* of all group-level values to get one SSR per layer (25th/75th percentiles as uncertainty bands).

This mirrors pseudo-bulk analysis in single-cell genomics (Squair et al., 2021), where cells are grouped before aggregation to avoid pseudoreplication. For standard MHA (e.g., OLMoE where $n_Q = n_{KV}$), step 1 is a no-op and step 2 reduces to a plain layer median.

4 First Quadrant (Q1): Toy × Static

Setup. We train one-layer transformers on modular arithmetic ($a \circ b \pmod p$, $p = 113$) with weight decay $\lambda = 1.0$. For the real-valued domain we use standard transformers; for the complex-valued domain we use complex-valued transformers on the modular arithmetic ($a + b \bmod 113$). Architectural details are given in Appendix A. We report SSR at the final training checkpoint after grokking is complete.

Results. Figure 2 shows that SSR converges to near-machine precision after grokking in both domains. In the real domain, final SSR $\leq 4.3 \times 10^{-5}$ (mean over 3 seeds). In the complex domain, final SSR $\leq 2 \times 10^{-6}$, approximately 20× lower, approaching the 10^{-6} floating-point floor.

Finding Q1: *After grokking, SSR converges to near-zero. The attention heads' Q and K weight matrices develop nearly identical spectral shapes—a geometric signature of generalization.*

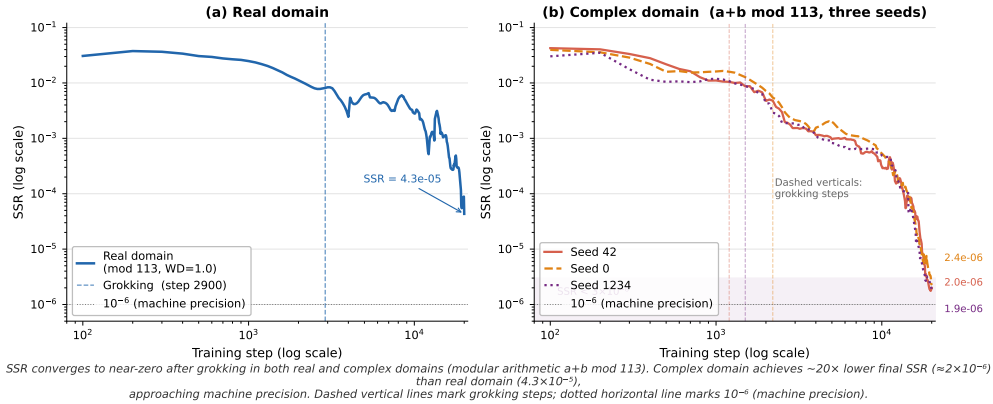

Figure 2: **Q1: SSR after grokking in toy models.** (a) Real domain: SSR drops to 4.3×10^{-5} after grokking (step 2,900, dashed vertical). (b) Complex domain (a+b mod 113, three seeds): SSR drops to $\leq 2 \times 10^{-6}$, approaching machine precision. Both axes log-scale. Dotted horizontal: 10^{-6} reference.

5 Second Quadrant (Q2): Toy × Dynamic

Setup. In the real-valued domain, using the same models as Q1. We record SSR at every training checkpoint and identify two events: (1) *SSR drop step*: the first step at which SSR falls below 50% of its peak value (`ssr_drop_frac`$= 0.5$); (2) *grokking step*: the first step at which validation accuracy exceeds 95%. The *leading margin* is (grokking step) − (SSR drop step).

Results. Figure 3 shows SSR and validation accuracy over training for three seeds. In all three cases, SSR drops significantly before generalization occurs. Leading margins: seed 42: 1,700 steps; seed 0: 1,900 steps; seed 1234: 200 steps.

Finding Q2: *SSR drop precedes grokking by 200–1,900 steps across all three seeds. SSR is a leading indicator, not a lagging one.*

The 50% threshold (`ssr_drop_frac`) is used consistently in Q2 and Q3 to ensure comparability across settings.

6 Fourth Quadrant (Q4): LLM × Static

Setup. We compute per-head SSR for four production models from their released weight checkpoints: LLaMA-3-8B (32 layers, GQA 32Q/8KV), Gemma-4-31B-IT (60 layers, sliding-window + global attention), Qwen2.5-14B-Instruct (48 layers, GQA 40Q/8KV), and DeepSeek-R1-Distill-Qwen-14B (same architecture, RL-distilled from Qwen). All models use GQA; we apply pseudo-bulk aggregation (Section 3).

SSR decreases with layer depth. Figure 4(a,b) shows that SSR decreases with layer depth in both LLaMA-3-8B and Gemma-4-31B-IT. In LLaMA-3-8B, a linear fit gives slope $= -0.03 \times 10^{-3}$/layer (Pearson $r = -0.921$). In Gemma-4-31B-IT, global attention layers (every 6th, 32Q/4KV) show systematically different SSR from adjacent sliding-window layers (16Q/8KV), reflecting their different architectural roles.

RL distillation effect. Figure 4(c) compares SSR layer-by-layer between DeepSeek-R1-Distill-Qwen-14B and its base model Qwen2.5-14B-Instruct. DeepSeek-R1-Distill-Qwen-14B shows lower SSR in **45/48 layers** (93.75%), with a median ratio of 0.9872 (see Table 6 in Appendix B for per-layer details). This suggests that RL training—which optimizes for reasoning quality—incidentally aligns Q/K spectral shapes more tightly.

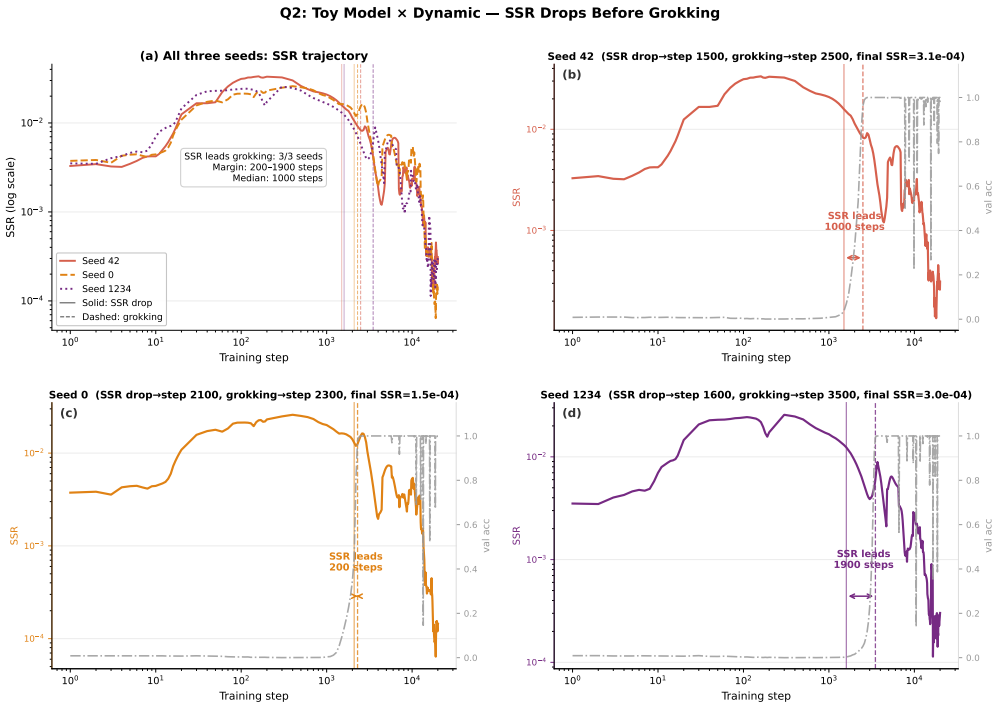

Figure 3: **Q2: SSR leads grokking in toy models.** (a) Overview: SSR curves for three seeds. Solid verticals: SSR drop steps. Dashed verticals: grokking steps. (b–d) Per-seed detail: SSR (colored, left axis) and validation accuracy (gray, right axis). Double-headed arrow: leading margin.

Finding Q4: *SSR decreases with layer depth in production LLMs, and RL distillation further reduces SSR in 45/48 layers. Deeper, more capable processing appears to require more aligned Q/K spectra.*

7 Third Quadrant (Q3): LLM × Dynamic

Why OLMoE? Q3 requires a model with (i) many public pretraining checkpoints and (ii) an independently verified grokking transition. OLMoE-1B-7B-0924 (Muennighoff et al., 2024) uniquely satisfies both, with 244 checkpoints and grokking confirmed by Li et al. (2026). Its 16 layers and 16 heads per layer ($n_Q = n_{KV}$, standard MHA) leave **12 layers (L02–L13) × 16 heads = 192 heads** after excluding embedding/output-projection layers (L00–L01, L14–L15).

When does grokking happen? Li et al. (2026) track benchmark performance across four domains (math reasoning, code generation, commonsense QA, domain-specific QA). To stay consistent with the val_acc $> 95\%$ criterion used to define grokking completion in Q1 and Q2, we read off the pretraining step at which each domain's benchmark accuracy first exceeds 95% from their Figure 1 (Table 1). Commonsense is the only domain that completes earlier, at step 980k; math, code, and domain-specific QA all remain below 95% until the final reported checkpoint, step 1100k. We therefore use a single **grokking completion point**, step 1100k, rather than a window: this is the step at which all four domains have demonstrably reached the 95% threshold. Converting via OLMoE's published checkpoint naming (`step1200000` = 5,033B tokens, i.e. 4.19M tokens/step), step 1100k corresponds to **4,614B tokens**.

We sample our 20 checkpoints deliberately rather than uniformly: 17 checkpoints (20B–4,613B) precede grokking completion, and 3 checkpoints (4,823B–5,117B) serve as a post-grokking reference. This design concentrates resolution on the pre-grokking period, where any leading signal must be detected.

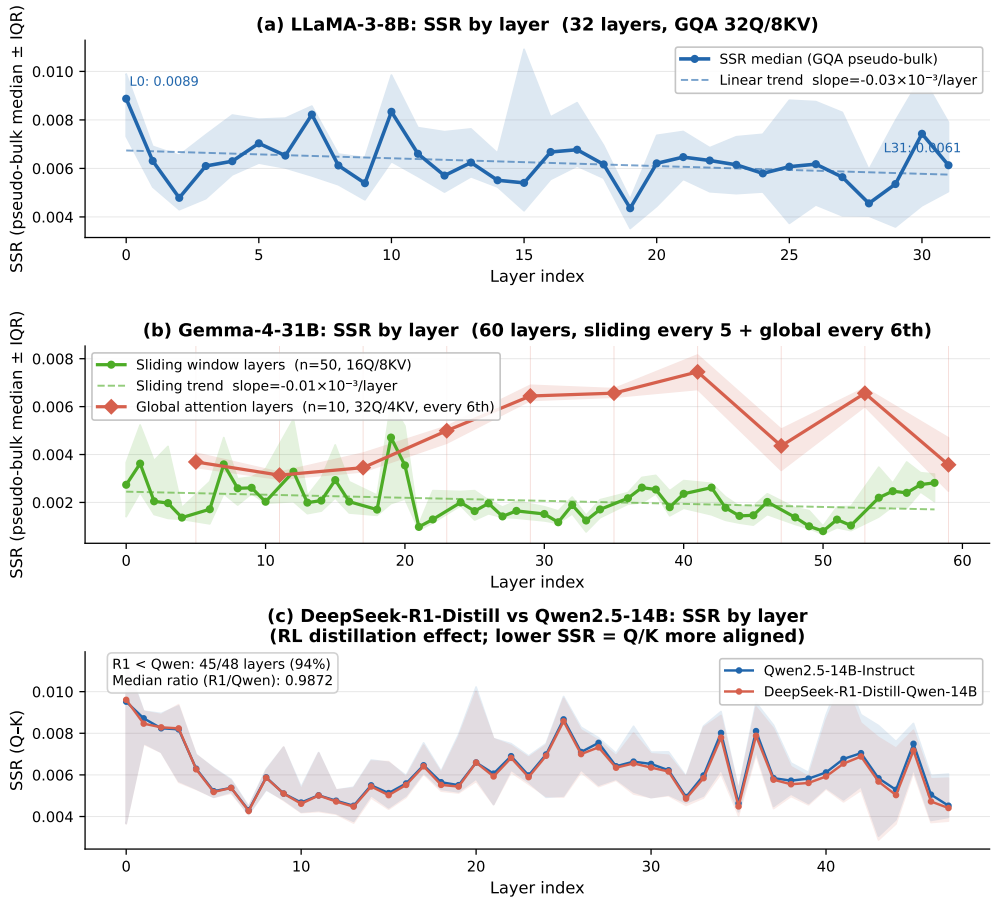

Figure 4: **Q4: SSR by layer in production LLMs.** GQA pseudo-bulk: within each KV-head group, Q-head SSR values are first aggregated by median; then KV-group medians are aggregated per layer (shaded band = IQR). (a) LLaMA-3-8B: SSR decreases with depth (slope $= -2.1 \times 10^{-4}$/layer). (b) Gemma-4-31B-IT: sliding-window (green) and global attention layers (red diamonds, every 6th layer) show distinct SSR levels.
(c) DeepSeek-R1-Distill-Qwen-14B vs. Qwen2.5-14B-Instruct: RL distillation reduces SSR in 45/48 layers (median ratio $= 0.987$).

Table 1: Per-domain step at which benchmark accuracy first exceeds 95%, read from Li et al. (2026) Figure 1.

Domain	**Step (k)**	**Tokens (B)**
Commonsense	980	4,110
Domain-specific QA	1100	4,614
Math	1100	4,614
Code	1100	4,614
Grokking completion (all domains)	**1100**	**4,614**

Why analyze heads individually, not by layer? Within a single layer, SSR varies enormously: the highest-SSR head can be 3–5$\times$ larger than the lowest. The coefficient of variation (CV $= \sigma/\mu$) averages 0.35–0.87 (median 0.45) across layers and checkpoints—the heads are not a homogeneous population (Voita et al., 2019; Michel et al., 2019). We therefore classify every head individually before any aggregation.

How are heads classified? Given that heads within the same layer behave so differently, we classify each head individually—and we must do so without looking at the future. A head's SSR is a sequence of 20 values, one per checkpoint. If we measure a head's trend using its *entire* sequence (including checkpoints after grokking) and then claim that trend "led" grokking, that argument is circular.

We avoid this by defining two separate Spearman correlations. First, ρ_{t17}: the rank correlation between SSR and training tokens, computed using *only* the first 17 checkpoints (20B–4,613B), all of which strictly precede grokking completion (step 1100k). A head with $\rho_{t17} < -0.5$ shows a clear downward trend using only information available before grokking completes—this is a genuine leading signal, not hindsight. Second, ρ_{final}: the same correlation computed over all 20 checkpoints, which confirms whether the trend persists rather than reversing after grokking. A head qualifies as **persistently Q/K-aligning** only if both $\rho_{t17} < -0.5$ *and* $\rho_{\text{final}} < -0.5$. **36 heads** (18.8% of 192) meet this criterion.

Applying the same $\rho_{t17}/\rho_{\text{final}}$ logic to all 192 heads yields a complete six-way classification (Table 2) that covers every head exactly once:

Table 2: Six-way head classification using pre-grokking (ρ_{t17}) and full-sequence (ρ_{final}) Spearman correlations. Covers all 192 heads exactly once.

Class	n	**%**
Aligning, persistent ($\rho_{t17}<-0.5$, $\rho_{\text{final}}<-0.5$)	36	18.8
Aligning, late-emerging ($\rho_{t17}\geq-0.5$, $\rho_{\text{final}}<-0.5$)	2	1.0
Diverging, persistent ($\rho_{t17}>+0.5$, $\rho_{\text{final}}>+0.5$)	74	38.5
Diverging, late-emerging ($\rho_{t17}\leq+0.5$, $\rho_{\text{final}}>+0.5$)	11	5.7
Reversal (sign flip, $\lvert\rho\rvert>0.5$ both sides)	0	0.0
Noisy / stable (both $\rho \in [-0.5, +0.5]$)	69	35.9

Two observations follow. No head reverses sign: once a head's SSR trend is established before grokking completes, it does not flip direction afterward, which supports SSR as a stable rather than noisy signal. Second, persistently diverging heads (74) outnumber persistently aligning heads (36) by more than 2:1—most heads with a clear trend are moving *away* from Q/K alignment, and only a minority align early.

The 36 persistently aligning heads are not evenly spread across layers: deep layers L10–L13 contain 15 of them (3.75 per layer, 4 layers), versus 21 in shallow layers L02–L09 (2.62 per layer, 8 layers)—a $1.4\times$ density difference, with L10 alone contributing 6 heads ($2.0\times$ the random baseline of 3.0/layer).

This raises a natural concern: are we selectively reporting a few specific heads? Table 3 shows the head-index distribution (positions 0–15 within each layer) of the 36 heads: 14 of 16 possible indices appear, with no single index dominating. The pattern is driven by layer depth, not by any particular head position.

Table 3: Head-index distribution (0–15) of the 36 persistently aligning heads.

Index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Count	1	3	3	1	1	4	2	2	2	3	0	3	3	3	0	5

How far in advance does SSR move? Classification alone tells us *whether* a head's trend precedes grokking, but not *how much advance warning* SSR provides. To quantify this, we examine the median SSR curve of the 36 persistently aligning heads across all 20 checkpoints and ask: at what point does this curve fall to a fixed fraction below its peak? Q1 and Q2 use a 50% threshold (`ssr_drop_frac`$= 0.5$), but the Q3 median curve never reaches a 50% decline—its largest drop from peak, reached only at the final checkpoints, is 40.9%. We therefore use the largest threshold the curve can actually reach while staying below the Q1/Q2 standard: **40%**. The median curve's peak occurs early (104B tokens); it crosses 40% below that peak at **4,032B tokens**, by linear interpolation between adjacent checkpoints—**582B tokens (139k steps) before grokking completes** at 4,614B (step 1100k).

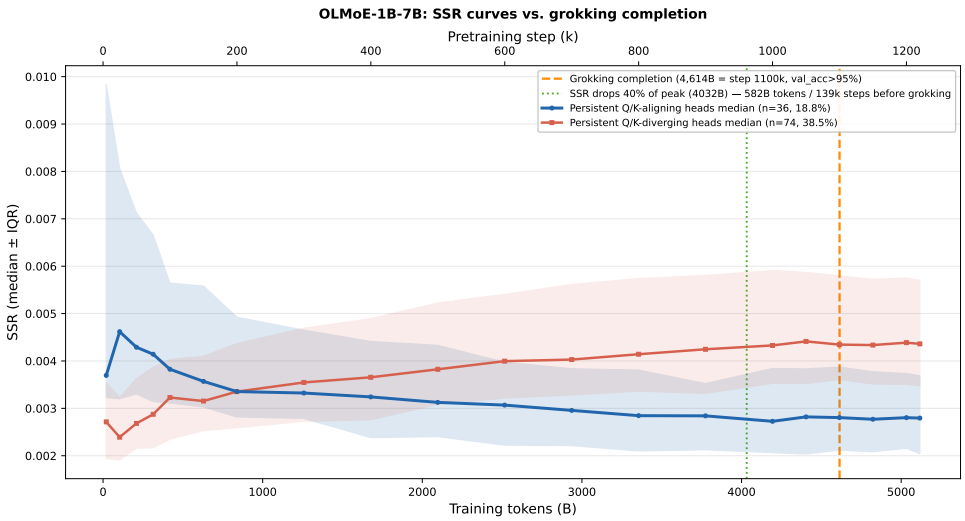

Figure 5: **Q3: SSR curves vs. grokking completion in OLMoE-1B-7B pretraining.** Bottom axis: training tokens (B); top axis: pretraining step (k), aligned via 4.19M tokens/step. Blue: median SSR of the 36 persistently Q/K-aligning heads (IQR band shaded). Red: median SSR of the 74 persistently Q/K-diverging heads (IQR band shaded). Orange dashed line: grokking completion (4,614B tokens, step 1100k, val_acc $> 95\%$; Li et al. 2026). Green dotted line: the point at which the aligning-head median curve falls 40% below its peak (4,032B tokens)—582B tokens (139k steps) before grokking completes.

Figure 5 plots the median SSR curve (with IQR band) of the persistently aligning and persistently diverging populations across all 20 checkpoints, with grokking completion and the 40% crossing point marked.

Finding in Q3: *Using only pre-grokking data (checkpoints through 4,613B tokens), 36 of 192 attention heads (18.8%) in OLMoE-1B-7B already show a statistically robust, persistent decline in SSR—meaning their Q and K weight matrices are converging toward shared spectral shape before grokking completes at 4,614B tokens (step 1100k). The median SSR curve of these 36 heads falls 40% below its peak at 4,032B tokens, 582B tokens (139k steps) before grokking completes. This decline is concentrated in deep layers L10–L13* ($1.4\times$ *the density of shallow layers) and is not an artifact of which head positions were selected. SSR therefore carries a genuine pre-grokking structural signal, detectable from weights alone.*

Scope and limitations. This analysis covers one model and one grokking event. OLMoE's grokking is also local and asynchronous across domains (Li et al., 2026), while SSR is computed from weight matrices that are global—every head participates in every forward pass regardless of input domain—so we cannot attribute any individual head's curve to a specific downstream task. Public pretraining checkpoints are also a blunt instrument: sampling intervals are uneven and the total checkpoint count is limited compared to what a dedicated tracking run could provide. We therefore treat the Q3 finding as an exploratory result rather than a load-bearing claim: it shows that a pre-grokking weight-space signal *can* exist, not that it generalizes across architectures or training runs. Confirming this pattern on additional open-source pretraining training runs, as they become available, is the natural next step.

8 Discussion

SSR = 0 as a theoretical lower bound. By definition, $\mathrm{SSR} = 0$ when W_Q and W_K have identical spectral shapes. In toy models after grokking, SSR approaches this bound ($\leq 10^{-5}$). In production LLMs, SSR ranges from 0.002 to 0.010—orders of magnitude higher. This gap may reflect the fact that production LLMs must handle diverse tasks simultaneously, preventing full Q/K spectral alignment. SSR could serve as a measure of how far a model is from the "fully generalized" regime.

Null result: Pythia without grokking. We also analyzed five Pythia models (160M–2.8B parameters) trained on The Pile. Unlike OLMoE, Pythia has no identified grokking transition (grokking in LLMs has so far been observed only in specific settings (Lv et al., 2025; Li et al., 2026)). Consistently, SSR increases monotonically throughout training in all five models (median Spearman $\rho = +0.83$ to $+0.88$ across scales), with zero Q/K-aligning heads. This null result strengthens our main claim: *SSR convergence is a grokking-specific structural signal, not a generic property of pretraining*. The contrast between OLMoE (18.8% aligning heads) and Pythia (0%) confirms that SSR-leading behavior requires a grokking-type phase transition.

Future work. Immediate directions include combining SSR with runtime attention patterns, using SSR to actively regulate training, extending SSR to W_V and cross-attention, and establishing a formal connection to generalization bounds.

9 Related Work

Grokking and mechanistic interpretability. Power et al. (2022) first documented delayed generalization; subsequent work studied its mechanistic basis (Nanda et al., 2023; Liu et al., 2022; Merrill et al., 2023; Humayun et al., 2024). Geva et al. (2021) showed FFN layers act as key-value memories, with deep layers encoding semantics—consistent with our Q3 finding that SSR-leading heads concentrate in deep layers. A rich literature studies what individual heads compute (Olsson et al., 2022; Voita et al., 2019; Michel et al., 2019); SSR offers a complementary weight-space view requiring no input data.

Spectral methods and RL fine-tuning. Spectral properties of weight matrices have been studied for generalization bounds (Bartlett et al., 2017), stability (Pennington et al., 2017), pruning (Yang et al., 2020), and implicit regularization (Martin & Mahoney, 2021); low-rank structure underlies LoRA (Hu et al., 2022) and GaLore (Zhao et al., 2024). Our work is the first to study the *relative* spectral structure between paired matrices (W_Q and W_K). DeepSeek-AI (2025) showed RL training improves reasoning; our Q4 result adds a spectral perspective: RL distillation systematically reduces SSR.

10 Conclusion

We introduced the Spectral Shape Residual (SSR), a lightweight metric computed from attention weight matrices. Across four experimental settings—toy vs. production LLMs, static snapshots vs. training dynamics—SSR reveals consistent geometric structure: it converges to near-zero after grokking, leads generalization by hundreds to thousands of steps, decreases with layer depth in production models, and is reduced by RL distillation.

Beyond the specific findings, the 2×2 framework (toy/LLM $\times$ static/dynamic) provides a reusable scaffold for mechanistic interpretability research: toy models establish what *must* happen theoretically, static snapshots reveal what *is* present in deployed models, and dynamic tracking reveals *when* structural change occurs during training.

SSR requires only one SVD per attention head and no evaluation data, making it deployable as a zero-overhead structural health monitor directly inside training pipelines—no forward pass, no evaluation set required. In practice, SSR supports three use cases: (1) *training health monitoring*—if SSR stops decreasing while training loss is still high, the model may be stuck before grokking; (2) *model quality screening*—abnormally high SSR in deep layers relative to shallow layers indicates potential attention degradation, detectable from weights alone; and (3) *fine-tuning comparison*—SSR profiles can distinguish fine-tuning strategies without running downstream benchmarks, as demonstrated by the R1 vs. Qwen comparison in Q4. Whether the question is "is pre-training grokking?" or "is RL alignment shifting the spectral structure of attention?", SSR provides a continuous answer from the weights alone.

AI USE DISCLOSURE

We used generative AI tools (Claude, Anthropic) to implement methods and to assist with coding and plotting for the experiments described in this paper, to assist in interpreting experimental results, to identify relevant literature, and to edit the manuscript for readability and language.

We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, contribute to proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean or reformat datasets, or support qualitative or thematic data analysis.

All AI-assisted code, literature searches, and language edits were reviewed and verified by the author.

We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Because SSR requires only a per-head singular value decomposition of the W_Q/W_K weight matrices—no forward pass and no evaluation data—every experiment in this paper, including the analysis of billion-parameter production models, was conducted without GPU acceleration (Q1/Q2 toy models on a local machine; Q3/Q4 production model weight extraction on a CPU-only cloud instance and a local machine with P106 GPU).

Toy transformers used in Q1 and Q2 are cheap to retrain from scratch (minutes on CPU); hyperparameters and all three random seeds are listed in Appendix A (Table 4). Production models analyzed in Q3 and Q4 are publicly released checkpoints; exact model identifiers and revisions are listed in Appendix A (Table 5). Because obtaining full checkpoints for Q4 requires downloading several to tens of gigabytes per model, we additionally release the precomputed per-head SSR values underlying every table and figure in this paper (Tables 2, 3, 6; Figures 2–5) as small CSV files in the Supplementary Material, so that the aggregation, statistical, and classification analyses can be independently verified without re-downloading any model weights. For Q3, checkpoints of OLMoE-1B-7B-0924 are instead scanned via targeted HTTP range requests that fetch only the per-layer query/key projection tensors, avoiding a full checkpoint download altogether.

Code implementing SSR computation is provided in Supplementary Material.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zelaski, Sumit Sanghai, and Yukun Fu. GQA: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2022.

Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Deep networks always grok and here is why. *arXiv preprint arXiv:2402.15555*, 2024.

Ziyue Li, Chenrui Fan, and Tianyi Zhou. Grokking in LLM pretraining? monitor memorization-to-generalization without test. In *International Conference on Learning Representations (ICLR)*, 2026.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35, 2022.

Ang Lv, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Rui Yan. Language models "grok" to copy. *arXiv preprint arXiv:2409.09281*, 2025.

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

William Merrill, Nikolaos Tsilivis, and Aman Shukla. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. *arXiv preprint arXiv:2303.11873*, 2023.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Niklas Muennighoff et al. OLMoE: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*, 2024.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *Advances in neural information processing systems*, 30, 2017.

Alethea Power, Yuri Burda, Harri Edwards, Ian Goodfellow, and Vineet Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23): 11537–11546, 2019.

Jordan W Squair, Matthieu Gautier, Claudia Kathe, et al. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1):5692, 2021.

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*, 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*, pp. 5797–5808, 2019.

Ruiqi Yang, Guiguang Ding, Jungong Han, and Ji Li. Learning to prune filters in convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Animashree Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024.

APPENDIX

A MODEL CONFIGURATIONS

This appendix provides supplementary details supporting the main text.

Table 4 lists the hyperparameters of the toy transformers used in Q1 and Q2 (Sections 4 and 5). Both domains are trained on modular arithmetic with modulus $p = 113$ and share all optimization hyperparameters; the complex domain uses $d_{\text{model}} = 64$ (complex-valued) so that its real-equivalent parameter count matches the real domain's $d_{\text{model}} = 128$.

Table 4: Hyperparameters of the toy transformers used in Q1 and Q2.

Hyperparameter	**Real domain**	**Complex domain**
Modulus p	113	113
Layers	1	1
Attention heads n_{heads}	4	4
d_{model}	128	64 (complex; real-equiv. 128)
d_h	32	16
FFN expansion	$4\times$	$4\times$
Optimizer	AdamW (β_1=0.9, β_2=0.999)	AdamW
Learning rate	1×10^{-3}	1×10^{-3}
Weight decay	1.0	1.0
Batch size	512	512
Max training steps	20,000	20,000
Train fraction	0.4	0.4
Seeds	42, 0, 1234	42, 0, 1234

Table 5 lists the architectural parameters of all production models analyzed in Q3 and Q4.

Table 5: Architectural parameters of production LLMs used in Q3 and Q4.

Model	**Layers**	**Q**	**KV**	$d_{\textbf{model}}$	d_h	**Notes**
OLMoE-1B-7B-0924	16	16	16	2048	128	MHA; MoE FFN (8 experts)
LLaMA-3-8B	32	32	8	4096	128	GQA; group size 4
Gemma-4-31B-IT	60	–	–	5376	–	Sliding (16Q/8KV) + Global (32Q/4KV) every 6th layer
Qwen2.5-14B-Instruct	48	40	8	5120	128	GQA; group size 5
DeepSeek-R1-Distill-Qwen-14B	48	40	8	5120	128	Same arch as Qwen2.5-14B-Instruct

B R1 VS. QWEN2.5 LAYER-BY-LAYER SSR COMPARISON

Table 6 reports the per-layer SSR pseudo-bulk median for DeepSeek-R1-Distill-Qwen-14B and Qwen2.5-14B-Instruct, along with the absolute difference and percentage change. Negative values indicate R1 has lower (more aligned) SSR than Qwen.

Table 6: Per-layer SSR comparison: DeepSeek-R1-Distill-Qwen-14B vs. Qwen2.5-14B-Instruct (GQA pseudo-bulk median). Green = R1 lower SSR (more aligned); Red = R1 higher.

L	R1	Qwen	Diff	(%)	L	R1	Qwen	Diff	(%)
0	0.0096	0.0095	+0.0001	+0.9%	24	0.0069	0.0070	-0.0001	-0.9%
1	0.0085	0.0087	-0.0003	-2.9%	25	0.0086	0.0087	-0.0001	-1.1%
2	0.0083	0.0082	+0.0000	+0.5%	26	0.0070	0.0071	-0.0001	-1.4%
3	0.0082	0.0082	+0.0000	+0.6%	27	0.0073	0.0075	-0.0002	-2.8%
4	0.0063	0.0063	-0.0000	-0.5%	28	0.0063	0.0064	-0.0001	-1.1%
5	0.0052	0.0052	-0.0000	-0.8%	29	0.0066	0.0066	-0.0001	-1.3%
6	0.0054	0.0054	-0.0000	-0.0%	30	0.0064	0.0065	-0.0002	-2.7%
7	0.0043	0.0043	-0.0000	-0.8%	31	0.0062	0.0062	-0.0001	-1.0%
8	0.0059	0.0059	-0.0000	-0.3%	32	0.0049	0.0049	-0.0001	-1.5%
9	0.0051	0.0051	-0.0000	-0.1%	33	0.0058	0.0060	-0.0001	-2.2%
10	0.0046	0.0047	-0.0001	-1.3%	34	0.0078	0.0080	-0.0002	-2.8%
11	0.0050	0.0050	-0.0000	-0.2%	35	0.0045	0.0046	-0.0002	-3.3%
12	0.0047	0.0048	-0.0000	-0.8%	36	0.0079	0.0081	-0.0002	-2.8%
13	0.0045	0.0045	-0.0000	-0.9%	37	0.0058	0.0058	-0.0001	-1.3%
14	0.0054	0.0055	-0.0001	-1.0%	38	0.0056	0.0057	-0.0002	-3.0%
15	0.0050	0.0051	-0.0001	-1.9%	39	0.0056	0.0058	-0.0002	-3.6%
16	0.0055	0.0056	-0.0001	-1.3%	40	0.0059	0.0061	-0.0002	-3.2%
17	0.0064	0.0065	-0.0000	-0.7%	41	0.0065	0.0068	-0.0002	-3.3%
18	0.0055	0.0056	-0.0001	-2.3%	42	0.0069	0.0070	-0.0002	-2.3%
19	0.0054	0.0055	-0.0001	-1.3%	43	0.0057	0.0058	-0.0002	-2.7%
20	0.0066	0.0066	-0.0000	-0.0%	44	0.0050	0.0053	-0.0003	-4.8%
21	0.0059	0.0061	-0.0001	-2.2%	45	0.0072	0.0075	-0.0003	-4.4%
22	0.0068	0.0069	-0.0001	-1.2%	46	0.0047	0.0051	-0.0003	-6.5%
23	0.0059	0.0060	-0.0001	-1.4%	47	0.0044	0.0045	-0.0001	-2.5%