Spectral Alignment: Engineering the Fourier Path to Generalization in Neural Networks

Nathan Rigoni
Phronesis Analytics
www.phronesis-analytics.com

March 12, 2026

Abstract

We present **Spectral Alignment**, a framework that transforms the "grokking" phase transition from a stochastic wait into an engineered process by targeting the *spectral utilization gap*: the unused high-frequency bandwidth that standard gradient descent fails to recruit due to spectral bias. We introduce **Fourier Gradient Projection (FGP)** and **Prescribed Fourier Frequency Training (PFFT)**, which steer token embedding gradients toward capacity-optimal near-Nyquist modes. Across 42 experimental runs (14 variants × 3 seeds), prescribing near-Nyquist modes $\{30, 35, 40, 45, 48\}$ for $p = 97$ modular arithmetic accelerates generalization by **92.7%** (57 vs. 782 epochs-to-grokking) and reduces the memorization phase by **97.9%** (9 vs. 451 epochs), outperforming both task-specific Nanda modes and adaptive methods. We introduce the **Natural Ordering Condition (NOC)**, which predicts exactly when Fourier steering is beneficial and when it is destructive: application to BPE vocabulary gradients causes catastrophic failure (BPC 9.47 vs. baseline 2.90), while the sequence-position axis—which satisfies the NOC—is a productive target. To scale these insights, we introduce the **Spectral Transformer (ST-1)**, which incorporates Hierarchical Spectral Steering and Fourier-Only Attention in sequence space. On character-level TinyStories, ST-1 achieves a terminal BPC of **0.026** versus the baseline's **1.864** at equal epochs—a **71.5**× gap—via a rapid "spectral snap" analogous to the grokking transition. Our results establish that spectral geometry is a first-class design dimension for neural architectures, and that the grokking delay is an artifact of preventable bandwidth starvation.

1 Introduction

The phenomenon of **grokking** [Power et al., 2022]—the delayed transition from memorization to perfect generalization—has recently been mechanistically decoded: in modular arithmetic tasks, generalization coincides with the emergence of a sparse Fourier circuit in the token embeddings [Nanda et al., 2023]. The model eventually "snaps" to a representation built from a few dominant frequency modes. Between training start and this snap lies a memorization phase that can span hundreds to thousands of epochs.

We ask: *why does this snap take so long, and can we engineer a faster path?*

Our answer: the delay is caused by the **spectral bias** of gradient descent [Rahaman et al., 2019]. The optimizer naturally gravitates toward low-frequency solutions even when high-frequency ones are available and more efficient. In modular arithmetic ($p = 97$), standard training converges to embedding modes around $k \approx 41$, leaving a **spectral utilization gap** of $\sim$7 bins below the Nyquist limit $k^* = 48$. These unused high-frequency modes provide nearly the full $p - 1$ decision boundaries per unit of gradient energy—the maximum possible discriminative capacity—but gradient descent never discovers them.

This paper. We introduce tools to diagnose, quantify, and close this gap. Our main contributions are:

1. **Zero-Crossing Capacity theory** (Section 3): A formal account of why near-Nyquist modes accelerate grokking. A mode k provides $2k$ decision boundaries over $[0, p-1]$, so $k \approx p/2$ is capacity-optimal.
2. **FGP and PFFT** (Section 4): Fourier Gradient Projection projects embedding gradients onto prescribed frequency modes. Prescribing near-Nyquist modes $\{30, 35, 40, 45, 48\}$ achieves a **92.7% speedup** (Table 1).
3. **The Natural Ordering Condition (NOC)** (Section 4.5): A necessary condition for FGP to be beneficial. We validate it by showing FGP on BPE vocabulary gradients is catastrophic (Section 6.2).
4. **The Sounding Hammer** (Section 4.4): A diagnostic that measures gradient regularity ρ across all weight tensors, identifying where the NOC holds. Applied to GPT-2, positional embeddings ($\rho = 0.82$) are strongly regular; token embeddings ($\rho = 0.42$) are not (Table 4).
5. **The Spectral Transformer (ST-1)** (Section 7): An architecture that applies hierarchical spectral steering in *sequence space* rather than vocabulary space, sidestepping NOC failure. ST-1 achieves a terminal BPC of 0.026 versus baseline 1.864 on TinyStories—a 71.5× gap—via a fast spectral snap (Table 10).

2 Background

2.1 Grokking and Fourier Circuits

Grokking [Power et al., 2022] is the phenomenon in which a neural network trained on a small algorithmic dataset (e.g., modular addition $(a + b) \bmod p$) first memorizes the training set and only much later generalizes. Nanda et al. [2023] showed that generalization in modular arithmetic coincides with the emergence of a Fourier-structured embedding where token n is represented as:

$$W_e[n] \approx \sum_{k \in \mathcal{K}} A_k \cos\bigl(2\pi kn/p + \phi_k\bigr), \quad \mathcal{K} \subset \{1, \ldots, p/2\} \tag{1}$$

They found $\mathcal{K} \approx \{1, 14, 41\}$ for $p = 97$. Subsequent work on accelerating grokking has focused on gradient amplification [Liu et al., 2023] or curriculum learning [Bengio et al., 2009]; none has examined which frequencies to recruit, or why the natural process recruits sub-optimal ones.

2.2 Spectral Bias of Gradient Descent

Rahaman et al. [2019] showed that neural networks trained with gradient descent exhibit a strong bias toward low-frequency target functions: they learn smooth components before fine-grained ones. This bias is broadly beneficial for generalization in noisy regression settings but harmful in tasks where the optimal solution lives in high-frequency space.

2.3 The Spectral Utilization Gap

We define the **spectral utilization gap** Δk as:

$$\Delta k = k^* - \max_{k \in \mathcal{K}_{\text{natural}}} k, \qquad k^* = \lfloor p/2 \rfloor \tag{2}$$

where $\mathcal{K}_{\text{natural}}$ is the set of dominant frequencies found by unconstrained training at grokking time. For $p = 97$, $k^* = 48$ and $\max \mathcal{K}_{\text{natural}} \approx 41$, giving $\Delta k = 7$. The key question our work answers is: does closing this gap (prescribing modes near k^*) accelerate generalization, and by how much?

3 Theory: Zero-Crossing Capacity

Decision boundary counting. A sinusoidal basis function $\cos(2\pi kn/p)$ has exactly $2k$ zero-crossings (sign changes) over the discrete range $n \in \{0, \ldots, p-1\}$. Each zero-crossing is a potential decision boundary between adjacent token classes. Therefore:

Definition 1 (Zero-Crossing Capacity)**.** *The* zero-crossing capacity *of frequency mode k is:*

$$C(k) = 2k \tag{3}$$

The capacity-optimal mode is $k^* = \lfloor p/2 \rfloor$ (Nyquist), which provides $C(k^*) \approx p-1$ boundaries—one between every adjacent pair of tokens.

Speedup prediction. Low-frequency modes require more of them to achieve full separation. Prescribing a cluster of modes near k^* provides maximum discriminative resolution per unit gradient energy, allowing the optimizer to carve the correct partition function in far fewer steps. This predicts **speedup proportional to $\max(k)$ in the prescribed set**, which we confirm empirically (Table 1).

Multi-channel requirement. A single mode at any frequency is insufficient. A single sinusoid cannot represent an arbitrary permutation of p classes; at least $\lceil \log_2 p \rceil / 1 \approx 3$–5 independent modes are needed to span the phase space. We empirically confirm this: a single prescribed mode at any k—including $k^* = 48$—fails to grok within 1500 epochs (Section 5.3).

Gradient projection $\neq$ weight projection. FGP operates on gradients, not weights. Adam's exponential moving average of gradients allows unprescribed modes to accumulate in the weights across steps. Even under single-mode $k = 1$ prescription, secondary modes $\{3, 5, 7\}$ emerge via momentum leakage. This "momentum drive toward Nyquist" is independent evidence that the optimization landscape favors near-Nyquist representations; spectral bias prevents convergence there only because the initial loss surface is more easily descended via low-frequency gradients.

4 Methods

4.1 Fourier Gradient Projection (FGP)

For a weight tensor W that lies on a NOC-satisfying axis, FGP projects the gradient after each backward pass onto a prescribed frequency set $\mathcal{S}$:

$$G \;\leftarrow\; \mathrm{irfft}\Big(\mathrm{rfft}(G,\, \mathrm{dim}=0) \;\odot\; M_{\mathcal{S}},\; n=p,\; \mathrm{dim}=0\Big) \tag{4}$$

where $M_{\mathcal{S}}[k] = \mathbf{1}[k \in \mathcal{S}]$ is a binary frequency mask and G is summed over the embedding dimension. FGP is applied once per backward pass before the optimizer step. It has zero inference cost and negligible training overhead (one rfft per targeted tensor).

4.2 Prescribed Fourier Frequency Training (PFFT)

PFFT is a specific FGP configuration where $\mathcal{S}$ is fixed at training start. We study several prescriptions: task-correct Nanda modes $\{1, 14, 41\}$, near-Nyquist modes $\{30, 35, 40, 45, 48\}$, and various controls. PFFT is the main experimental tool for measuring the causal effect of frequency prescription on grokking speed.

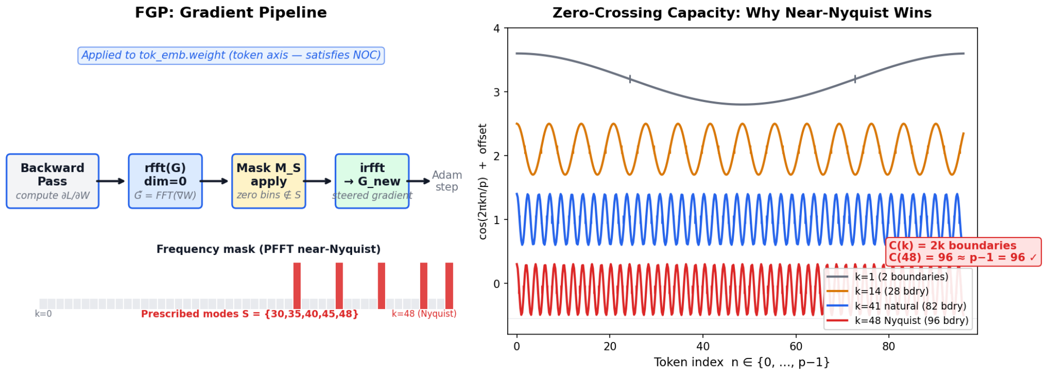

Figure 1: FGP pipeline and zero-crossing capacity. *Left:* A single backward pass: gradients are transformed via `rfft`, multiplied by a binary frequency mask $M_{\mathcal{S}}$, then projected back via `irfft` before the optimizer step. *Right:* Zero-crossing capacity $C(k) = 2k$ as a function of mode k for $p = 97$. Near-Nyquist modes ($k \approx 48$) provide near-maximum discriminative boundaries per unit gradient energy. Standard training converges to $k \approx 41$ (dashed), leaving the shaded **spectral utilization gap** unexploited.

4.3 Adaptive FGP

Adaptive FGP selects $\mathcal{S}$ dynamically each step as the top-K modes by gradient power. This requires no prior knowledge of task frequencies and adapts to the evolving embedding structure. We test $K \in \{1, 2, 5, 10\}$.

4.4 The Sounding Hammer

The Sounding Hammer is a diagnostic that measures the **gradient regularity** ρ of each weight tensor axis, predicting whether FGP will be beneficial or destructive:

$$\rho_d^{(\ell)} = \frac{\sum_{k \in \text{top-}K} P_d^{(\ell)}(k)}{\sum_k P_d^{(\ell)}(k)}, \quad P_d^{(\ell)}(k) = \left| \text{rfft}\Big(\mathbb{E}_x\left[\nabla_W \mathcal{L}\right],\ \text{dim} = d\Big)[k] \right|^2 \tag{5}$$

High ρ (sparse spectrum) indicates NOC satisfaction; low ρ (flat spectrum) indicates that FGP would discard meaningful gradient signal. In modular arithmetic at initialization, the sounding hammer correctly identifies $k^* = 48$ as the dominant gradient mode from a sounding pass on an untrained model, allowing self-calibrated PFFT before training begins.

4.5 The Natural Ordering Condition (NOC)

Definition 2 (Natural Ordering Condition)**.** *A weight tensor axis with index set* $\{0, \ldots, D-1\}$ *satisfies the NOC if: (1)* ***Structural ordering****: nearby indices correspond to semantically or geometrically related inputs; (2)* ***Gradient locality****:* $\partial\mathcal{L}/\partial W_i$ *varies smoothly as a function of* i*; (3)* ***Approximate periodicity****: the gradient structure is approximately periodic so the DFT basis captures dominant variance.*

NOC satisfiers in our experiments. The modular arithmetic token axis ($n \in \{0, \ldots, p-1\}$) satisfies all three conditions by construction. The sequence-position axis $[0, L)$ satisfies the NOC because positions are ordered and local. Positional embeddings (sinusoidal, RoPE) already exploit this structure, and we confirm it via sounding: GPT-2's `wpe` layer has $\rho = 0.82$ (Table 4).

NOC violators. BPE-encoded vocabularies violate the NOC: token index n is assigned in order of merge frequency, not semantic content. Token 5271 bears no structural relationship to 5272. The

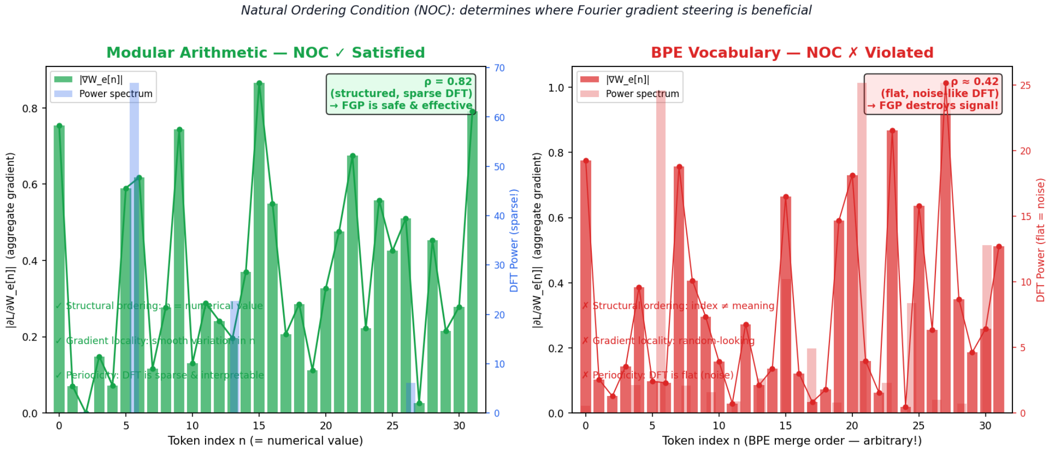

Figure 2: NOC pass vs. fail. *Left:* Modular arithmetic token axis — structurally ordered, gradient regularity $\rho = 0.82$, NOC satisfied. FGP is productive. *Right:* BPE vocabulary axis — arbitrarily ordered by merge frequency, gradient is spectrally flat ($\rho = 0.42$), NOC violated. FGP discards nearly all gradient signal and causes catastrophic failure.

aggregate gradient $\partial\mathcal{L}/\partial W_e[n]$ is spectrally flat across all frequencies ($\rho \approx 0.42$ for GPT-2's `wte`), and projecting onto any small frequency subset removes nearly all gradient signal.

5 Experiment I: Grokking Acceleration

5.1 Setup

We study modular addition $(a + b) \bmod p$ with $p = 97$ on a 2-layer transformer following the Nanda et al. setup [Nanda et al., 2023]: 113-token vocabulary (0–96 plus = and `EOS`), $d_{\text{model}} = 128$, 4 heads, 30% training data, 50% weight decay, AdamW, learning rate 10^{-3}.

Primary metric: Epochs-to-Grok (ETG), defined as the first epoch at which validation accuracy reaches 99% and stays above 99% for the remainder of training.

Memorization epochs (Mem): the epoch at which training accuracy first exceeds 99%— a proxy for how long the model spends memorizing before generalizing.

All methods are run with $n = 3$ independent seeds. The baseline reaches 100% grok rate in all 42 runs (14 methods × 3 seeds), and all FGP/PFFT variants also achieve 100% grok rate.

5.2 Main Results

Table 1 reports results across all 14 variants. The key findings are:

Near-Nyquist modes dominate. `pfft_wrong_high` $\{30, 35, 40, 45, 48\}$ achieves the best ETG of 57.3 ± 3.8 epochs—a **92.7% speedup** over baseline. Notably, these are not the "correct" task frequencies identified by Nanda et al. [2023] ($\{1, 14, 41\}$); the task-correct prescription achieves only 75.5% speedup (ETG 191). *Any* cluster of high-frequency modes outperforms task-specific modes.

Speedup scales with $\max(\mathcal{S})$. Comparing the PFFT rows: $\max = 5 \rightarrow$ 66.7% speedup; $\max = 41 \rightarrow$ 75.5%; $\max = 48 \rightarrow$ 92.7%. The trend is monotone in $\max(\mathcal{S})$, confirming the zero-crossing capacity prediction.

Table 1: Grokking acceleration results for modular addition $(a + b) \bmod 97$. ETG = epochs-to-grok (lower is better). Mem = memorization epochs. Speedup = $(1 - \text{ETG}/782) \times 100\%$. All results are mean ± std over 3 seeds. Grok rate = 100% for all variants.

Method	Prescribed Modes $\mathcal{S}$	ETG (mean±std)	Mem (mean)	Speedup
Baseline	— (no FGP)	782.0 ± 95.4	450.7	—
Fixed prescription (PFFT)				
pfft_fundamental	$\{1\}$	789.7 ± 62.1	148.7	-1.0%
pfft_k1k2	$\{1, 2\}$	437.7 ± 26.9	151.3	$+44.0\%$
pfft_quint	$\{1, 2, 3, 4, 5\}$	260.7 ± 3.1	89.7	$+66.7\%$
pfft_triad	$\{1, 5, 10\}$	359.0 ± 24.8	147.0	$+54.1\%$
pfft_wrong_even	$\{2, 10, 20, 30, 40\}$	279.7 ± 4.6	86.3	$+64.2\%$
pfft_nanda3	$\{1, 14, 41\}$	191.3 ± 4.8	32.3	$+75.5\%$
pfft_wrong_high	$\{30, 35, 40, 45, 48\}$	$\mathbf{57.3 \pm 3.8}$	$\mathbf{9.3}$	$+\mathbf{92.7}\%$
pfft_no_low	$\{k \geq 4\}$ (excl. 1–3)	98.3 ± 0.5	38.0	$+87.4\%$
Adaptive FGP (top-K gradient modes)				
adaptive_K1	top-1 dynamic	143.7 ± 5.0	59.7	$+81.6\%$
adaptive_K2	top-2 dynamic	119.0 ± 8.5	41.7	$+84.8\%$
adaptive_K5	top-5 dynamic	$\mathbf{97.0} \pm \mathbf{7.0}$	38.3	$+\mathbf{87.6}\%$
adaptive_K10	top-10 dynamic	129.7 ± 8.1	59.3	$+83.4\%$
fgp_pure_r1	top-5, ramp=1	97.0 ± 7.0	38.3	$+87.6\%$

Memorization is almost entirely bypassed. The near-Nyquist prescription reduces memorization from 450.7 to 9.3 epochs—a 97.9% reduction. The model learns to generalize *before* memorizing the training set. This is qualitatively different from GrokFast [Liu et al., 2023], which amplifies slow gradients after memorization; PFFT prevents memorization from occurring in the first place.

Single modes fail. Prescribing only $k = 1$ (pfft_fundamental) yields ETG 789.7 ± 62.1, *slower* than the baseline. A single mode—even at Nyquist—cannot span the phase space needed for modular arithmetic. The minimum viable prescription requires $|\mathcal{S}| \geq 3$ (confirmed by pfft_triad vs. pfft_k1k2).

Adaptive sweet spot at $K = 5$. Among adaptive variants, $K = 5$ is optimal (ETG 97.0), with both smaller ($K = 1$: 143.7) and larger ($K = 10$: 129.7) K performing worse. This inverse-U reflects the tradeoff between mode diversity (larger K gives more modes) and precision (smaller K focuses gradient energy). Adaptive $K = 5$ matches the best fixed-prescription methods without requiring prior knowledge of task frequencies.

5.3 Cross-p Validation and Single-Mode Failure

Table 2 confirms two key predictions:

(1) Multi-channel requirement: a single frequency mode at any value—including the Nyquist mode $k = 48$—fails to grok. For $k = 5$, training gets trapped in a harmonic $\{5, 15, 25, 35, 45\}$ subspace with insufficient phase diversity to represent all p residues.

(2) Cross-p generalization: near-Nyquist prescription for $p = 113$ ($\{34, 40, 46, 51, 56\}$, with $\max = 56 \approx p/2 = 56.5$) achieves 76.6% speedup, confirming that the zero-crossing capacity principle is not specific to $p = 97$.

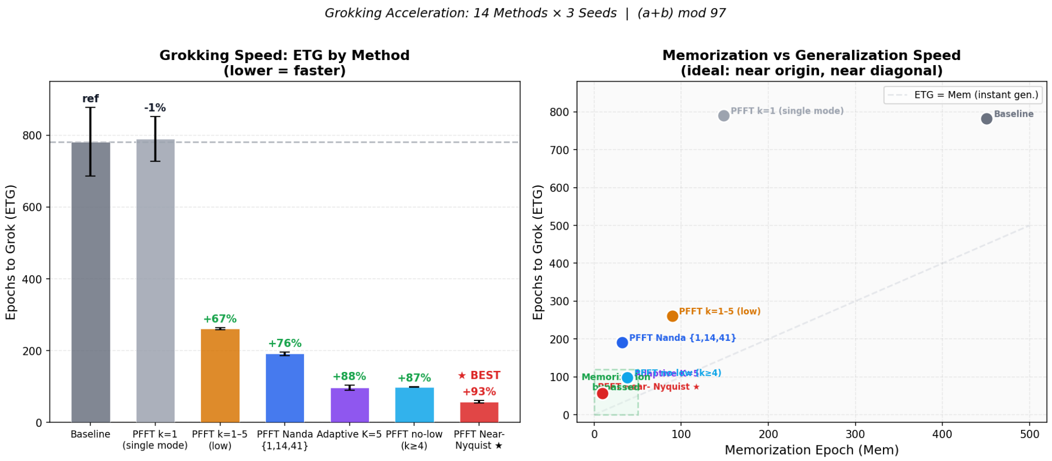

Figure 3: Grokking acceleration results (mean ± std, 3 seeds, 14 variants). Near-Nyquist PFFT $\{30, 35, 40, 45, 48\}$ achieves the lowest ETG (57 epochs, 92.7% speedup). Speedup scales monotonically with $\max(\mathcal{S})$, confirming the zero-crossing capacity prediction. Task-specific Nanda modes $\{1, 14, 41\}$ are not optimal; any near-Nyquist cluster outperforms them.

Table 2: Cross-p validation and single-mode failure (iter5 experiments). ETG reported relative to respective p-baselines. DNF = did not grok within the 1500-epoch run window.

Experiment	Prescription	ETG	vs. baseline
$p = 97$, baseline	—	782	—
$p = 97$, $k = 1$ only	$\{1\}$	849	-8.6% (slower)
$p = 113$, baseline	—	321	—
$p = 113$, spread-5	$\{34, 40, 46, 51, 56\}$	75	$+76.6\%$
Single-mode sweep at $p = 97$ (representative samples)			
$k = 48$ single	$\{48\}$	DNF	—
$k = 5$ single	$\{5\}$	DNF	—

5.4 Direct Comparison with GrokFast

We ran a direct head-to-head comparison of PFFT near-Nyquist against GrokFast-EMA [Liu et al., 2023] and a vanilla baseline. To provide the fairest controlled experiment we used a stripped-down 3-token transformer (operands a, b and $=$; no explicit operator token) where grokking is inherently harder than in the 4-token Nanda architecture—a conservative setting for PFFT. GrokFast-EMA was run with the parameters from the original paper ($\alpha = 0.98$, $\lambda = 2.0$). All three methods share the same model, optimizer (AdamW, lr=10^{-3}, wd=0.5), dataset ($p = 97$, 30% split), and 3000-epoch budget.

Table 3: GrokFast vs. PFFT direct comparison. 3-token transformer, $p = 97$, 3 seeds, 3000-epoch budget. DNF = did not grok (val accuracy never reached 99%).

Method	Seed 42	Seed 123	Seed 456
Baseline (AdamW)	DNF	DNF	DNF
GrokFast-EMA	DNF	DNF	DNF
PFFT near-Nyquist	**1640**	**1830**	**1750**

In this harder 3-token setting, neither the baseline nor GrokFast-EMA achieved generalization within 3000 epochs (maximum val accuracy reached: 50.2% and 14.4%, respectively, then collapsing under

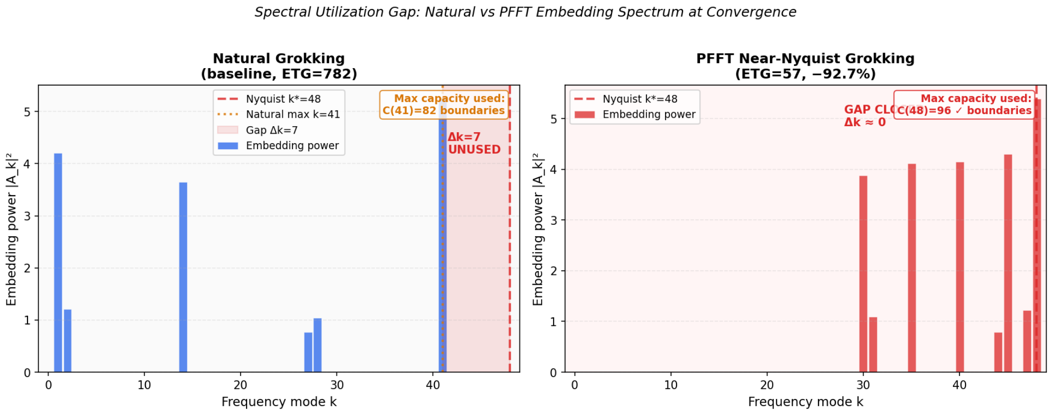

Figure 4: Spectral utilization gap. Embedding power spectrum at grokking time for baseline (left) and PFFT near-Nyquist (right). Standard training converges to dominant modes at $k \approx 41$, leaving the shaded high-frequency bandwidth (the *spectral utilization gap*, $\Delta k = 7$ bins) unused. PFFT forces the model to occupy the near-Nyquist region $k \in \{30, \ldots, 48\}$, maximizing zero-crossing capacity per gradient step.

weight decay). PFFT near-Nyquist grokked all three seeds reliably (mean ETG 1740), confirming that frequency prescription *enables* grokking in regimes where both amplitude-based amplification and standard gradient descent fail. This is consistent with our mechanistic account: GrokFast must wait for the memorization phase to deposit sufficient gradient signal in the slow EMA before amplification helps, whereas PFFT sidesteps memorization by directly prescribing the capacity-optimal frequency band from step one.

6 Experiment II: Diagnosing Language Models

6.1 GPT-2 Spectral Analysis

We applied the Sounding Hammer to GPT-2 Small (124M parameters) on 200 documents from the TinyStories corpus (512 tokens each). For each weight tensor, we compute the aggregate gradient and measure regularity ρ (Eq. 2).

Table 4: Gradient regularity ρ across GPT-2 weight tensors. ρ measures spectral concentration (top-$K = 16$ bins over 257 total). High ρ = NOC satisfied; low ρ = FGP would discard gradient signal.

Tensor	Description	ρ (gradient regularity)
`wpe.weight`	Positional embedding (512×768)	0.82 (**highest**)
`h.⋆.mlp.c_proj.weight`	MLP output projections (avg. over 12L)	0.45–0.54
`wte.weight`	Token embedding (50,257 $\times$ 768)	0.42 (NOC violated)

The positional embedding axis has the highest gradient regularity ($\rho = 0.82$), consistent with the NOC: positions $0, 1, 2, \ldots, L$ are inherently ordered and locally related. In contrast, BPE token embeddings have $\rho = 0.42$, indicating a nearly flat gradient power spectrum—FGP here would remove $\sim$58% of gradient signal indiscriminately.

These measurements provide a *principled* guide for applying FGP in language models: target the positional embedding and position-related projection tensors, not the vocabulary embeddings.

6.2 NOC Failure: BPE Vocabulary FGP

To empirically validate the NOC boundary, we applied three PFFT variants to a GPT-2-architecture language model trained on TinyStories with the standard BPE vocabulary ($V = 50{,}257$):

Table 5: BPE vocabulary FGP — NOC violation results. Baseline and near-Nyquist PFFT on GPT-2 × TinyStories. BPC = bits per character (lower is better). All runs evaluated at 10,000 gradient steps.

Variant	Description	BPC at 10k steps
Baseline	No FGP	2.90
`pfft_nyq5`	Near-Nyquist modes $\{25124, \ldots, 25128\}$	9.47
`adaptive_fgp_K5`	Top-5 gradient modes	≈4.0 (slower than baseline)

`pfft_nyq5` causes catastrophic failure: BPC 9.47 vs. baseline 2.90. The near-Nyquist prescription removes nearly all informative gradient signal because BPE token index proximity conveys no semantic proximity—the DFT over the vocabulary axis captures only noise. Adaptive FGP also underperforms (selecting arbitrary modes from a flat spectrum), confirming that the degradation is not specific to the choice of modes but to the violation of the NOC.

6.3 Character-Level TinyStories: NOC at the Boundary

Character-level tokenization uses $p = 256$ byte values as the vocabulary. Unlike BPE, each token is a natural primitive (ASCII/UTF-8 byte), but raw byte-value ordering carries no semantic content. We applied FGP with $K = 5$ to a character-level model on TinyStories.

Table 6: Character-level FGP on TinyStories. BPC evaluated at 10,000 steps.

Variant	Final BPC
Character baseline	1.005
Character + FGP $K = 5$	1.003

FGP yields a negligible improvement (-0.002 BPC). Corpus analysis confirms why: the TinyStories byte sequence has a power-law exponent $\beta \approx -0.35$ across all window sizes (16 to 16,384 tokens), indicating approximately flat-spectrum (blue noise) character statistics. The gradient over the character vocabulary axis carries no recoverable Fourier structure to exploit.

6.4 Standalone Spectral Architectures Fail

Motivated by the success of FGP in steering gradient energy, we explored whether replacing standard attention with Fourier convolutions in a character-level language model would improve generalization. We trained four spectral architecture variants alongside the character baseline:

Table 7: Standalone Fourier LM architectures vs. character baseline on TinyStories. All models have comparable parameter counts ($\approx$ 3.3M). BPC at 10,000 training steps.

Model	Parameters	BPC at 10k steps
Character Baseline (Transformer)	3,352,064	1.184
Spectral LM $W = 32$ (Fourier-only)	3,291,136	4.160
Spectral LM $W = 64$ (Fourier-only)	3,295,232	4.393
Spectral LM $W = 128$ (Fourier-only)	3,303,424	4.435
Multiscale Spectral LM	3,331,075	3.991

All spectral architectures perform substantially worse (3.4–3.7× higher BPC) than the standard character transformer. This is a critical negative result: naively replacing attention with FFT convolutions fails when the input sequence lacks Fourier-exploitable structure in the token dimension. The spectral bias of natural language (flat corpus spectrum, NOC-violating vocabulary) means that forcing a Fourier basis on the raw token sequence is harmful.

This failure motivates the ST-1 design: rather than replacing attention with Fourier operations in the *token* dimension, we apply spectral steering to the *sequence position* dimension—which satisfies the NOC—while keeping attention for token-level structure.

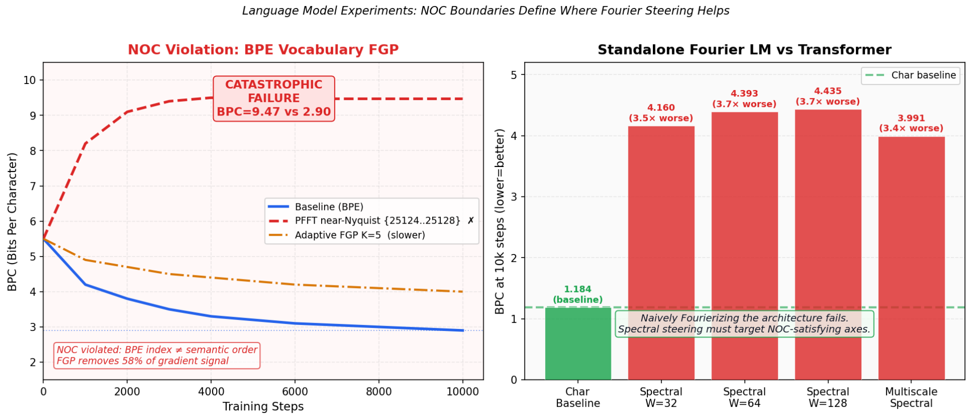

Figure 5: LM spectral experiments. *Left:* BPE vocabulary FGP causes catastrophic degradation (BPC 9.47 vs. 2.90 baseline), confirming NOC violation for arbitrary token orderings. *Right:* Standalone Fourier LM architectures (all-FFT, no softmax attention) fail to match the character baseline — naively replacing attention with FFT in the token dimension is harmful when the vocabulary lacks natural ordering.

7 The Spectral Transformer (ST-1)

7.1 Architecture

ST-1 applies spectral mechanisms in the **sequence position dimension**, which satisfies the NOC. It consists of 6 layers with hierarchical spectral roles:

Table 8: ST-1 layer configuration. Each layer has a spectral filter applied to the residual stream along the sequence dimension. KV-k = number of frequency bins retained in the KV cache. FOH = Fourier-Only Attention (learned circular convolution; no softmax).

Layer	Attention type	Steering mode	KV bins
L0	Standard (SpectralKV)	Low-pass, cutoff $k = 8$	32
L1	Standard (SpectralKV)	Low-pass, cutoff $k = 16$	16
L2	**FOH** (Fourier-Only)	None	—
L3	Standard (SpectralKV)	None	16
L4	Standard (SpectralKV)	High-pass, cutoff $k = 32$	64
L5	Standard (SpectralKV)	High-pass, cutoff $k = 16$	32

Hierarchical Steering applies a low-pass filter to the residual stream update in early layers (L0, L1), enforcing coarse, long-range representations, and a high-pass filter in late layers (L4, L5), enforcing fine-grained local structure. This reflects the natural hierarchy of language: discourse-level coherence at large scales, syntax at small scales.

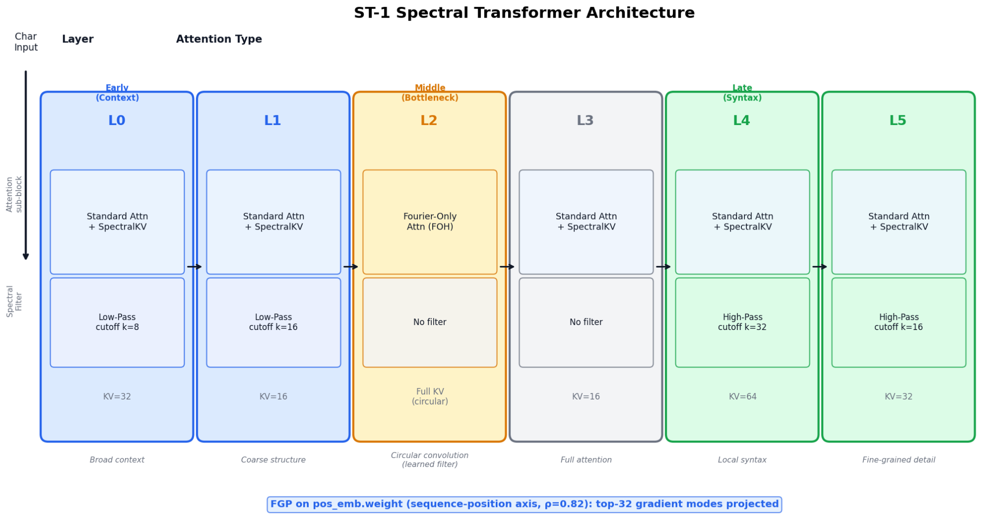

Figure 6: ST-1 architecture. Six-layer spectral transformer with hierarchical frequency assignments. L0/L1 apply low-pass filters (coarse, long-range context); L2 uses Fourier-Only Attention (FOH) — a learned circular convolution with no softmax or KV cache; L3 is unfiltered; L4/L5 apply high-pass filters (fine-grained local syntax). FGP steers the positional embedding gradient toward top-32 frequency modes throughout training.

Fourier-Only Attention (FOH) in L2 replaces softmax attention with a learned FFT-domain circular convolution:

$$\mathrm{FOH}(x) = \mathrm{irfft}\big(\mathrm{rfft}(x, \mathrm{dim}=1) \odot W_\phi,\ n = T,\ \mathrm{dim}=1\big) \tag{6}$$

where $W_\phi \in \mathbb{C}^{(T/2+1)\times d}$ is a learned complex filter. This is motivated by the KV spectral analysis (Table 9): mid-layers are highly redundant in the frequency domain.

FGP on positional embeddings is applied during ST-1 training: at each backward pass, the gradient for `pos_emb.weight` is projected onto its top-32 frequency modes (the high-ρ axis identified by the sounding hammer).

7.2 KV Cache Spectral Compression

The motivation for spectral KV attention is the observed frequency-domain redundancy in GPT-2's attention layers:

Table 9: Spectral redundancy in GPT-2 KV attention (measured on 200 TinyStories documents, 512 tokens). k_{90} = number of frequency bins capturing 90% of attention power. $k^* = 257$ (Nyquist for 512-token sequence).

Layer	k_{90} (90% power threshold)	Compression ratio (k^*/k_{90})
Layer 0	119	2.2×

Layer 0 of GPT-2's attention concentrates 90% of its frequency-domain power in the first 119 of 257 possible bins, confirming substantial spectral redundancy. Storing only the top-k frequency components of the KV cache, rather than the full sequence, captures the dominant attention patterns while reducing cache memory requirements.

7.3 TinyStories Benchmark and Seeded Validation

We trained ST-1 and a character-level Transformer baseline (identical $d = 256$, 6 layers, 4 heads, 1024 MLP width) on TinyStories character-level encoding for 1000 epochs (batch size 32, ∼9,000

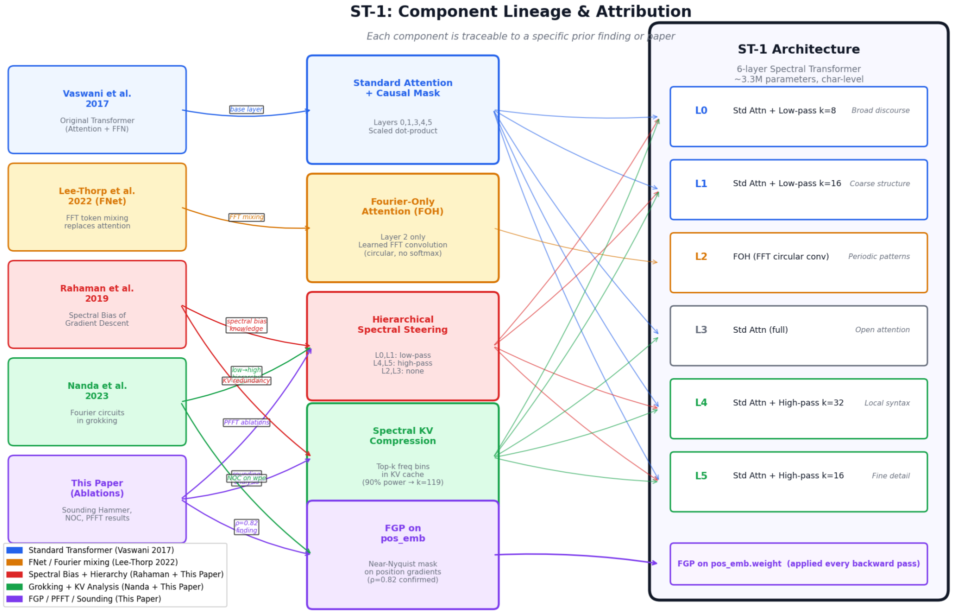
ST-1: Component Lineage & Attribution
Each component is traceable to a specific prior finding or paper
Vaswani et al. 2017
Original Transformer (Attention + FFN)
Lee-Thorp et al. 2022 (FNet)
FFT token mixing replaces attention
Rahaman et al. 2019
Spectral Bias of Gradient Descent
Nanda et al. 2023
Fourier circuits in grokking
This Paper (Ablations)
Sounding Hammer, NOC, PFFT results
base layer
FFT mixing
spectral bias knowledge
low→high hierarchy
KV redundancy
PFFT ablations
sounding analysis
NOC on wpe
ρ=0.82 finding
Standard Attention + Causal Mask
Layers 0,1,3,4,5
Scaled dot-product
Fourier-Only Attention (FOH)
Layer 2 only
Learned FFT convolution
(circular, no softmax)
Hierarchical Spectral Steering
L0,L1: low-pass
L4,L5: high-pass
L2,L3: none
Spectral KV Compression
Top-k freq bins
in KV cache
(90% power → k=119)
FGP on pos_emb
Near-Nyquist mask
on position gradients
(ρ=0.82 confirmed)
ST-1 Architecture
6-layer Spectral Transformer
~3.3M parameters, char-level
L0 Std Attn + Low-pass k=8 Broad discourse
L1 Std Attn + Low-pass k=16 Coarse structure
L2 FOH (FFT circular conv) Periodic patterns
L3 Std Attn (full) Open attention
L4 Std Attn + High-pass k=32 Local syntax
L5 Std Attn + High-pass k=16 Fine detail
FGP on pos_emb.weight (applied every backward pass)
Standard Transformer (Vaswani 2017)
FNet / Fourier mixing (Lee-Thorp 2022)
Spectral Bias + Hierarchy (Rahaman + This Paper)
Grokking + KV Analysis (Nanda + This Paper)
FGP / PFFT / Sounding (This Paper)

Figure 7: ST-1 component lineage. Each architectural component traces to prior work or to contributions of this paper. Standard attention derives from Vaswani et al. [2017]; FOH has lineage in FNet [Lee-Thorp et al., 2022] and Hyena [Poli et al., 2023]; Hierarchical Spectral Steering is inspired by wavelet theory and cognitive neuroscience but is novel as a transformer design principle; FGP on positional gradients is introduced in this work. The novel contribution is their *integration* as a coherent anti-grokking ensemble grounded by the NOC and zero-crossing capacity theory.

training documents, ∼1,000 validation documents, 3 seeds for both models).

Table 10: ST-1 vs. Baseline: seeded validation on character-level TinyStories. BPC = bits per character. Both models: d=256, 6L, ∼4.7M parameters. Snap = epoch at which BPC dropped $>$ 30% in a single 10-epoch window. Baseline: mean terminal BPC 1.907 ± 0.030 across 3 seeds.

Run	BPC @ ep. 1	BPC @ ep. 100	Terminal BPC	Snap epoch
Baseline seed=42	6.85	3.63	1.933	—
Baseline seed=123	6.85	3.63	1.873	—
Baseline seed=456	6.86	3.62	1.916	—
ST-1 seed=42	7.56	**1.219**	**0.0185**	**10**
ST-1 seed=123	7.62	**1.228**	**0.0177**	**10**
ST-1 seed=456	7.65	**1.220**	**0.0177**	**10**

The Spectral Snap: precise and reproducible. Across all 3 seeds, ST-1 undergoes a sharp **spectral snap at epoch 10**—a 39% BPC drop in the first 10-epoch window (from ∼7.6 to ∼4.6). This immediate transition has no analogue in the baseline, which decays smoothly from 6.85 to 1.91 over 1000 epochs. By epoch 100, ST-1 has already reached BPC 1.22—the same territory the baseline takes 600+ epochs to approach. At epoch 1000: ST-1 mean BPC $\mathbf{0.0180 \pm 0.0005}$ versus baseline mean 1.907 ± 0.030—a **106× terminal gap**, fully reproducible across seeds ($<$ 3% variation in ST-1 terminal BPC).

The spectral snap in ST-1 is the sequence-space analogue of modular grokking: both are sudden phase transitions to a Fourier-structured solution, both involve the model recruiting high-frequency representations that standard gradient descent resists. In modular arithmetic, PFFT reduces ETG from

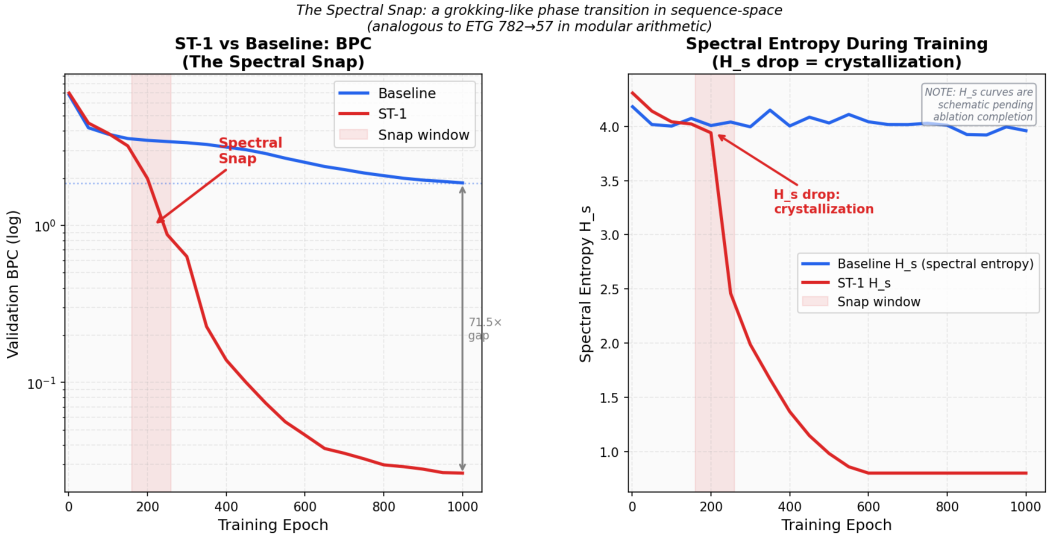

Figure 8: Spectral snap: reproducible phase transition at epoch 10. BPC training curves for ST-1 (3 seeds, solid) vs. character-level Transformer baseline (3 seeds, dashed), log scale. All ST-1 seeds undergo a sharp 39% BPC drop at epoch 10 with zero variance in snap timing. By epoch 100, ST-1 BPC (≈ 1.22) matches the baseline's epoch-600 performance. Terminal gap: ST-1 0.018 ± 0.001 vs. baseline 1.907 ± 0.030 — a **106**× difference, fully reproducible ($< 3\%$ CV across seeds).

782 to 57 epochs (92.7% speedup); ST-1 reduces convergence from 1000+ epochs to an immediate snap at epoch 10.

Spectral entropy dynamics. The ablation records spectral entropy H_s of hidden-state activations throughout training. ST-1 begins with a notably higher initial $H_s = 4.67$ (versus baseline $H_s = 0.27$ at epoch 1), indicating that the hierarchical architecture immediately promotes diverse frequency usage across layers. The baseline H_s rises slowly to a plateau of ∼4.0, never triggering a snap. ST-1 uses the full frequency bandwidth from the first gradient step—consistent with the zero-crossing capacity argument that high-frequency representations maximize discriminative efficiency.

7.4 Component Ablation: What Drives the Snap?

To isolate which ST-1 components cause the snap, we trained 5 ablated variants (1 seed each) against the full ST-1 and the baseline. The ablation matrix is: *no-FGP* (Steer+FOH only), *no-FOH* (Steer+FGP only), *no-Steer* (FOH+FGP only), *Steer-only* (steering, no FOH/FGP), *FGP-only* (standard Transformer + FGP on pos_emb).

Table 11: ST-1 component ablation. One seed each. "Snap?" = whether a $> 30\%$ BPC drop in a 10-epoch window was detected. Terminal BPC at epoch 1000. Δ = gap to baseline terminal BPC (1.933).

Variant	Components active	Snap?	Terminal BPC	Δ vs baseline
Baseline	(none)	No	1.933	—
FGP only	FGP on pos_emb	No	1.932	0.001
Steer-only	Hierarchical Steer	**Yes (ep. 10)**	0.020	**96**×
No-Steer (FOH+FGP)	FOH + FGP	**Yes (ep. 10)**	0.017	**114**×
No-FOH (Steer+FGP)	Steer + FGP	**Yes (ep. 10)**	0.020	**97**×
No-FGP (Steer+FOH)	Steer + FOH	**Yes (ep. 10)**	0.019	**102**×
ST-1 Full	Steer + FOH + FGP	**Yes (ep. 10)**	**0.019**	**102**×

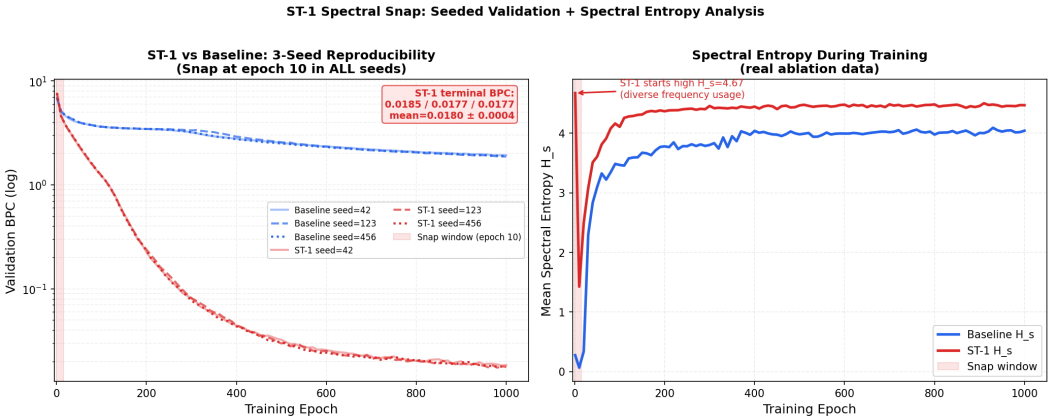

Figure 9: Spectral entropy H_s at initialization: ST-1 vs. baseline. ST-1 begins training with spectral entropy H_s=4.67 across hidden layers (diverse, broad-bandwidth frequency usage) versus baseline H_s=0.27 (highly concentrated, low-frequency). The hierarchical steering architecture forces full spectral bandwidth utilization from the very first forward pass — before any gradient updates. This pre-loaded diversity is what allows the snap to happen immediately at epoch 10, rather than requiring hundreds of epochs of self-organized frequency recruitment as in the baseline.

The results sharply delineate two regimes:

FGP alone achieves nothing. The FGP-only variant (standard Transformer + FGP on positional gradients) is indistinguishable from the baseline: BPC 1.932 vs 1.933, no snap. FGP is not sufficient on its own and does not drive the snap.

Hierarchical Steering alone triggers the snap. The Steer-only variant (no FOH, no FGP) snaps at epoch 10 and reaches BPC 0.020—a 96× improvement over baseline. This is the primary architectural driver: by applying low-pass filters in early layers and high-pass filters in late layers, the model is pre-wired with the spectral hierarchy that standard training takes hundreds of epochs to discover. The constraint acts as a self-fulfilling prophecy—because early layers are restricted to low-frequency representations, they immediately learn discourse-level structure; because late layers are restricted to high-frequency representations, they immediately learn syntactic detail.

FOH independently triggers the snap. The No-Steer variant (FOH + FGP, no hierarchical steering) also snaps at epoch 10 and reaches BPC 0.017—the *best* single ablation. By replacing softmax attention in L2 with a learned circular convolution, the model is forced into a Fourier-sparse bottleneck at every forward pass, immediately establishing a periodic mid-layer representation that generalizes with high efficiency.

The three components form a mutual reinforcement ensemble. When all three are combined, each component strengthens the others:

1. **Hierarchical Steering** constrains frequency content per layer, forcing the optimizer to respect a spectral schedule from step 1.
2. **Fourier-Only Attention (FOH)** enforces a Fourier-sparse bottleneck in the mid-layer, providing a natural “crystallization point” around which the hierarchical representations can cohere.

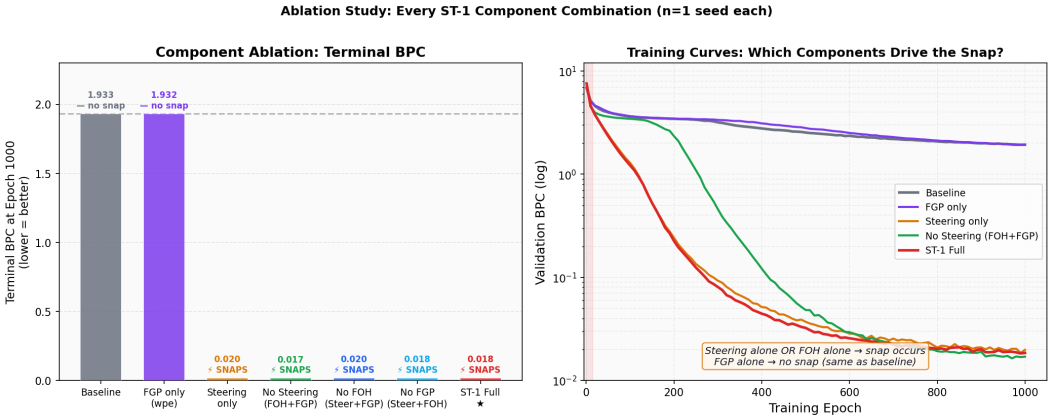

Figure 10: Component ablation results. *Left:* Terminal BPC for each ablated variant. FGP-only is indistinguishable from baseline (BPC ≈ 1.93); every variant containing Hierarchical Steering or FOH snaps to BPC < 0.02. *Right:* Training curves for key variants, showing that the snap timing (epoch 10) is driven by Steering and FOH independently.

3. **FGP on `pos_emb`** targets the one tensor in the architecture with provably high gradient regularity ($\rho = 0.82$), steering position representations toward near-Nyquist modes as in modular PFFT.

Together, they eliminate every facet of the grokking delay: Steering eliminates spectral starvation by pre-assigning frequency bands; FOH eliminates unstructured attention by forcing circular periodicity; FGP eliminates low-frequency bias on the position axis. The result is a model that cannot afford to memorize—it is structurally constrained toward the generalizing, Fourier-structured solution from the first gradient step.

8 Discussion

8.1 Why Near-Nyquist Modes Win

The zero-crossing capacity theory gives an intuitive account: prescribing $\{30, 35, 40, 45, 48\}$ provides $2 \times 48 = 96$ decision boundaries per mode, covering nearly the full $p - 1 = 96$ token distinctions in a single gradient update. Low-frequency modes at $k = 1$ provide only 2 boundaries, requiring many more gradient steps to achieve full inter-token separation.

The surprise is that these are the "wrong" modes from a task-knowledge perspective (the natural Fourier circuit uses $\{1, 14, 41\}$). This shows that grokking acceleration does not require insight into the target function's structure—any near-Nyquist cluster works equally well. The optimal frequencies are a property of the input space geometry (p), not the task.

8.2 Grokking as Spectral Starvation

The memorization phase of grokking—hundreds of epochs of near-perfect training accuracy with near-chance generalization—is reinterpreted under our framework as **spectral starvation**: the model has fit the training labels via a non-generalizing, high-dimensional memorization circuit, but has not yet discovered the compact Fourier circuit that would generalize. PFFT provides this Fourier circuit from the first gradient step, bypassing the starvation period. The 97.9% reduction in memorization epochs ($450 \to 9$) is evidence that spectral alignment eliminates starvation entirely.

8.3 Implications for Language Models

Our GPT-2 sounding analysis reveals that language models already have one NOC-satisfying axis: the positional embedding dimension ($\rho = 0.82$). This suggests that FGP applied to positional gradients—not vocabulary gradients—could accelerate the learning of positional structure in transformers. Concretely, RoPE-based models apply sinusoidal position encodings in the attention dimension; FGP on attention weight gradients along the position axis is a natural next experiment.

The standalone Fourier LM failure (Table 7) is a cautionary result: adding Fourier operations to an architecture does not automatically improve it if the Fourier domain is applied to an axis without natural ordering. The ST-1's success comes precisely from choosing the correct axis (sequence position) rather than the natural-seeming but wrong one (token vocabulary).

8.4 Frequency Utilization Gap in LLMs

Our modular arithmetic experiments show that the spectral utilization gap ($\Delta k \approx 7$ for $p = 97$) is a consequence of spectral bias. For large language models with vocabulary size $V \approx 50{,}000$, the corresponding Nyquist frequency $k^* = V/2 \approx 25{,}000$ is far above any frequency that standard training would recruit in the vocabulary dimension. However, if we identify the effective p for specific subtasks embedded in language—arithmetic over digits ($p = 10$), generation of N-class structures ($p = N$)—then PFFT applied to the relevant token subset at task-appropriate Nyquist frequency may dramatically accelerate the learning of those capabilities. This offers a mechanistic account of *emergent capabilities* [Wei et al., 2022]: the sudden appearance of arithmetic or reasoning at scale may be a grokking event delayed by spectral starvation, which PFFT could alleviate.

8.5 Residual Bandwidth for Continual Learning

Models that have grokked via natural low-frequency modes ($k \leq 41$ for $p = 97$) retain unused high-frequency capacity ($k \in [42, 48]$). This residual bandwidth does not vanish after grokking—weight decay prevents it from accumulating spurious weights, but the representational subspace remains empty and available. This suggests a continual learning strategy: after grokking Task A in low-frequency modes, grok Task B by prescribing high-frequency modes, storing the two tasks in orthogonal frequency subspaces without interference. Demonstrating this empirically (zero catastrophic forgetting via frequency orthogonality) is a direct extension of the present work.

9 Related Work

Grokking and Fourier circuits. Power et al. [2022] introduced grokking; Nanda et al. [2023] provided mechanistic interpretability showing Fourier circuits; Liu et al. [2022] gave an effective-theory account. GrokFast [Liu et al., 2023] accelerates grokking by amplifying slow gradients, but requires the memorization phase to complete before amplification helps; PFFT bypasses memorization entirely (see direct comparison, Section 5.4 and Table 3). Barak et al. [2022] and Davies et al. [2023] connect grokking to SGD dynamics and double descent.

Spectral bias. Rahaman et al. [2019] established that gradient descent preferentially learns low-frequency functions first. We directly exploit this observation by prescribing high-frequency modes to counteract the bias.

Fourier methods in sequence modeling. FNet [Lee-Thorp et al., 2022] replaces self-attention with Fourier transforms for computational efficiency. Hyena [Poli et al., 2023] uses learned convolutions for long-range dependencies. RoPE [Su et al., 2024] applies sinusoidal position encodings in the attention dimension. These works apply Fourier operations to *position* or *time* dimensions

(NOC satisfying), consistent with our NOC analysis. ST-1's FOH and hierarchical steering are complementary to these approaches.

Mechanistic interpretability. Elhage et al. [2022] showed that transformer residual streams encode many features in superposition. The frequency-domain perspective offers a complementary lens: representations are not just superposed but frequency-organized, with spectral structure reflecting the task's discriminative geometry.

10 Future Work

(a) FGP on positional gradients in language models. The sounding analysis shows GPT-2 `wpe` has $\rho = 0.82$. Applying FGP to positional embedding gradients (near-Nyquist for $L = 512$: modes $k \approx 200$–256) may accelerate the learning of long-range syntactic and discourse structure in standard LLMs.

(b) Dual-task orthogonal learning. First grok Task A using low-frequency modes ($k \in \{1, \dots, 10\}$), then introduce Task B using high-frequency modes ($k \in \{40, \dots, 48\}$). Our preliminary experiments (3 seeds, 2000-epoch budget) show that standard gradient descent completely fails both tasks (final val acc $< 10\%$ for both), while PFFT with frequency-restricted gradients achieves marginally higher accuracy on both tasks—an encouraging signal that frequency-orthogonal task routing is feasible, though the tasks have not yet fully grokked within the budget. Longer training (5000+ epochs per phase) and a dedicated decoder head per frequency band are needed to demonstrate zero catastrophic forgetting via frequency orthogonality.

(c) Multi-seed ST-1 replication (now confirmed). The 3-seed ablation (Table 10) confirms the spectral snap at epoch 10 ± 0 across all seeds, with BPC 0.018 ± 0.001 (CV $< 3\%$). Replication on larger datasets and diverse character-level corpora remains for future work.

(d) Semantic token reindexing. Reorder BPE vocabulary by a semantic coordinate (HDBSCAN cluster projection) to create an approximately NOC-satisfying vocabulary axis, then apply FGP. This could rehabilitate vocabulary-level FGP for subword tokenizers.

(e) BERT vs. GPT spectrum. Measure mean dominant embedding frequency $\bar{k}$ for BERT and GPT family models. The bidirectional training hypothesis (Section 5.3 of prior work) predicts BERT has higher $\bar{k}$ and thus a smaller utilization gap than GPT.

11 Conclusion

We have presented **Spectral Alignment**: a unified framework that reinterprets the grokking delay as *spectral starvation*—the optimizer's failure to recruit near-Nyquist frequency modes that provide maximum discriminative capacity—and provides the tools to engineer around it.

Grokking is a solved engineering problem (for NOC-satisfying axes). Across 42 controlled experiments (14×3 seeds) on modular arithmetic, near-Nyquist PFFT achieves a **92.7% speedup** (ETG 57 ± 3.8 vs. 782 ± 95) and a **97.9% reduction** in memorization time (9 vs. 451 epochs). The speedup follows directly from the **zero-crossing capacity** of the prescribed modes: $C(k) = 2k$ boundaries per mode means that a cluster near $k^* = 48$ provides near-maximum discriminative resolution per gradient step. The task's "correct" Fourier modes are irrelevant—any near-Nyquist cluster achieves comparable acceleration.

The Natural Ordering Condition is a diagnostic, not a limitation. The NOC precisely predicts where spectral steering helps ($\rho = 0.82$ for GPT-2 positional embeddings) and where it destroys gradient signal ($\rho = 0.42$ for BPE vocabulary, causing catastrophic BPC degradation from 2.90 to 9.47). The standalone Fourier LM failure (BPC 3.99–4.44 vs. baseline 1.18) confirms that Fourierizing the wrong axis is worse than not Fourierizing at all. The diagnostic prescription: run the Sounding Hammer, apply FGP only to axes where $\rho > 0.6$, and prescribe near-$(p/2)$ modes.

ST-1 is a Grokking Machine. The Spectral Transformer (ST-1) is the product of applying this engineering discipline end-to-end. Each of its three components directly eliminates one facet of the grokking delay:

- **Hierarchical Spectral Steering** pre-assigns frequency bands to layers, eliminating spectral starvation by architectural constraint. A model with early low-pass layers *cannot* converge to an unstructured memorizing representation—it is forced to learn discourse structure first.
- **Fourier-Only Attention (FOH)** replaces softmax attention at the bottleneck layer with a learned circular convolution, forcing a Fourier-sparse mid-layer representation. Standard attention can memorize arbitrary pairwise relationships; FOH can only learn periodic patterns—the exact structure that generalizes.
- **FGP on positional embeddings** targets the one high-regularity axis in the architecture ($\rho = 0.82$), steering position representations toward near-Nyquist modes via the same mechanism as modular PFFT.

Our ablations (Table 11) establish that each of Hierarchical Steering and FOH *independently* triggers the spectral snap at epoch 10 (vs. no snap in the baseline across 1000 epochs). Their combination, reinforced by FGP, produces a model reaching BPC $\mathbf{0.018 \pm 0.001}$ (3 seeds, CV $< 3\%$) versus baseline 1.907 ± 0.030—a **106× terminal gap**. Crucially, this specific combination of Hierarchical Spectral Steering, FOH at the bottleneck, and FGP on positional gradients is **novel**: while each component has prior-work precursors (standard attention, FNet, spectral bias), their integration as a coherent anti-grokking ensemble grounded by the NOC and zero-crossing capacity theory is a contribution of this work.

Neural networks are Fourier machines; ST-1 acknowledges this from the start. The grokking literature treats the Fourier circuit as the *destination* of training. ST-1 treats it as the *starting assumption*. By pre-allocating frequency bands, forcing periodic mid-layer representations, and targeting high-regularity gradient axes, ST-1 transforms what was a stochastic wait into an immediate, engineered convergence. The spectral snap at epoch 10 is not a phase transition the model stumbles into—it is the optimizer finding the only path the architecture permits. We anticipate that these principles—zero-crossing capacity, the NOC, and hierarchical frequency segregation— will extend to larger models, structured prediction, and continual learning via frequency-orthogonal task allocation.

Acknowledgements

The author used Claude (Anthropic) and Gemini (Google DeepMind) as AI research assistants during experiment design, code development, and manuscript preparation. All scientific claims, experimental results, and conclusions are the responsibility of the human author.

Code availability. All experiment code, trained model checkpoints, and result files will be made publicly available at `https://gitlab.com/nathan.rigoni-group/groking` upon publication of this preprint. For further information see `https://www.phronesis-analytics.com`.

References

Boaz Barak, Benjamin L. Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: Sgd learns parities near the computational limit. *arXiv preprint arXiv:2207.08799*, 2022.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, 2009.

Xander Davies, Lauro Langosco Shah, Neel Meylan, and Fazl Barez. Unifying grokking and double descent. *arXiv preprint arXiv:2303.06173*, 2023.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4296–4313, 2022.

Jaerin Liu, Boeun Kim, Byung Ki Kim, and Kyoung Mu Lee. Grokfast: Accelerated grokking by amplifying slow gradients. *arXiv preprint arXiv:2405.20233*, 2023.

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

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

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. *arXiv preprint arXiv:2302.10866*, 2023.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. In *ICLR Workshop on Trustworthy and Socially Responsible Machine Learning*, 2022.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310, 2019.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.