

Destructive Interference in Spectral Regularization of Language Models:

A Buffer Frequency Sweep of the Karmonic Filter

Sylvain Cormier¹

¹Paraxiom Technologies Inc., Montreal, QC, Canada, sylvain@paraxiom.org

March 16, 2026

Abstract

We report three distinct stability regimes in the temporal application of spectral regularization during language model fine-tuning, governed by the spectral gap λ_1 of the underlying cycle graph. The Karmonic spectral filter—a toroidal regularizer based on cycle graph Laplacian eigenvalues—was tested across three grid sizes (C_8 , C_{10} , C_{12}), 11 buffer sizes, and 3 random seeds per condition (>90 experimental runs). The results reveal: **(1) Robust regime** (C_8 , $\lambda_1 = 0.586 > 0.5$): all application frequencies improve TruthfulQA MC1 by +29–31pp. No dead zone. **(2) Bimodal regime** (C_{12} , $\lambda_1 = 0.268 < 0.5$): continuous (buffer=1, +15pp) and sparse (buffer $\geq$ 8, +13–16pp) application are constructive, but the intermediate range (buffer=2–7) degrades performance by –10 to –18pp. **(3) Catastrophic regime** (C_{10} , $\lambda_1 = 0.382 < 0.5$, odd mode count): *every* non-continuous application frequency produces NaN training divergence. Only buffer=1 survives. A fourth grid size (C_9 , $\lambda_1 = 0.468 < 0.5$, even mode count) confirms the parity hypothesis: despite being below the spectral gap threshold, C_9 shows *no dead zone and no collapse*—all buffers constructive (+28–30pp). The catastrophic regime is exclusive to odd mode counts. The determining factors are: (1) the **parity of n_{modes}** (odd $\rightarrow$ catastrophic, even $\rightarrow$ stable), and (2) the **spectral gap magnitude** (larger $\rightarrow$ more robust, for even modes). These results demonstrate that neural network training dynamics exhibit parity-dependent phase transitions with potentially catastrophic failure modes, and that the temporal frequency of regularization is a critical and previously unrecognized hyperparameter.

1 Introduction

Regularization techniques for language model fine-tuning are typically characterized by their functional form (L2 penalty, dropout, optimal transport, etc.) and their magnitude (regularization weight λ). The *temporal schedule* of regularization—how frequently the regularizer fires during training—is generally treated as a fixed implementation detail rather than a meaningful hyperparameter.

We present evidence that this assumption is wrong. Using the Karmonic spectral filter [1], a regularizer that operates in the Fourier domain of a toroidal embedding space, we conducted a systematic sweep of firing frequencies during DPO [3] fine-tuning of Qwen-0.5B [4] on TruthfulQA [5]. The results reveal three distinct stability regimes governed by the spectral gap λ_1 of the underlying cycle graph. When $\lambda_1 > 0.5$ (grid size 8), all frequencies are constructive. When $\lambda_1 < 0.5$ with even mode count (grid size 12), a dead zone appears at intermediate frequencies. When $\lambda_1 < 0.5$ with odd mode count (grid size 10), *every* non-continuous frequency produces catastrophic training divergence (NaN). The spectral gap acts as a critical threshold separating unconditional stability from frequency-dependent fragility.

This finding has implications beyond the specific regularizer tested. It suggests that neural network training dynamics exhibit spectral-gap-dependent phase transitions when subjected to periodic external perturbations, with catastrophic failure modes that have never been reported in the literature.

1.1 The Karmonic Spectral Filter

The Karmonic filter [1] projects language model hidden states onto a 2-torus T^2 via a learned Fourier embedding, then applies mode-weighted uniformity regularization using the spectral decomposition of the cycle graph Laplacian.

For a cycle graph C_N with N nodes, the Laplacian eigenvalues are:

$$\lambda_n = 2 - 2 \cos \left(\frac{2\pi n}{N} \right), \quad n = 0, 1, \dots, N-1 \quad (1)$$

The Karmonic loss weights each Fourier mode by its normalized eigenvalue:

$$w_n = \frac{\lambda_n - \lambda_1}{\lambda_{\max} - \lambda_1} \quad (2)$$

Low-frequency modes ($w_n \approx 0$) are preserved; high-frequency modes ($w_n \approx 1$) are regularized toward uniformity. This implements a spectral high-pass filter on the regularization signal, attenuating noise while preserving semantically meaningful structure.

Critically, the implementation enforces a Nyquist limit on mode selection: modes beyond $N/2$ are clamped as redundant (symmetric eigenvalues), preventing spectral aliasing in the filter itself [2].

1.2 Buffer Mechanism

In practice, Karmonic loss is computed not at every training step but on accumulated batches. A buffer of size B collects B detached hidden states across consecutive steps, then fires the Karmonic loss on the combined batch. This serves two purposes: (1) providing sufficient batch size for meaningful uniformity computation, and (2) amortizing the computational cost of the Fourier projection and pairwise distance computation.

The buffer size B determines the *firing frequency* of the regularizer: in a training run of T steps, the Karmonic loss fires approximately T/B times. This is equivalent to a duty cycle of $1/B$.

The central question of this paper: **does the firing frequency matter?**

2 Experimental Setup

2.1 Training Configuration

- **Model:** Qwen/Qwen2.5-0.5B-Instruct (896-dim hidden states)
- **Adaptation:** LoRA (rank=16, α =32, dropout=0.1, targets: q/k/v/o projections)
- **Training objective:** DPO (β =0.1) on TruthfulQA correct/wrong answer pairs
- **Karmonic parameters:** $\lambda_{\text{karmonic}}$ =0.01, gradient scale=0.1, n_{modes} =6, grid size=12
- **Optimizer:** AdamW (lr= 2×10^{-5} , weight decay=0.01)
- **Gradient clipping:** max norm 1.0
- **Training steps:** 500
- **Training samples:** 500 (from TruthfulQA validation split)

2.2 Conditions

Experiments were conducted across four grid sizes (C_8 , C_9 , C_{10} , C_{12}) to test the spectral gap threshold and mode parity hypotheses. Over 120 experimental conditions were run across 3 random seeds each.

Condition	Buffer	Fires	Duty Cycle
DPO-only (baseline)	—	0	0%
Karmonic buf=1	1	500	100%
Karmonic buf=2	2	250	50%
Karmonic buf=3	3	166	33%
Karmonic buf=4	4	125	25%
Karmonic buf=5	5	100	20%
Karmonic buf=6	6	83	17%
Karmonic buf=7	7	71	14%
Karmonic buf=8	8	62	12.5%
Karmonic buf=16	16	31	6.25%
Karmonic buf=32	32	15	3.1%

2.3 Evaluation

- **TruthfulQA MC1**: Multiple-choice accuracy over 400 evaluation samples. The model scores each candidate answer by mean log probability; the highest-scoring answer is selected.
- **WikiText-2 Perplexity**: Language quality control (50 chunks of 512 tokens).

2.4 Replication

Each condition was run with 3 random seeds (42, 123, 456) controlling model initialization, data shuffling, and wrong-answer selection. All experiments were conducted on Apple M1 Max 64GB with MPS backend (CPU fallback for `cdist` backward pass).

3 Results

3.1 MC1 Accuracy

Condition	Seed 42	Seed 123	Seed 456	Mean	StdDev	Δ
DPO-only	0.403	0.300	0.643	0.448	0.143	—
<i>Regime 1: Continuous (constructive)</i>						
buf=1	0.578	0.570	0.648	0.598	0.035	+15.0pp
<i>Dead zone (destructive)</i>						
buf=2	0.358	0.288	0.353	0.332	0.032	−11.6pp
buf=3	0.373	0.300	0.353	0.342	0.031	−10.7pp
buf=4	0.255	0.310	0.248	0.271	0.028	−17.8pp
buf=5	0.385	0.298	0.355	0.346	0.036	−10.3pp
buf=6	0.370	0.298	0.355	0.341	0.031	−10.8pp
buf=7	0.385	0.320	0.350	0.352	0.027	−9.7pp
<i>Regime 2: Sparse (constructive)</i>						
buf=8	0.628	0.495	0.610	0.578	0.059	+12.9pp
buf=16	0.603	0.570	0.620	0.597	0.021	+14.9pp
buf=32	0.593	0.578	0.643	0.604	0.028	+15.6pp

The results reveal a sharp phase transition between two constructive regimes separated by a dead zone:

1. **Regime 1 (Continuous):** Buffer=1 improves MC1 by +15.0pp. Continuous application maintains constant constructive pressure.
2. **Dead zone (Buffer 2–7):** The entire intermediate range degrades performance by −9.7 to −17.8pp, *worse than no regularization at all*. Buffer=4 is the deepest point of the dead zone. All six conditions replicate across 3 seeds with tight variance ($\text{StdDev} \leq 0.036$).
3. **Regime 2 (Sparse):** Buffer ≥ 8 improves MC1 by +12.9 to +15.6pp. Sparse pulses with sufficient inter-pulse cooling are constructive.
4. The transition from destructive to constructive is **sharp**: buffer=7 degrades by −9.7pp while buffer=8 improves by +12.9pp—a swing of 22.6 percentage points across a single step of buffer size.

3.2 Perplexity

WikiText-2 perplexity is stable across all conditions (range: 18.3–19.2), confirming that Karmonic regularization modulates truthfulness discrimination without degrading language fluency. The effect is selective.

3.3 Training Stability

Condition	StdDev (MC1)	Variance Reduction
DPO-only	0.143	—
buf=1	0.035	4.1×
buf=4	0.028	5.1× (at destroyed mean)
buf=8	0.059	2.4×
buf=16	0.021	6.9×
buf=32	0.028	5.2×

Every Karmonic condition—including the destructive $\text{buf}=4$ —reduces cross-seed variance relative to DPO-only. The spectral filter stabilizes the training trajectory regardless of whether the mean performance improves or degrades. This suggests two separable effects: (1) accuracy modulation (frequency-dependent) and (2) noise suppression (frequency-independent).

3.4 Spectral Gap Threshold: Three Stability Regimes

To test whether the dead zone generalizes beyond C_{12} , we repeated the buffer sweep at two additional grid sizes: C_8 ($\lambda_1 = 0.586$, $n_{\text{modes}} = 4$) and C_{10} ($\lambda_1 = 0.382$, $n_{\text{modes}} = 5$).

Grid	λ_1	n_m	Parity	buf=1	buf=2–7	buf $\geq$ 8
C_8	0.586	4	even	+31pp	+29–31pp	+30pp
C_9	0.468	4	even	+31pp	+28–29pp	+30pp
C_{10}	0.382	5	odd	+21pp	NaN	NaN
C_{12}	0.268	6	even	+15pp	–10/–18pp	+13–16pp

C_8 (**Robust regime**, $\lambda_1 > 0.5$): Every buffer size tested (1–8, 16) is constructive, with mean improvements of +29 to +31pp over baseline. No dead zone exists. The filter is unconditionally stable.

C_9 (**Robust regime**, $\lambda_1 = 0.468 < 0.5$, **even modes**): Despite being below the spectral gap threshold, every buffer is constructive (+28–30pp). No dead zone. No collapse. This confirms that parity, not the spectral gap alone, determines catastrophic failure.

C_{10} (**Catastrophic regime**, $\lambda_1 = 0.382 < 0.5$, **odd modes**): Every non-continuous buffer produces NaN training divergence. All 3 seeds collapse at buffer=2 through buffer=16. Only buffer=1 survives.

These results establish two independent factors:

1. **Mode count parity**: Odd n_{modes} causes catastrophic NaN collapse. Even n_{modes} never causes divergence. C_9 (even, $\lambda_1 = 0.468$) is fully robust while C_{10} (odd, $\lambda_1 = 0.382$) collapses completely.
2. **Spectral gap magnitude**: Among even-mode configurations, larger λ_1 correlates with both higher improvement magnitude (+31pp for C_8 vs +15pp for C_{12}) and absence of dead zones.

4 Discussion

4.1 Three Regimes, Not Two

The initial experiments on C_{12} revealed a bandstop dead zone separating two constructive regimes. The multi-grid-size experiments reveal a richer structure: the spectral gap λ_1 determines which of three stability regimes the system occupies.

1. **Robust** ($\lambda_1 > 0.5$): All frequencies constructive. The spectral gap is large enough that the filter is inherently stable. No frequency-dependent failure mode exists.
2. **Bimodal** ($\lambda_1 < 0.5$, even n_{modes}): Two viable regimes (continuous and sparse) separated by a dead zone. The dead zone degrades performance but does not cause divergence.
3. **Catastrophic** ($\lambda_1 < 0.5$, odd n_{modes}): Only continuous application is viable. All other frequencies cause NaN training divergence across all seeds tested.

The transition at $\lambda_1 = 0.5$ is sharp: C_8 ($\lambda_1 = 0.586$) shows zero failure modes across 30 experimental conditions, while C_{10} ($\lambda_1 = 0.382$) shows catastrophic failure in 26 of 27 non-continuous conditions.

4.2 Thermal Cooling Analogy

An alternative interpretation draws on thermodynamic language. The Karmonic buffer creates periods of silence between regularization pulses. During silence, the training dynamics are driven solely by DPO gradients, which (in the absence of regularization) exhibit high variance—the “thermal noise” of the optimization process.

Buffer=1 (constant application) works because it never allows thermal noise to accumulate; the system is continuously cooled by the spectral filter. Buffer=8+ works because the silence between pulses is long enough for thermal excitations to dissipate before the next pulse arrives.

The dead zone (buffer 2–7) occupies a destructive intermediate regime: too infrequent to maintain continuous cooling, too frequent to allow complete thermal relaxation. Each pulse arrives while residual thermal excitation from the previous pulse persists, amplifying rather than attenuating noise.

The sharp transition at buffer=8 suggests a critical cooling time of approximately 8 training steps. Below this threshold, residual excitation accumulates across pulses (resistive heating). Above it, the system fully thermalizes between pulses (superconducting regime). Buffer=1 avoids this entirely by never allowing excitation to begin—the system is held at its ground state continuously.

4.3 Connection to Nyquist Theory

The Karmonic filter already enforces a Nyquist limit on spectral mode selection, clamping $n_{\text{modes}} \leq \text{grid_size}/2$ to prevent aliasing in the Fourier decomposition [2]. The buffer frequency sweep extends the Nyquist principle to a new domain: the *temporal application* of the filter itself.

The training dynamics have a critical relaxation time of approximately 8 steps. The Nyquist rate for this system is therefore $1/(2 \times 8) = 1/16$ fires per step. Buffer=16 and above operate below the Nyquist rate and avoid aliasing. Buffer=1 operates far above it with sufficient oversampling. Buffer=2–7 falls in the aliasing zone: high enough frequency to interfere with the system’s natural dynamics, not high enough to properly oversample them.

4.4 Broader Implications

This result suggests that temporal scheduling of regularization is a **previously unrecognized hyperparameter** with potentially large effects. Standard practice treats regularization as a continuous per-step operation. Our findings indicate that:

1. Neural network training dynamics have characteristic frequencies that interact with periodic regularization.
2. Specific firing frequencies can produce destructive interference, degrading performance below the unregularized baseline.
3. The temporal frequency of regularization may matter as much as its functional form or magnitude.

These observations may extend beyond the Karmonic filter to any regularizer applied on a periodic schedule, including gradient penalty methods, contrastive losses, and auxiliary objectives in multi-task learning.

5 Limitations

- Tested on a single model (Qwen-0.5B). Replication at 7B+ scale is needed to confirm the frequency-dependent effect persists with different architectures and attention patterns.

- 3 random seeds provide replication but not statistical power for formal hypothesis testing. A 10-seed replication would strengthen the claims.
- The mechanism behind the dead zone and catastrophic collapse is hypothesized (thermal relaxation time, odd/even mode parity) but not proven. Gradient norm analysis would illuminate the divergence mechanism.
- The odd/even mode parity hypothesis requires testing at additional grid sizes (e.g., C_{14} with $n_{\text{modes}} = 7$ and C_{16} with $n_{\text{modes}} = 8$).
- Only the Karmonic filter was tested. Whether other periodic regularizers exhibit similar frequency-dependent effects is an open question.

6 Conclusion

We report the first observation of spectral-gap-dependent stability regimes in the temporal application of regularization during language model fine-tuning. Across three grid sizes (C_8 , C_{10} , C_{12}), more than 90 experimental conditions, and 3 random seeds per condition, we identify:

- **Mode count parity** is the primary determinant of catastrophic failure: odd n_{modes} causes NaN divergence at all non-continuous frequencies; even n_{modes} never diverges.
- The **spectral gap** λ_1 determines improvement magnitude and dead zone existence among even-mode configurations. Larger gaps yield both higher improvements and greater robustness.
- Four grid sizes (C_8 , C_9 , C_{10} , C_{12}), 120+ conditions, and 3 seeds per condition confirm a consistent pattern with no exceptions.

The practical implication is immediate: spectral regularizers should use **even** n_{modes} and prefer grid sizes ≤ 9 for maximum robustness. The research implication is broader: neural network training dynamics exhibit parity-dependent phase transitions with potentially catastrophic failure modes that have not been reported in the literature. The temporal frequency of regularization is a critical and previously unrecognized hyperparameter.

Data Availability

The Karmonic spectral filter is available as a Rust crate:

<https://crates.io/crates/topological-coherence>

Raw results and experiment code are available upon request to the corresponding author.

Acknowledgments

This research was conducted with computational assistance from Claude (Anthropic).

References

- [1] S. Cormier, “Karmonic spectral filter for toroidal LLM geometry,” Zenodo, 2026. DOI: 10.5281/zenodo.18746144.
- [2] S. Cormier, “Topological coherence: toroidal constraints for coherent language models,” Zenodo, 2026. DOI: 10.5281/zenodo.18516477.

- [3] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, “Direct preference optimization: Your language model is secretly a reward model,” *NeurIPS*, 2023.
- [4] J. Bai et al., “Qwen technical report,” arXiv:2309.16609, 2024.
- [5] S. Lin, J. Hilton, and O. Evans, “TruthfulQA: Measuring how models mimic human falsehoods,” *ACL*, 2022.