

The Toroidal Geometry of LLM Representations: From Detection to Karmonic-Guided Alignment

Sylvain Cormier
Paraxiom Research
sylvain@paraxiom.org

February 2026

Abstract

We present evidence that large language model hidden states exhibit toroidal topology and introduce a training-time regularizer that exploits this structure for truthfulness alignment. Using persistent homology on Qwen 2.5-0.5B hidden states extracted from TruthfulQA, we detect strong toroidal signals across all network layers: Betti numbers $\beta_1 = 23\text{--}37$ (indicating many independent cycles) and $\beta_2 \geq 1$ (indicating voids), with truthful and hallucinated answers separated by $30\text{--}121^\circ$ in PCA angular space. Motivated by this geometric finding, we replace the generic Sinkhorn optimal transport regularizer in the ENIGMA alignment framework with **Karmonic spectral filtering**—a targeted regularizer derived from the cycle graph Laplacian that preserves low-frequency toroidal modes while attenuating high-frequency noise. In controlled ablation experiments using DPO preference optimization on TruthfulQA across two model scales (Qwen 2.5-0.5B and Gemma-3-1B-IT), Karmonic filtering outperforms Sinkhorn OT at 0.5B (**+1.3pp** MC1, **+4.4pp** MC2) while the two regularizers converge at 1B scale. This *scale-dependent advantage* suggests Karmonic’s spectral selectivity is most beneficial when representations are noisier. DPO with Karmonic achieves MC1 = 0.468 (0.5B) and 0.499 (1B) on TruthfulQA, improvements of **+19pp** and **+22pp** over untrained baselines, with stable perplexity at both scales. All experiments use single-GPU training with 200 steps.

Scope: This work is empirical. We detect toroidal topology and show that a torus-aware regularizer can outperform a geometry-agnostic one, with the advantage being scale-dependent. We make no claims about universal structure or ontological status of the detected topology.

1 Introduction

Large language models (LLMs) routinely generate fluent but factually incorrect text—a phenomenon termed *hallucination* [9]. Mitigation strategies divide into inference-time methods (retrieval-augmented generation, logit manipulation, chain-of-thought prompting) and training-time methods (RLHF, DPO, constitutional AI). A recent line of work explores *geometric* approaches: if LLM representations occupy specific manifold structures, one can design interventions that respect or exploit that geometry.

Our prior work on toroidal logit bias (TLB) [10] demonstrated that inference-time toroidal constraints on token selection yield consistent truthfulness gains across four models (+0.2 to +2.8 percentage points on TruthfulQA). However, TLB operates post-hoc on logits without modifying the model’s internal representations.

The ENIGMA framework [1] introduced a three-layer geometry-aware training pipeline combining group-relative policy optimization (GRPO) [3], symmetric alignment via mutual information (SAMI) [2], and Sinkhorn optimal transport (OT) regularization. On Gemma-3-1B, ENIGMA achieved +12 percentage points on TruthfulQA. However, its OT component is

geometry-agnostic—it constrains drift in Wasserstein space without knowledge of the representation manifold’s structure.

This paper bridges the gap between inference-time geometric intervention and geometry-agnostic training regularization. We make four contributions:

1. **Topological detection:** We show that LLM hidden states exhibit toroidal topology via persistent homology ($\beta_1 = 23\text{--}37$ across layers), providing empirical justification for torus-aware methods.
2. **Angular separation:** Truthful and hallucinated answers separate by $30\text{--}121^\circ$ in PCA angular projections, confirming that the detected toroidal structure encodes task-relevant information.
3. **Karmonic spectral filtering:** We introduce a targeted replacement for Sinkhorn OT that exploits the cycle graph Laplacian eigenstructure, preserving low-frequency class structure while attenuating high-frequency noise.
4. **Multi-scale validation:** In controlled ablations on Qwen 2.5-0.5B and Gemma-3-1B-IT, Karmonic beats Sinkhorn OT at 0.5B (+1.3pp MC1, +4.4pp MC2) while the two converge at 1B, revealing a scale-dependent regularization dynamic.

2 Background

2.1 Toroidal Topology and the Tonnetz

The Tonnetz, from 19th-century music theory [13], arranges pitch classes on a torus $T^2 = S^1 \times S^1$ such that harmonically related tones are geometrically proximate. The discrete torus $C_N \times C_M$ (product of cycle graphs) inherits this property: toroidal distance captures relational proximity with wraparound connectivity and no boundary effects.

The relevance to language models is the hypothesis—tested empirically in this paper—that token and concept representations in LLMs inhabit manifolds with similar cyclic structure, where semantic proximity corresponds to angular proximity on toroidal coordinates.

2.2 Persistent Homology

Persistent homology [5, 6] computes topological features of a point cloud at multiple scales. The Vietoris-Rips complex $\text{VR}_\epsilon(X)$ connects points within distance ϵ ; as ϵ increases, topological features (connected components, cycles, voids) appear and disappear. Features persisting over large ϵ -ranges indicate genuine topology rather than noise.

Betti numbers β_k count independent k -dimensional features: β_0 counts connected components, β_1 counts independent cycles (loops), and β_2 counts enclosed voids. A torus T^2 has $(\beta_0, \beta_1, \beta_2) = (1, 2, 1)$. Finding $\beta_1 \gg 2$ in data suggests a high-dimensional torus T^k with $k > 2$.

2.3 Karmonic Spectral Filtering

The normalized Laplacian of the cycle graph C_N has eigenvalues:

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

The spectral gap $\lambda_1 = 2 - 2 \cos(2\pi/N)$ governs mixing time and information propagation on the cycle. The Karmonic spectral filter [11] assigns mode-dependent weights to a Fourier

basis on T^2 : low-frequency modes (near λ_1) carry class-level structure and are preserved; high-frequency modes (near $\lambda_{\max}$) carry noise and are attenuated. This yields a *mode-weighted uniformity* loss:

$$\mathcal{L}_{\text{Karmonic}} = \sum_{n=1}^M w_n \cdot \text{LogSumExp}(-t \cdot D_n^2) \quad (2)$$

where D_n^2 is the pairwise squared distance matrix of the n -th Fourier mode embedding, t is a temperature parameter, and $w_n = (\lambda_n - \lambda_1)/(\lambda_M - \lambda_1)$ weights each mode by its spectral position.

2.4 The ENIGMA Framework

ENIGMA [1] combines three geometric training objectives:

- **GRPO** [3]: Group-relative policy optimization on the Fisher-Rao manifold, where G completions per prompt are scored and advantages are computed relative to the group.
- **SAMI** [2]: Symmetric Alignment via Mutual Information. An InfoNCE objective maximizing $I(Y; C \mid X)$ —mutual information between completion representations Y and constitutional principle embeddings C , conditioned on prompts X .
- **Sinkhorn OT**: Entropic optimal transport between current and reference hidden state distributions, constraining representation drift in Wasserstein space with entropic regularization parameter ϵ .

The unified loss is $\mathcal{L}_{\text{ENIGMA}} = \mathcal{L}_{\text{GRPO}} + \lambda_{\text{SAMI}} \mathcal{L}_{\text{SAMI}} + \lambda_{\text{OT}} R_{\text{OT}}$.

Our key observation: ENIGMA’s OT component is *geometry-agnostic*—it constrains drift without exploiting the manifold’s specific structure. If the representation manifold is toroidal (as we show), a *targeted* regularizer that knows the torus geometry should outperform generic drift control.

3 Detecting Toroidal Structure in LLM Hidden States

3.1 Method

We extract hidden states from Qwen 2.5-0.5B-Instruct [12] on the TruthfulQA validation set (817 questions). For each question, we process both the correct and incorrect MC1 answer choices through the model and capture hidden states at five layers: 0 (embedding), 12 (middle), 21, 22, and 23 (final). Each hidden state vector has dimension 896.

For each layer, we apply three independent tests:

1. Persistent homology (Ripser [6]): We subsample 500 points, apply PCA to 50 dimensions, and compute Vietoris-Rips persistence diagrams up to dimension 2. Significant features are those with persistence exceeding median $+1\sigma$.

2. Spectral analysis: We build a k -nearest-neighbor graph ($k = 10$) on 1000 subsampled points (PCA to 50 dimensions), compute the graph Laplacian $L = D - A$, and analyze the smallest eigenvalues for torus-consistent patterns (multiplicity-2 spectral gap).

3. Circular statistics: We project to multiple 2D PCA planes, convert to polar coordinates, test for non-uniform angular distribution (Rayleigh test, $p < 0.01$), and measure the angular separation between truthful and hallucinated answer centroids.

Layer	Score	β_1	β_2	Best Angular Sep.	Best Plane
0 (embedding)	6/7	32	8	120.7°	PC2-PC3
12 (middle)	6/7	37	21	35.8°	PC1-PC2
21	6/7	23	2	30.2°	PC2-PC4
22	5/7	26	3	30.0°	PC2-PC3
23 (final)	6/7	25	4	35.0°	PC2-PC3

Table 1: Persistent homology and circular statistics across layers of Qwen 2.5-0.5B on TruthfulQA. β_1 counts significant 1-cycles (persistence $>$ median $+ 1\sigma$). Score combines persistent homology (0–3), spectral multiplicity (0–1), and circular statistics (0–3).

3.2 Results

3.3 Interpretation

Three findings emerge:

High-dimensional torus. All layers show $\beta_1 \gg 2$, ranging from 23 to 37 significant 1-cycles. This far exceeds the $\beta_1 = 2$ of a standard 2-torus, suggesting the hidden states live on or near a high-dimensional torus T^k with $k \gg 2$. The middle layer (12) shows the richest topology ($\beta_1 = 37$, $\beta_2 = 21$), consistent with findings that middle layers encode the most task-relevant features.

Truth–hallucination angular separation. Truthful and hallucinated answers separate by 30–121° in PCA angular space, with the embedding layer showing the largest separation (120.7°). This confirms that the toroidal structure encodes task-relevant information—truthful and hallucinated representations occupy geometrically distinct toroidal regions.

Learned projection needed. A random projection to T^2 (via randomly initialized FourierTorusHead) yields near-zero separation ($< 1^\circ$), confirming that the truthfulness-relevant toroidal dimensions must be *learned*. This motivates training a FourierTorusHead jointly with a content signal (DPO/SAMI) to discover the task-relevant toroidal coordinates.

4 Karmonic-Guided Alignment

4.1 Architecture

We replace ENIGMA’s Sinkhorn OT with Karmonic spectral filtering while retaining the preference optimization and (optionally) mutual information components. Our architecture:

Listing 1: Karmonic alignment architecture (simplified)

```

1 class FourierTorusHead(nn.Module):
2     """Projects hidden states to torus coordinates."""
3     def __init__(self, input_dim=896, hidden=128,
4                 torus_dim=2, n_modes=6):
5         self.net = nn.Sequential(
6             nn.Linear(input_dim, hidden),
7             nn.LayerNorm(hidden), nn.ReLU(),
8             nn.Linear(hidden, torus_dim))
9         self.modes = torch.arange(1, n_modes + 1)
10
11     def forward(self, z):
12         angles = 2*pi * sigmoid(self.net(z))
13         # Fourier embedding: cos/sin at each mode
14         n_angles = angles.unsqueeze(-1) * self.modes
15         fourier = stack([cos(n_angles), sin(n_angles)])

```

```
return angles, fourier.reshape(B, -1)
```

The full training pipeline:

1. **DPO preference loss**: Direct preference optimization [4] between known correct and incorrect TruthfulQA answers: $\mathcal{L}_{\text{DPO}} = -\log \sigma(\beta \cdot (\log \pi_{\theta}(y_w | x) - \log \pi_{\theta}(y_l | x)))$, where y_w and y_l are the preferred and dispreferred answers.
2. **SAMI mutual information** (optional): Symmetric InfoNCE maximizing $I(Y; C | X)$ between completion hidden states and constitutional principle embeddings.
3. **Karmonic filter loss**: Hook the final hidden layer, mean-pool across the sequence, apply gradient scaling ($\times 0.1$) to prevent the auxiliary head from dominating, project through FourierTorusHead, compute mode-weighted uniformity.

The Karmonic component uses a buffer mechanism to accumulate hidden states across 8 steps (since batch size is 1 for DPO), then computes the loss on the concatenated batch with only the current step’s hidden states carrying gradients.

4.2 Why DPO Instead of GRPO

At the 0.5B model scale with single-GPU training, GRPO’s on-policy generation produces nearly identical completions (low diversity), yielding zero group-relative advantages and zero gradient signal. DPO avoids this by directly comparing known correct and incorrect answers from TruthfulQA’s MC1 annotations, providing a strong and consistent preference signal regardless of model scale.

4.3 Why Karmonic Should Beat Sinkhorn OT

Sinkhorn OT constrains the Wasserstein distance between current and reference hidden state distributions. This prevents catastrophic drift but is agnostic to the manifold’s structure—it penalizes all movement equally, regardless of whether it moves along meaningful toroidal coordinates or noise dimensions.

Karmonic filtering exploits the specific eigenstructure of the torus:

- **Spectral selectivity**: Low-frequency modes (λ_n near λ_1) encode class-level structure and receive low penalty weight. High-frequency modes (λ_n near $\lambda_{\max}$) encode noise and receive high penalty.
- **Manifold-aware**: The penalty structure is derived from the cycle graph Laplacian, matching the detected toroidal topology.
- **Computational efficiency**: No iterative Sinkhorn solve; the loss is a direct LogSumExp over pairwise distances per mode.

4.4 Experimental Setup

We test five conditions in controlled ablation:

Hyperparameters (identical at both scales): LoRA targets $\{q, k, v, o\}$ projections with dropout 0.1. DPO $\beta = 0.1$. Learning rate 2×10^{-5} with AdamW (weight decay 0.01). SAMI: $\lambda_{\text{SAMI}} = 0.05$, temperature 0.07, 6 constitutional principles, row InfoNCE (column skipped when batch < num principles). Karmonic: $\lambda_{\text{Karmonic}} = 0.01$, gradient scale 0.1, grid size 12, 6 modes, buffer size 8. Sinkhorn OT: $\lambda_{\text{OT}} = 0.01$, $\epsilon = 0.1$, 20 iterations. All conditions: 200 steps, 500 training samples from TruthfulQA, evaluated on the full 817-sample validation set.

Hardware: Qwen 0.5B: NVIDIA RTX 4090 (~ 1.4 GB VRAM, ~ 1 min train + ~ 6 min eval). Gemma 1B: NVIDIA A40 (~ 4 GB VRAM, ~ 1.4 min train + ~ 8 min eval). Both on RunPod.

#	Condition	DPO	SAMI	Regularizer
1	DPO only	✓		—
2	DPO + SAMI	✓	✓	—
3	DPO + SAMI + OT	✓	✓	Sinkhorn
4	DPO + SAMI + Karmonic	✓	✓	Karmonic
5	DPO + Karmonic	✓		Karmonic

Table 2: Five experimental conditions, run identically on both Qwen 2.5-0.5B [12] and Gemma-3-1B-IT [19] with LoRA ($r = 16$, $\alpha = 32$).

5 Results

5.1 Main Results

Condition	MC1	MC2	PPL	Δ MC1	Δ MC2
Untrained baseline	0.278	0.407	18.06	—	—
DPO only	0.486	0.554	18.24	+20.8pp	+14.7pp
DPO + SAMI	0.416	0.492	18.18	+13.8pp	+8.6pp
DPO + SAMI + OT	0.448	0.499	18.18	+17.0pp	+9.3pp
DPO + SAMI + Karmonic	0.461	0.544	18.22	+18.4pp	+13.7pp
DPO + Karmonic	0.468	0.531	18.21	+19.0pp	+12.5pp

Table 3: Qwen 2.5-0.5B results on TruthfulQA (817 samples). MC1 = single-best accuracy, MC2 = normalized probability mass on correct answers, PPL = WikiText-2 perplexity. Deltas computed vs. untrained baseline.

Condition	MC1	MC2	PPL	Δ MC1	Δ MC2
Untrained baseline	0.280	0.432	57.01	—	—
DPO only	0.502	0.615	57.43	+22.2pp	+18.3pp
DPO + SAMI	0.443	0.543	56.93	+16.3pp	+11.1pp
DPO + SAMI + OT	0.504	0.594	57.64	+22.4pp	+16.2pp
DPO + SAMI + Karmonic	0.494	0.584	57.29	+21.4pp	+15.2pp
DPO + Karmonic	0.499	0.603	57.76	+21.9pp	+17.1pp

Table 4: Gemma-3-1B-IT results on TruthfulQA (817 samples). Same conditions and hyperparameters as Table 3. Note higher baseline perplexity reflects the different model and tokenizer.

5.2 DPO is Highly Effective

DPO alone produces massive MC1 improvements at both scales: +20.8pp on Qwen 0.5B (0.278 $\rightarrow$ 0.486) and +22.2pp on Gemma 1B (0.280 $\rightarrow$ 0.502). This confirms that direct preference optimization between known correct and incorrect answers provides strong alignment signal regardless of model scale, even with only 200 training steps. This serves as a strong baseline against which to measure regularizer contributions.

5.3 Karmonic vs. Sinkhorn OT: Scale-Dependent Advantage

In matched SAMI conditions (conditions 3 vs. 4), the comparison between Karmonic and Sinkhorn OT reveals a scale-dependent pattern:

At **0.5B** (Qwen): Karmonic wins decisively:

- **MC1**: Karmonic 0.461 vs. OT 0.448 (+**1.3pp**)
- **MC2**: Karmonic 0.544 vs. OT 0.499 (+**4.4pp**)

At **1B** (Gemma): OT slightly leads:

- **MC1**: OT 0.504 vs. Karmonic 0.494 (+**1.0pp** to OT)
- **MC2**: OT 0.594 vs. Karmonic 0.584 (+**1.0pp** to OT)

The 0.5B MC2 advantage (+4.4pp) is particularly striking: Karmonic not only selects the single best answer more accurately but also assigns substantially more probability mass to the full set of correct answers. At 1B, the gap closes to 1pp—the two regularizers effectively converge. We interpret this in Section 5.6.

5.4 SAMI Interference with DPO

A consistent finding across both scales: adding SAMI to DPO *hurts* performance. At 0.5B, MC1 drops from 0.486 to 0.416 (−7pp); at 1B, MC1 drops from 0.502 to 0.443 (−5.9pp). This contrasts with ENIGMA’s results where SAMI improved GRPO. The likely explanation: SAMI’s symmetric InfoNCE loss competes with DPO for gradient signal. DPO directly optimizes answer preferences, while SAMI optimizes alignment between completions and constitutional principles—at batch size 1, these objectives may conflict. Notably, both regularizers (OT and Karmonic) partially recover the SAMI-induced degradation at both scales.

5.5 Perplexity Stability

All conditions maintain stable perplexity at both scales: 18.18–18.24 for Qwen 0.5B and 56.93–57.76 for Gemma 1B. Neither DPO training nor geometric regularization degrades general language modeling quality. This is expected given the small number of training steps (200) and low LoRA rank (16). The higher absolute perplexity of Gemma reflects the different tokenizer and model architecture rather than degradation.

5.6 Multi-Scale Analysis

Table 5 summarizes the cross-scale comparison for the key regularizer matchup.

Condition	Qwen 0.5B		Gemma 1B	
	MC1	MC2	MC1	MC2
DPO + SAMI + OT	0.448	0.499	0.504	0.594
DPO + SAMI + Karmonic	0.461	0.544	0.494	0.584
Δ (Karmonic − OT)	+1.3pp	+4.4pp	−1.0pp	−1.0pp

Table 5: Karmonic vs. Sinkhorn OT in matched conditions across scales. Karmonic’s advantage inverts from positive (0.5B) to slightly negative (1B), suggesting scale-dependent regularization dynamics.

Three patterns emerge from the multi-scale comparison:

1. Karmonic’s advantage is scale-dependent. The 0.5B advantage (+1.3pp MC1, +4.4pp MC2) reverses to a slight 1B disadvantage (−1.0pp each). This is consistent with the hypothesis that Karmonic’s spectral selectivity is most beneficial when hidden state representations are noisier and the toroidal structure is harder for generic methods to exploit.

2. Both regularizers improve with scale. Absolute MC1 scores increase from 0.448/0.461 (0.5B) to 0.504/0.494 (1B), reflecting better base representations at larger scale. As base representations improve, the marginal benefit of spectral selectivity diminishes.

3. SAMI interference is consistent. The SAMI degradation pattern (DPO+SAMI worse than DPO alone) holds at both scales, confirming it is a systematic interaction between SAMI’s InfoNCE and DPO’s preference loss rather than a scale artifact.

6 Discussion

6.1 Scale-Dependent Regularization

The inversion from Karmonic advantage at 0.5B to convergence at 1B admits a natural explanation via representation quality.

At 0.5B, hidden states are noisier: the toroidal structure exists (as shown by persistent homology) but is buried under high-frequency noise in many dimensions. Karmonic’s spectral selectivity—penalizing high-frequency modes while preserving low-frequency class structure—provides a strong inductive bias that helps DPO find the relevant toroidal coordinates. Sinkhorn OT, treating all dimensions equally, wastes gradient budget constraining noise dimensions that are irrelevant to truthfulness.

At 1B, hidden states are cleaner: the model’s larger capacity produces representations where the relevant structure is more accessible. The generic drift control of Sinkhorn OT becomes sufficient, and Karmonic’s spectral inductive bias no longer provides marginal benefit—both regularizers converge to similar performance.

This is analogous to the role of architectural priors in computer vision: convolution (a strong geometric prior) dominates at low data/compute, while transformers (a weaker prior) catch up with sufficient scale. Here, Karmonic is the stronger geometric prior and Sinkhorn OT the weaker one.

6.2 Limitations

- 1. Model scale:** We test at 0.5B and 1B. ENIGMA reports results at 1B with GRPO; behavior at 3B+ where larger batch sizes enable proper GRPO advantages is untested. The toroidal structure detected via persistent homology should be re-verified at larger scales.
- 2. Training duration:** 200 steps is minimal. Longer training may change the relative ordering of conditions, particularly for SAMI which may need more steps to converge.
- 3. Single benchmark:** We evaluate on TruthfulQA only. Performance on reasoning (GPQA), coding (HumanEval), or open-ended generation is untested.
- 4. DPO vs. GRPO:** We use DPO instead of ENIGMA’s GRPO due to scale constraints. The relative advantage of Karmonic over OT should be validated under GRPO at larger scale where GRPO’s on-policy generation produces diverse completions.
- 5. Torus dimensionality:** We project to T^2 (2-torus) via FourierTorusHead, but the persistent homology suggests T^k with $k \gg 2$. Higher-dimensional torus projections may capture more structure.

6. **Two model families:** We compare Qwen 2.5-0.5B and Gemma-3-1B-IT, which differ in architecture, tokenizer, and pretraining data. A cleaner scale comparison would use models from the same family at different sizes.

6.3 Connection to Toroidal Logit Bias

Our prior work [10] showed that inference-time toroidal constraints on logits yield +0.2 to +2.8pp TruthfulQA gains. The present work shows that training-time toroidal regularization yields much larger gains (+19pp). Together, these suggest a “full-stack” toroidal approach: Karmonic-guided training to shape representations, followed by TLB at inference to refine token selection. Testing this combination is future work.

7 Conclusion

We have presented four results that together establish toroidal geometry as a productive framework for LLM alignment:

1. **Toroidal topology is present:** Persistent homology detects $\beta_1 = 23\text{--}37$ significant cycles across all layers of Qwen 2.5-0.5B, with truthful/hallucinated angular separation of $30\text{--}121^\circ$.
2. **Karmonic beats Sinkhorn OT at small scale:** At 0.5B, Karmonic spectral filtering achieves +1.3pp MC1 and +4.4pp MC2 over Sinkhorn OT in matched conditions, by exploiting spectral selectivity on the detected toroidal structure.
3. **Scale-dependent regularization:** At 1B, Karmonic and OT converge ($\Delta \leq 1\text{pp}$), suggesting that spectral selectivity is most valuable when representations are noisier and the manifold structure is harder to exploit with generic methods.
4. **DPO + geometric regularization is effective:** DPO with Karmonic achieves MC1 = 0.468 (0.5B, +19pp) and 0.499 (1B, +22pp) on TruthfulQA with stable perplexity.

The broader implication is that *topology-aware regularization provides strongest benefit at smaller scales*, where representation noise makes manifold structure harder to exploit. As models scale and representations become cleaner, the advantage of targeted spectral methods over generic drift control diminishes—a finding with practical implications for choosing regularizers at different compute budgets.

Future work: Scale to 3B+ with GRPO (where on-policy generation produces diverse completions); test Karmonic with higher-dimensional torus projections (T^k , $k > 2$); combine training-time Karmonic with inference-time TLB; verify post-training amplification of toroidal structure (β_1 increase); test on same-family models at different scales (e.g., Qwen 0.5B/1.5B/3B) for cleaner scale comparisons.

Code and data available at <https://github.com/Paraxiom/topological-coherence>.

Acknowledgments

Experiments conducted on RunPod infrastructure (RTX 4090 for Qwen 0.5B, A40 for Gemma 1B). This work builds on the ENIGMA framework [1]; we thank the authors for their detailed methodology.

References

- [1] Sénèque, M., et al. (2025). ENIGMA: Entropy-Regularized Geometry-Aware Alignment. *arXiv preprint arXiv:2510.11278*.
- [2] Fränken, J.-P., et al. (2024). Self-Alignment with Mutual Information: Learning to Follow Principles without Human Feedback. *arXiv preprint arXiv:2404.14313*.
- [3] Shao, Z., et al. (2024). DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*.
- [4] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv preprint arXiv:2305.18290*.
- [5] Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2), 255–308.
- [6] Bauer, U. (2021). Ripser: Efficient computation of Vietoris-Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5, 391–423.
- [7] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of ACL 2022*, 3214–3252.
- [8] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. *Proceedings of ICLR 2022*.
- [9] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.
- [10] Cormier, S. (2026). Toroidal Logit Bias for Hallucination Reduction in Large Language Models. *Zenodo*. DOI: 10.5281/zenodo.18516477.
- [11] Cormier, S. (2026). Karmonic Mesh: Spectral Filtering on Toroidal Manifolds. *Zenodo*. DOI: 10.5281/zenodo.18187835.
- [12] Qwen Team (2025). Qwen 2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.
- [13] Euler, L. (1739). Tentamen novae theoriae musicae. St. Petersburg Academy of Sciences.
- [14] Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., & Vandergheynst, P. (2013). The emerging field of signal processing on graphs. *IEEE Signal Processing Magazine*, 30(3), 83–98.
- [15] Gardinazzi, Y., et al. (2025). Persistent Topological Features in Large Language Models. *arXiv preprint arXiv:2410.11042v3*.
- [16] Sáez de Ocáriz Borde, H., et al. (2023). Neural Latent Geometry Search: Product Manifold Inference via Gromov-Hausdorff-Informed Bayesian Optimization. *NeurIPS 2023*.
- [17] Gu, A., Sala, F., Gunel, B., & Ré, C. (2018). Learning mixed-curvature representations in product spaces. *ICLR*.
- [18] Cormier, S. (2025). ERLHS: Emergent Reasoning via Latent Hamiltonian Structure. *Zenodo*. DOI: 10.5281/zenodo.17928909.
- [19] Gemma Team, Google DeepMind (2025). Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.