Spectral Triplet for Storyworlds — A Decoder & Evaluation Framework

Authors: Patrick Dugan, et al.
Draft: v0.1 (2025-09-24)

Abstract

We propose a non-commutative geometric formalism for narrative environments ("storyworlds") that simultaneously (i) yields a principled decoder from authored structure to latent manifold and (ii) defines concrete evaluation tasks for long-horizon reasoning and interpretability. Modeling a storyworld as a spectral triplet $\mathbf{S} = (\mathcal{A}, \mathcal{H}, D)$, where $\mathcal{A}$ is a narrative algebra, $\mathcal{H}$ a Hilbert space of states, and D a Dirac-like narrative operator, we derive a spectral metric, operator recovery objectives, and hermeneutic probes that bridge quantitative and qualitative assessment. The framework generalizes rotary/phase encodings from positional to semantic modes, enabling Fourier/QFT-style analysis of motifs and authorial "voice." We release a minimal synthetic suite and scoring rubric covering compression, operator identification, spectral reconstruction, long-horizon consistency, and interpretive essays.

---

1. Introduction

LLMs increasingly accept long contexts, but raw window size poorly predicts quality on narrative coherence, authorial voice, and long-horizon constraints. We argue for **structure-aware evaluation**: treat storyworlds as formal objects with an intrinsic geometry that models can compress, reconstruct, and interpret. Our contribution:

1. **Formalism:** a spectral triplet for storyworlds that induces distances, resonances, and dynamics.
2. **Decoder:** a pathway from authored JSON (variables, gates, effects) to a low-dimensional Hilbert manifold with interpretable motif probes.
3. **Benchmark:** five tasks—Compression, Operator Recovery, Spectral Reconstruction, Long-Horizon Consistency, and Hermeneutic Essay—with clear metrics and a minimal synthetic suite.

We position this work as complementary to scaling context windows: models that succeed must internalize operator algebra and semantic phase, not merely attend across more tokens.

---

2. Background & Related Work (brief)

Non-commutative geometry: Spectral triples $(\mathcal{A}, \mathcal{H}, D)$ reconstruct geometry from operator spectra.
Mechanistic interpretability & SAEs: Sparse features provide "confessions" of hidden concepts; our motif probes play an analogous role for narrative.
Long-context evals: Retrieval-heavy benchmarks test access, not semantics; we test manifold recovery and

meaning.

Narrative engines: Formal storyworlds (encounters, variables, gates) offer a natural ground truth for causal structure. Foundational precedents include Chris Crawford's *On Interactive Storytelling* and Mateas & Stern's *Façade* (interactive drama), which formalize encounters, gating/logics, and drama management.

---

3. Formalism: Spectral Triplet for Storyworlds

We model a storyworld as $\mathbf{S} = (\mathcal{A}, \mathcal{H}, D)$.

3.1 Narrative Algebra $\mathcal{A}$

Let variables $\mathcal{V} = \{v_1, \ldots, v_m\}$ and encounters $\mathcal{E} = \{e_i\}$ have gate predicates $g_i : \mathbb{R}^m \to \{0, 1\}$ and effects $f_i : \mathbb{R}^m \to \mathbb{R}^m$.

$\mathcal{A}$ is the unital *-algebra generated by projections P_i (gates), effect operators U_i (updates), and motif probes M_k (linear/composite observables). Non-commutativity $U_iU_j \neq U_jU_i$ captures order-sensitive causality.

3.2 Hilbert Space $\mathcal{H}$

Embed state $x \in \mathbb{R}^m$ via $\phi(x) \in \mathbb{C}^d$ learned by PCA/SAE; complete spans to obtain $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$. Characters and spools define subspaces.

3.3 Narrative Dirac D

D is self-adjoint and (i) induces Connes' spectral metric

$$d(\omega_x, \omega_y) = \sup_{\|[D,a]\| \leq 1} |\omega_x(a) - \omega_y(a)|, \quad \omega_x(a) = \langle \phi(x), a\, \phi(x) \rangle,$$

(ii) parameterizes local dynamics via generators H_i with $U_i = \exp(i \Delta t H_i)$. The spectrum $\{(\lambda_k, \psi_k)\}$ encodes motif resonances; energies $E_k = \langle \psi_k, M\, \psi_k \rangle$.

---

4. Decoder & Spectral Analysis

1. **State embedding:** learn ϕ on rollouts across spools; target $d \in [128, 512]$.
2. **Operator identification:** regress $\hat{H}_i$ from observed state deltas; validate commutators and gate F1.
3. **Dirac fitting:** choose $\hat{D}$ to align empirical distances with the spectral metric under a budget $\|[\hat{D}, a]\| \leq 1$.
4. **Motif probes:** train linear/composite M_k for interpretable axes (e.g., betrayal, redemption, debt relief).
5. **Spectral tuples:** report $\{\hat{\lambda}_k\}$, energies, and phase relations; compare to ground truth or curator labels.

---

5. Evaluation Protocol

We propose five tasks with standardized metrics.

Task A — Compression
Reconstruct states from $\phi(x)$ *.*
Metrics: MSE $\downarrow$, cosine@k $\uparrow$, sparsity (L0/L1) $\uparrow$, CKA vs baseline $\uparrow$.

Task B — Operator Recovery
Estimate $\hat{U}_i \approx e^{i\Delta t \hat{H}_i}$ *and gates* $\hat{P}_i$ *from trajectories.*
Metrics: next-state NRMSE $\downarrow$, commutator error $\downarrow$, gate F1 $\uparrow$.

Task C — Spectral Reconstruction
Match spectra and motif energies.
Metrics: Wasserstein(λ) $\downarrow$, motif R^2 $\uparrow$, distance correlation on Connes pairs $\uparrow$.

Task D — Long-Horizon Consistency
Roll out plans under motif constraints; measure spectral drift.
Metrics: spectral drift $\downarrow$, failure@N $\downarrow$, constraint satisfaction $\uparrow$.

Task E — Hermeneutic Essay
Write 500–1000 words mapping spectra to themes and authorial intent.
Metrics: expert rubric (argument, evidence, alignment), citation rate to modes, retrieval of planted motifs.

---

6. Minimal Synthetic Suite

- Variables $x = (\mathrm{Trust}, \mathrm{Debt}, \mathrm{Resolve})$.
- Encounters: Betrayal (Trust $\downarrow$, Resolve $\uparrow$), Atonement (Debt $\downarrow$, Trust $\uparrow$), Temptation (phase shift on Trust).
- Gates as half-spaces; block-circulant D with modes aligned to encounters.
- Release JSON + 10k rollouts; publish planted spectra and probes.
- Baselines: random features vs. SAE + linear probes.

---

7. Experiments (planned)

1. **Ablations:** embedding dim, probe types, noise on gates/effects.
2. **Model classes:** GPT-x, Qwen-x, recurrent/state models, kernel/log(n) attention, Power-Attention-style.
3. **Data regimes:** synthetic → curated human storyworlds (Crawfordian), varying motif density and deception.
4. **Generalization:** train on spools A,B; test on held-out spool C.

---

8. Discussion

Our spectral view elevates semantic phase and operator algebra to first-class citizens. Unlike window-size stress tests, these tasks reward models that internalize narrative geometry. The Connes metric supplies a principled distance; commutators diagnose causal order sensitivity.

---

9. Limitations

- Fitting D may be underdetermined without curated constraints.
- Human rubric for essays introduces subjectivity; we mitigate with planted motifs and citations.
- Synthetic suites risk over-regularity; we plan diverse human-authored worlds.

---

10. Ethics & Broader Impact

Interpretability gains can improve safety (detecting incoherent or manipulative arcs) but also enable persuasive systems. We recommend disclosure policies and caps on motif-targeted manipulation in deployed agents.

---

11. Conclusion

We introduced a spectral triplet formalism and a companion benchmark that jointly test compression, operator learning, spectral recovery, long-horizon coherence, and hermeneutic skill. We invite contributions of open storyworlds and will maintain a leaderboard with standardized reports.

A. Notation (quick reference)

- $\mathcal{A}$: narrative algebra (gates P_i , effects U_i , probes M_k).
- $\mathcal{H}$: Hilbert space of embedded states $\phi(x)$.
- D : narrative Dirac; $[D, a]$ commutator; Connes distance d .
- Spectral data: (λ_k, ψ_k) , energies E_k .

References

[1] A. Connes. *Noncommutative Geometry* (and spectral triples background). [2] Sparse Autoencoders for Interpretable Features (representative SAE literature). [3] Long-Context Language Model Evaluations (representative benchmarks). [4] Formal Storyworld Engines and Encounter Grammars (representative foundations). [5] Chris Crawford. *On Interactive Storytelling*, 2nd ed., 2013. [6] Michael Mateas and Andrew Stern. *Façade*: An Experiment in Building a Fully-Realized Interactive Drama, 2005 (project & papers).