Storyworlds as Sparse Autoencoders — Benchmark & Training Protocol

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

Abstract

We propose storyworlds as structured corpora for discovering sparse features and evaluating long-horizon reasoning in LLMs. Treating encounters, gates, and effects as generators of a latent manifold, we define tasks that (i) induce **Sparse Autoencoder (SAE)** features aligned with narrative motifs, (ii) test generalization via multi-spool worlds with secret endings, and (iii) enable **QFT-style decoders** to read spectral patterns from large-context rollouts. We introduce an agent ensemble (Author-Bots) tuned to different optimization criteria—**Tolstoy** (character depth), **Finemann** (variable richness), and **Tokien** (ending multiplicity)—to generate controlled datasets spanning motif density and combinatorial branching. We release evaluation rubrics spanning compression, manifold recovery, puzzle-language competence, and hermeneutic essays.

---

1. Motivation

Long context alone does not guarantee narrative coherence, motif control, or interpretable internal state. Storyworlds offer a ground-truth operator algebra—**Nudge, Blend, Reverse, Proximity**—over variables and gates. SAEs applied to rollouts reveal sparse features (motifs), while QFT-style analysis exposes spectral structure beyond token locality.

---

2. Storyworld Formalism (concise)

- **Variables**: $x \in \mathbb{R}^m$ (traits, debts, loyalties, beliefs).
- **Encounters**: $e_i = (g_i, f_i)$ with gate predicate $g_i : \mathbb{R}^m \to \{0,1\}$ and effect $f_i : \mathbb{R}^m \to \mathbb{R}^m$.
- **Options & Reactions**: deterministic reactions with weighted effects; after-effects include *Set*, *Nudge*, *Blend*, *Reverse*, *Proximity* (v1.9 Sweepweave-compatible).
- **Spools & Endings**: DAG of encounters partitioned by spools; endings defined by Boolean/weighted formulae.

---

3. Author-Bot Ensemble

We define three controllable **Author-Bots** to generate worlds under distinct objectives: - **Tolstoy (Character Maximizer)**: maximize per-character manifold depth; high mutual information between encounters and

character vectors. - **Finemann (Variable Maximizer)**: maximize variable count/entropy and inter-variable coupling; stress SAEs with high-dimensional sparsity. - **Tokien (Ending Maximizer)**: maximize count/diversity of endings and secret branches; emphasize gating logic and reachability constraints.

Each bot exposes knobs for motif density, deception (false affordances), and cross-spool allusions.

---

4. Benchmark Tasks

A. SAE Compression & Alignment
Train SAEs on rollout states; evaluate sparse feature interpretability against planted motif probes.
Metrics: recon MSE, feature sparsity (L0/L1), probe R^2 , CKA vs baseline.

B. Manifold Recovery & Operator Identification
Estimate local generators from observed Δx ; validate commutators and gate classification.
Metrics: next-state NRMSE, commutator error, gate F1.

C. Secret Endings Discovery (Needle-in-Haystack)
Evaluate ability to discover planted endings with and without hints (breadcrumbs).
Metrics: success rate@N episodes, sample complexity, regret.

D. Constructed Languages & Puzzle Ciphers
Introduce invented language variants and deterministic machine-code puzzles embedded in encounters.
Metrics: translation accuracy, cipher solve rate, transfer to new spools.

E. Spectral Decoder (QFT)
Apply Fourier/QFT to trajectories in latent space to recover motif spectra and phase relations.
Metrics: Wasserstein(λ), spectral coherence, phase stability under perturbations.

F. Hermeneutic Essay
Generate an interpretive analysis mapping sparse features and spectra to themes and authorial intent.
Metrics: expert rubric (argument, evidence), motif citation density, consistency across spools.

---

5. Dataset Generation Protocol

1. **Bot sampling:** draw worlds from Tolstoy/Finemann/Tokien with curriculum on motif density and branching factor.
2. **Sweepweave v1.9 compliance:** all after-effects wrapped in *Set* with operator expressions.
3. **Hints curriculum:** train/test splits with/without breadcrumbs for secret endings.
4. **Language variants:** inject invented language and machine-code puzzles with paired keys for evaluation.

---

6. Training Recipes

- **SAE stage:** train sparse encoders on states and token-level summaries; select $d \in [128, 512]$.
- **Decoder stage:** train linear probes for motifs; fit local generators; optional spectral Dirac fitting (see companion Spectral Triplet paper).
- **RL fine-tune:** optimize agents for endings discovery under constraint satisfaction; regularize spectral drift.

---

7. Evaluation & Leaderboard

Report per-task metrics plus a composite score:

$$\text{Score} = \alpha A + \beta B + \gamma C + \delta D + \epsilon E + \zeta F,$$

with weights set by target use (interpretability vs. exploration).

---

8. Extensions

- **Multi-spool generalization:** train on spools A,B; test on held-out C.
- **Adversarial probes:** perturb gate thresholds; inject decoy motifs.
- **Human curation:** include Crawfordian worlds; integrate drama-manager baselines.

---

9. Related Work

- **Interactive narrative:** Chris Crawford, *On Interactive Storytelling*.
- **Interactive drama:** Mateas & Stern, *Façade*.
- **Mechanistic interpretability:** SAEs exposing latent features.
- **Long-context evals:** retrieval-heavy vs. semantic-structure benchmarks.

---

10. Ethics

Managing persuasive capacity and deceptive affordances; require disclosure and guardrails for motif-targeted conditioning.

---

11. Conclusion

Storyworlds provide a controllable substrate to surface sparse features and test semantic long-horizon reasoning. By combining SAEs, operator identification, secret-ending discovery, constructed languages, and QFT-style decoding, we offer a concrete research path beyond raw context scaling.