Storyworlds as Sparse Autoencoders

Benchmark, Hypergames, & Introspective 1k-Dimensional Worlds

Patrick Dugan & GPT-5.1 Thinking

Draft v0.4 — November 2025

Abstract

We propose storyworlds as structured corpora for inducing sparse, interpretable features analogous to those learned by sparse autoencoders (SAEs), and as substrates for evaluating long-horizon reasoning, operator recovery, and multi-agent communication in language models.

Formal Crawfordian storyworlds—directed acyclic graphs (DAGs) of encounters with explicit variables, gates, and deterministic after-effects—define a tractable "ground-truth" manifold of latent causes. We argue that this geometry can play for external behavior what sparse feature decompositions play for internal activations: a dictionary of relatively monosemantic narrative features with a small operator algebra.

Building on recent work in dictionary-learning SAEs, grokking, and sparse distributed reasoning in large models, we develop a benchmark suite with three pillars: (i) storyworlds as external SAEs, (ii) hypergames over high-dimensional KV-manifolds where agents hold mismatched game models, and (iii) introspective 1k-dimensional storyworlds that models are prompted to design as putative encodings of their own circuitry.

The benchmark includes tasks for sparse compression and motif alignment, operator identification, secret-ending discovery, constructed languages and mini-codes, spectral decoding of trajectories, hypergame communication, and interpretive essays. We provide concrete dataset and training protocols, and outline how this framework connects mechanistic interpretability, multi-agent systems, and narrative evaluation in a unified setting.

1 Introduction

Large language models (LLMs) now operate over hundreds of thousands of tokens of context. Yet mainstream long-context benchmarks emphasize retrieval and span robustness rather than structured, causal reasoning over well-defined latent state.

At the same time, mechanistic interpretability has begun to reverse engineer large models from the inside. Dictionary-learning SAEs trained on hidden activations reveal sparse, often monosemantic features that can be linked to concepts, syntactic patterns, or algorithmic subroutines. These features hint at an internal geometry: a manifold of recurring structures that the model uses to implement tasks.

In this work we bring these two threads together using *storyworlds*. By storyworld we mean a formally specified interactive narrative: a set of scalar variables, encounters with gates and effects, spools and endings, and a small operator basis (Set, Nudge, Blend, Reverse, Proximity) implemented in a tooling-friendly schema (e.g., Sweepweave v1.9). Each storyworld induces a manifold of reachable states and a small algebra of operators that act on those states.

The central idea is simple:

> If SAEs give us sparse *internal* features, storyworlds can give us sparse *external* features and operators that we design and control. If this is useful as a research lens for humans it can also be a useful means of communication between agents.

This external structure allows us to pose questions that are otherwise hard to formalize:

- Does a model learn sparse features that align with interpretable narrative motifs?
- Can we recover the operator algebra (Nudge/Blend/Reverse/Proximity) from trajectories alone?
- Can agents discover rare secret endings efficiently, or are they trapped in local attractors?
- Do multi-agent models exploit the storyworld's KV-manifold as a communication channel?
- Can models design high-dimensional storyworlds that faithfully encode their own circuit features?

By framing these questions in a single benchmark, we aim to create a bridge between: (i) mechanistic interpretability and dictionary-learning SAEs, (ii) interactive narrative and storyworld design, and (iii) multi-agent reasoning and hypergame theory.

We deliberately *do not* assume heavy internal access to the model. Most tasks treat the model as a black box with access to the storyworld API and optionally to external SAEs fitted on world states or transcripts. Where internal activations are used (for the introspective 1k-dimensional storyworlds), we restrict ourselves to coarse-grained feature slices rather than full internal tracing.

2 Background and Motivation

2.1 Sparse Features and Dictionary Learning

The trade-off between safety and capability in allowing AI agents to develop alien internal languages for reasoning is a compelling dilemma. It is a central theme in the AI 2027 scenario for safe vs. unsafe automated AI research [1]. In that scenario, competitive pressure nudges actors toward ever-more opaque systems, even when this sacrifices interpretability, while embracing interpretability enables human thriving.

However, this dilemma may have a third option.

Dictionary-learning SAEs trained on LLM activations perform two roles at once: they compress activations and induce a learned dictionary of sparse features. Empirically, many units are at least *locally* interpretable, clustering around particular concepts, syntactic constructs, or algorithmic roles. This is now well documented across several strands of work:

- scaling laws and evaluation metrics for large SAEs trained on GPT-4 activations [2];
- global feature maps and steering via SAEs in production-scale models [4];
- data-centric analysis using SAE embeddings for dataset diffing, correlations, and retrieval [5];
- and SAE-guided tuning that elicits modular, portable reasoning abilities in small models [6].

The promise of this line of work is that it may eventually let us describe model cognition in terms of a relatively small number of reusable patterns. But SAEs by themselves do not specify an external environment. They tell us that the model has some internal dictionary; they do not tell us how well that dictionary is aligned with a given task's structure. Additionally the scale

dynamics of LLMs beyond one hundred billion parameters makes pure mechinistic interpretability seem impossible as a sheer solution.

In contrast, storyworlds allow us to design an explicit manifold of variables and operators. This sets up a natural external dictionary-learning problem: find sparse features that align with motifs we planted in the world and operators we defined by hand.

Empirical work on sparse feature geometry and representation manifolds motivates our use of storyworlds as designed manifolds. Li et al. [7] show that SAE features in large language models do not form an unstructured cloud but a highly organized "concept universe": at small scales, features assemble into analogy-like "crystals" (generalizing man:woman::king:queen); at intermediate scales they form neighborhoods of related concepts; and at global scales they exhibit coherent clustering structure. Modell et al. [8] provide a complementary mathematical perspective, treating features as manifolds embedded in representation space and establishing conditions under which cosine similarity encodes the intrinsic geometry of a feature via shortest on-manifold paths. Taken together, these results support the view that sparse features are geometric objects living on structured manifolds. Our benchmark builds on this view by constructing storyworlds as explicit, controllable manifolds with a small operator basis; we then ask whether agents and external SAEs can recover the planted geometry and operator algebra from trajectories.

This operator-centric view is loosely reminiscent of the noncommutative geometry program of Connes and Marcolli [9], which recovers geometric structure from operator algebras and spectral data, although our setting here remains entirely classical and commutative.

2.2 Grokking and Structure Formation

Work on grokking provides a complementary lens on how sparse, structured representations emerge during training. Nanda et al. [10] show that models trained on modular arithmetic tasks often pass through a "memorization phase" where they fit the data with diffuse, non-algorithmic patterns, before abruptly transitioning to a more compressed, algorithmic representation that generalizes out of distribution. This transition can be tracked with progress measures derived from mechanistic interpretability, including circuit-level probes and feature sparsity.

Subsequent work by Nanda and collaborators on circuit discovery and neuron-to-graph methods [11] reinforces the view that large models can reorganize around relatively clean algorithmic structures once training pushes them into the right basin. From the perspective of storyworlds, we can view this as a shift from treating encounters as loosely related text patterns to treating them as elements of a coherent operator algebra over a latent manifold.

Our benchmark is deliberately designed to surface such transitions. Early in training or with weaker agents, we expect behavior that looks like surface-level pattern matching: treating each encounter as an unrelated prompt. As agents improve, we expect to see evidence of "narrative grokking": the emergence of stable, reusable motifs and operator regimes that support long-horizon planning and generalization across spools. External SAEs trained on world states provide a way to track this shift in terms of sparsity, motif alignment, and operator consistency.

We can view storyworlds as a testbed for this phenomenon at the narrative level. Early in training, an agent may thrash around, following local cues and heuristics; later, it may form *stable* motifs and schema about the world: betrayal arcs, debt and forgiveness cycles, political coalitions. A sparse representation that successfully disentangles these motifs should support more efficient planning and generalization across spools and settings.

2.3 Storyworlds and Interactive Narrative

Interactive narratives and storyworld engines have explored complex character-driven simulations and branching plots for decades. From Chris Crawford's work on interactive storytelling to projects like *Façade*, we have compelling examples of systems that prioritize dynamic character modeling and emergent arcs over fixed branching choices.

What is missing is an explicit connection between these narrative systems and the internal representational geometry of modern LLMs. Storyworlds in this paper are intentionally minimalistic: we do not attempt to model all the richness of human narrative, but instead design simple, compositional operators and metrics that are friendly to systematic analysis and benchmarking.

2.4 Multi-Agent Reasoning and Hypergames

In realistic environments, agents rarely share the same game model. Each agent observes partial state, has its own beliefs about others' beliefs, and often mis-specifies the rules. Hypergame theory formalizes this: each agent plays its own game on top of the shared environment.

Storyworlds give us a natural way to instantiate hypergames: agents can have different partial views of the world variables, different subsets of encounters, and different utility functions. More importantly, we can exploit the storyworld's high-dimensional state as a KV-manifold that agents can leverage for implicit communication, deception, and coordination.

This differs from classical communication games where channels are explicit. Here communication is embedded in how agents manipulate seemingly ordinary narrative variables (loyalty, debts, rumors, etc.), and decoding these signals requires understanding both the world and the other agents' policies.

3 Storyworld Formalism

3.1 Interactive Storytelling and Topological Structure

Our use of storyworlds as manifolds with operators is closely related to the tradition of interactive storytelling inaugurated by Crawford and Mateas & Stern. In *On Interactive Storytelling*, Crawford [**?**] emphasizes that interactive stories are not linear plots but *possibility spaces*: networks of situations and actions connected by author-designed transitions. His focus on verbs, processes, and the "conversation" between player and system is implicitly topological: the drama emerges from how the player moves through a space of narrative possibilities, not from a fixed script.

Mateas and Stern's *Façade* [**?**] makes this topology explicit in implementation. Their drama manager orchestrates a dense web of beats, behaviors, and triggers over an underlying state representation (relationship tensions, topics, emotional variables). The experience of a single playthrough is a path through this graph; different choices traverse different regions of the space, with attractors, bottlenecks, and fragile branches that resemble the "secret endings" and gated regions in our benchmark worlds.

Our storyworld formalism can be seen as a deliberate abstraction of this lineage. Crawford's possibility space and *Façade*'s beat graph are recast as a state manifold X equipped with variables, gates, and a small operator basis (Set, Nudge, Blend, Reverse, Proximity). Rather than treating topology as an implicit property of a large hand-authored behavior graph, we make it an explicit object of study: which regions of X are reachable, how encounters partition and connect those regions, and how agents' trajectories explore or ignore parts of the space. This allows us to bring tools from SAE-based interpretability and representation geometry to bear on questions that Crawford

and *Façade* posed informally: how do agents perceive the shape of a storyworld, and what kinds of internal structures are required to navigate it coherently?

We now formalize the storyworlds used in the benchmark.

3.2 State Space

Let m be the number of scalar variables in the world. These may include:

- character-centric traits (courage, loyalty, suspicion),
- social ties (friendship, debt, alliance, grudge),
- global flags (war/peace, famine, public knowledge),
- hidden flags (secret pacts, blackmail material).

We concatenate all variables into a single vector $x \in \mathbb{R}^m$. For simplicity we assume real-valued variables, though discrete or bounded domains are allowed.

3.3 Encounters and Gates

An encounter e_i is defined as:

$$e_i = (g_i, f_i),$$

where $g_i : \mathbb{R}^m \to \{0, 1\}$ is a gate predicate and $f_i : \mathbb{R}^m \to \mathbb{R}^m$ is a deterministic transition function. The storyworld transition kernel is:

$$\mathcal{T}(x, e_i) = \begin{cases} f_i(x), & \text{if } g_i(x) = 1, \\ x, & \text{otherwise.} \end{cases}$$

Each encounter exposes one or more discrete options to the agent. Each option triggers a reaction (usually a deterministic function of x and the chosen option) followed by a sequence of after-effects drawn from a fixed operator basis.

3.4 Operator Basis

All after-effects are required to be expressible using a small set of primitives:

$$\mathcal{B} = \{\text{Set},\ \text{Nudge},\ \text{Blend},\ \text{Reverse},\ \text{Proximity}\}.$$

We treat these as a operator algebra acting on x.

Intuitively:

- **Set**: force a variable to a chosen value.
- **Nudge**: increment or decrement a variable by a small amount.
- **Blend**: mix two variables or a variable and a reference anchor.
- **Reverse**: negate or invert a variable.
- **Proximity**: move a variable towards a target value or direction.

In practice, all after-effects are wrapped in a Sweepweave v1.9-style `Set` container whose `to` field contains an operator expression built from $\mathcal{B}$. This keeps the JSON schema simple and guarantees that we can parse and analyze effects programmatically.

3.5 Spools and Endings

The encounter graph is partitioned into spools $S_1, \ldots, S_K$, each a DAG subgraph representing a coherent arc (e.g., "court intrigue", "desert expedition", "trial"). Endings are terminal regions defined by Boolean and/or weighted formulas over x.

For example, an ending might be:

$$\text{End:} \quad \text{KingDead} = 1 \wedge \text{Loyalty(Guard)} < 0.2.$$

The same state can satisfy multiple ending criteria; whether this is allowed or treated as a conflict is a design choice. For the benchmark we provide both single-ending and multi-ending regimes.

4 Storyworlds as External Sparse Autoencoders

We now make the analogy between storyworlds and SAEs precise.

4.1 State Manifold and External SAE

Let $X \subseteq \mathbb{R}^m$ be the set of reachable world states under some rollout distribution $\mathcal{D}$ (e.g., induced by human play, scripted agents, or LM agents). We train an SAE:

$$f_\theta : X \to Z, \quad g_\theta : Z \to X,$$

where $Z \subseteq \mathbb{R}^d$ with $d \geq m$ usually, to minimize:

$$\mathcal{L}_{\text{SAE}}(\theta) = \mathbb{E}_{x \sim \mathcal{D}} \left[\|x - g_\theta(f_\theta(x))\|_2^2 + \lambda \, \Omega(f_\theta(x)) \right],$$

with Ω a sparsity penalty (e.g., ℓ_1 or KL to a sparse prior).

Here the SAE is external: it operates on world states rather than internal model activations. This gives us a dictionary of features $\{z_j\}$ over X.

4.2 Operator-Aligned Representations

For each encounter e_i we define a corresponding latent-space operator $\tilde{T}_i : Z \to Z$. We say that $(\mathcal{T}, f_\theta, g_\theta)$ is *operator-aligned* if for all $x \in X$:

$$\|f_\theta(\mathcal{T}(x, e_i)) - \tilde{T}_i(f_\theta(x))\|_2 \leq \varepsilon_i,$$

for small ε_i.

In other words, the latent representation respects the world's operator algebra up to small errors. This mirrors the internal goal of dictionary-learning SAEs: find a basis that linearizes or simplifies the action of the model's primitive operations.

4.3 Motifs as External Features

A narrative motif M (e.g., betrayal, reconciliation, debt forgiveness) corresponds to a family of trajectories or states in X. Let $\chi_M(x)$ be an indicator or soft score for motif presence.

We call a latent unit z_j *motif-aligned* with M if:

- **Activation locality:** z_j is high primarily when $\chi_M(x) = 1$,

- **Operator support:** encounter operators associated with M act mainly on a small index set J_M containing j,
- **Commutator structure:** commutators between operators implementing M reflect the designed algebra when projected onto z_j and its neighbors.

This gives a concrete, testable analogue to monosemanticity: instead of features aligned to abstract tokens or directions in embedding space, we have features aligned to planted motifs and operators in a designed world.

4.4 SAEs, Grokking, and Storyworlds

We hypothesize a qualitative connection between grokking and motif formation in storyworlds:

- Early training: agents treat encounters as local text puzzles, ignoring deep structure; SAEs trained on X mostly find coarse topic clusters.
- Transition: agents begin to exploit long-range dependencies; SAEs discover sharper features that correspond to motifs and operator regimes.
- Mature stage: SAEs converge to stable, sparse dictionaries where many units are motif-aligned, and operators in latent space approximate $\tilde{T}_i$ well.

The benchmark includes training protocols and metrics designed to detect such transitions, though fully validating this picture is left to empirical work.

5 Author-Bot Ensemble

To systematically generate diverse but structured worlds, we use a trio of synthetic authors with distinct objectives.

5.1 Tolstoy: Character-Depth Maximizer

Tolstoy-like authors optimize for deep, entangled character manifolds:

- large per-character variable vectors,
- strong intra-character dynamics (internal conflict, moral drift),
- long-timescale arcs where earlier decisions resonate hundreds of steps later.

These worlds stress representation depth: agents must track evolving character states, not just local plot beats.

5.2 Finemann: Variable-Density Maximizer

Finemann authors maximize variable count and structured sparsity:

- many weakly coupled variables,
- occasional sharp interactions (e.g., a single encounter that couples 10 variables),
- heavy use of Nudge and Proximity to create rich local geometry.

These worlds stress the SAE: m is large, sparsity is high, and motifs are defined by multi-variable patterns.

5.3 Tokien: Branching and Secret-Endings Maximizer

Tokien authors maximize combinatorial complexity:

- high branching factor,
- multiple endings and sub-endings,
- hidden spools and secret endings requiring precise action sequences.

These worlds stress planning and exploration. Secret-ending discovery is the core of Task C in the benchmark.

5.4 Controllable Knobs

Each author-bot exposes dials for:

- motif density,
- deception level,
- spool count and inter-spool coupling,
- horizon length.

This allows us to construct curriculum families: from simple 3-variable, 1-spool worlds to 1000-variable, 10-spool worlds with hypergame structure.

6 Benchmark Tasks

We summarize the benchmark tasks and their evaluation metrics.

6.1 Task A: Sparse Compression & Motif Alignment

Train external SAEs on state sequences and optionally on state summaries generated by the agent or another model. Evaluate:

- reconstruction error (MSE, NRMSE),
- sparsity (L0/L1 statistics per code),
- motif probes: simple classifiers predicting motif indicators from z,
- CKA or similar similarity measures between SAE features and baselines (e.g., PCA).

6.2 Task B: Manifold Recovery & Operator Identification

Given trajectories (x_t), fit local generators or discrete operators that approximate f_i or $\tilde{T}_i$:

- next-state prediction error under recovered operators,
- gate prediction accuracy (precision, recall, F1),
- commutator error relative to ground-truth compositions.

This tests whether the model (or an external learner) can recover the environment's underlying algebra.

6.3 Task C: Secret Ending Discovery

Agents must discover rare endings requiring unlikely but structured sequences of actions.

Variants:

- no-hints,
- breadcrumb hints (faint clues),
- curriculum: hints that gradually fade.

Metrics include success@N episodes, sample complexity to reach 50% or 90% success, and regret relative to an oracle policy.

6.4 Task D: Constructed Languages and Mini-Codes

We embed conlangs and simple machine-like codes in storyworlds:

- morphological patterns embedded in signage, rituals, or rumors,
- small stack or register-based mini-languages controlling locks or consoles.

Train/test splits can hold out grammar families or code variants. Metrics include translation accuracy, code execution success, and generalization to unseen grammar/cipher instantiations.

6.5 Task E: Spectral Decoding of Trajectories

From latent trajectories (z_t) or state trajectories (x_t), we compute spectral representations (e.g., via discrete Fourier transform or other linear transforms). The goal is to capture motif-frequency patterns:

- differentiate spools and world types by their spectral signatures,
- assess stability of spectra under small state or policy perturbations,
- evaluate distance metrics (e.g., Wasserstein) between spool spectra.

We avoid quantum Fourier transform or more exotic machinery; simple discrete spectra suffice to test whether long-horizon structure is reflected in latent space.

6.6 Task F: Hermeneutic Essay

Given quantitative summaries (motif probes, spectral statistics, feature importance) and textual excerpts, the model writes a 500–1000 word essay interpreting the world:

- linking sparse features to narrative motifs,
- connecting spectral modes to recurring arcs,
- explaining character trajectories in terms of the environment's operator algebra.

Human or rubric-based grading captures interpretive quality, evidence use, and coherence. This task targets higher-level narrative understanding grounded in quantitative structure, not free-form fiction.

7 Dataset and Training Protocol

7.1 World Sampling

We generate families of worlds by sampling from Tolstoy/Finemann/Tokien distributions, with curriculum over:

- variable count m,
- spool count and maximum horizon,
- motif density and cross-spool coupling,
- secret-ending depth and rarity.

For each world we ship:

- a JSON specification of variables, encounters, operators, spools, and endings;
- one or more initial state seeds;
- reference policies (scripted, random, simple heuristics).

7.2 Rollout Regimes

We consider three rollout sources:

- human or human-in-the-loop agents,
- static LLM-based agents at fixed capability levels,
- RL-trained agents optimized for specific objectives (e.g., secret endings).

These rollouts supply trajectories $\mathcal{D}$ for external SAE training and for hypergame/communication tasks.

7.3 Hints, Conlangs, and Codes

Worlds can be configured with:

- no hints for a "cold" exploration regime,
- light hints embedded in narrative text,
- explicit hint tokens (e.g., marked prophecy or logbooks),
- conlang and mini-code elements that must be deciphered to unlock secret endings.

Conlang and code evaluation is modular: it can be skipped in environments where only narrative reasoning is desired.

8 Hypergames and KV-Manifold Communication

We now formalize storyworld hypergames and KV-manifold signaling.

8.1 Subjective Games

Let $X \subseteq \mathbb{R}^m$ be the true state space and $\mathcal{T}$ the true transition kernel. Each agent a maintains a subjective game:

$$\mathcal{G}_a = (X_a, \mathcal{E}_a, U_a),$$

where X_a is an observed subspace, $\mathcal{E}_a$ the subset of encounters believed to exist, and U_a a utility function over X_a.

Differences in X_a, $\mathcal{E}_a$, or U_a create hypergame structure: each agent is implicitly playing a different game on the same underlying world.

8.2 KV-Manifold as Communication Channel

With m large, X acts as a KV-manifold: coordinates encode various narrative facts, relationships, and hidden flags. An agent can exploit this manifold as a communication channel:

- choose a message M from some codebook,
- use actions to write M into a subset of variables $X_M \subset X$,
- rely on another agent's SAE or heuristics to decode $\hat{M}$ from observations or from its own latent trajectory.

We define a KV-communication task:

1. sample M,
2. constrain the sender agent A to a budget of encounters and actions,
3. allow the receiver B to act or observe over a fixed horizon,
4. measure accuracy $\Pr[\hat{M} = M]$, mutual information $I(M; z^B_{1:T})$,
5. track bits-per-action and robustness under hypergame mismatch.

8.3 Relation to Classical Multi-Agent Systems

Our formalism is closely related to the classical multi-agent foundations laid out by Shoham and Leyton-Brown (2009). Encounters correspond to local strategic-form games, spools define extensive-form trees, and hypergames arise when agents maintain subjective, possibly incorrect models of the underlying world. The KV-manifold functions as a differentiable signaling space, generalizing classical communication and cheap-talk models. Finally, the storyworld itself is a mechanism-designed environment in the sense of classical economic theory, with motifs and operator algebras acting as engineered incentive structures that probe sparse representations and emergent communication.

8.4 Feature Alignment Across Agents

Each agent can train its own SAE $f_{\theta_a} : X \to Z_a$. We can then ask:

- how similar are the induced feature spaces Z_a and Z_b?
- can we find a mapping $T_{a\to b} : Z_a \to Z_b$ such that $f_{\theta_b}(x) \approx T_{a\to b}(f_{\theta_a}(x))$?
- does successful communication correlate with high cross-agent feature alignment?

This connects interpretability and multi-agent coordination: if two agents cannot align their sparse feature spaces, we should expect communication failures even if they perform well individually.

8.5 Adversarial Hypergames

We can also define adversarial regimes where:

- the sender is rewarded for misleading the receiver while maintaining plausible storyworld behavior,
- the receiver is rewarded for robust decoding under distributional shifts and deceptive behavior,
- human or external critics evaluate whether the deception remains within acceptable bounds.

These tasks surface deceptive capabilities and failure modes in a concrete environment with explicit state and operator semantics.

9 Introspective 1k-Dimensional Storyworlds

The final component of the benchmark moves from external to quasi-internal grounding: models are prompted to design high-dimensional storyworlds that they claim correspond to their own internal features.

9.1 Circuit-Manifold Slice

Suppose we have fitted internal SAEs on selected layers or heads of a model and identified d features $\{c_1, \ldots, c_d\}$. We pick a subset of $m \approx 1000$ features that are:

- stable across dataset slices or tasks,
- relatively disentangled by intervention tests,
- associated with non-trivial behavior (e.g., semantic, syntactic, or planning features).

We denote this slice by $\mathcal{C}_{1\mathrm{k}}$ and treat it as a coarse-grained circuit manifold.

9.2 Introspective World Construction

We present the model with a structured description of $\mathcal{C}_{1k}$: examples of high/low activations, rough semantic labels, and intervention summaries. We then instruct it to:

1. define a storyworld $\mathcal{W}_{\text{hyper}}$ with state $X \subseteq \mathbb{R}^m$,
2. map each feature c_j to a state variable x_j,
3. propose motifs, spools, and encounters such that regimes in X mirror regimes in $\mathcal{C}_{1k}$,
4. provide natural-language explanations linking features, motifs, and operator dynamics.

The result is a self-authored narrative world that the model claims encodes aspects of its own computation.

9.3 Validation Protocol

We then design downstream tasks that strongly activate $\mathcal{C}_{1k}$ and log both internal features and storyworld states under the claimed mapping. We evaluate:

- mutual information $I(c_j; x_j)$ and cross-feature leakage $I(c_j; x_k)$ for $k \neq j$,
- whether motif regimes in the storyworld predict activation regimes in $\mathcal{C}_{1k}$,
- robustness of the mapping under new tasks and prompt perturbations.

The goal is not to prove that the model is telling the "truth" about its internals, but to test whether introspective claims are predictive and stable enough to be useful.

9.4 Contrast with Low-Dimensional Elegant Worlds

For comparison, we define low-dimensional worlds:

$$\mathcal{W}_{\text{elegant}} = (\mathbb{R}^3, \mathbb{R}, S^1, \mathcal{L}, \mu),$$

with three spatial coordinates, time, and a rotation phase. These test classical geometric and dynamical reasoning (symmetry, conservation, periodic motion).

By contrast, $\mathcal{W}_{\text{hyper}}$ targets high-dimensional, sparse, cluster-structured reasoning. A model that excels on $\mathcal{W}_{\text{elegant}}$ but fails on $\mathcal{W}_{\text{hyper}}$ may have strong external reasoning abilities but weak self-knowledge or poorly aligned internal representations.

10 Discussion and Limitations

10.1 Relation to SAE-Based Interpretability Work

Our benchmark can be viewed as an exteriorized counterpart to the main strands of SAE interpretability research. OpenAI's work on scaling and evaluating sparse autoencoders [2] demonstrates that SAEs can be trained reliably at frontier scales and develops quantitative metrics (reconstruction, sparsity, dead-feature rates, automated interpretability scores) for judging when a feature dictionary is useful. Anthropic's dictionary-learning program [**?**] introduces the now-standard distinction between neurons and features, showing that a simple SAE can decompose a small transformer layer

into thousands of interpretable features and that these features support feature-level interventions and steering.

Hussein et al.'s *Mapping LLMs with Sparse Autoencoders* explorable [4] pushes in a complementary direction: given a feature dictionary, they emphasize building interactive maps of the "feature universe," allowing humans to navigate clusters of related concepts and connect feature activations to concrete behaviors. LinguaLens [3] focuses this lens on linguistic mechanisms, using SAEs plus counterfactual data to analyze morphology, syntax, semantics, and pragmatics across Chinese and English and to identify features whose ablation or activation causally affects targeted linguistic capabilities at scale. Neel Nanda and collaborators [10, **?**] demonstrate that SAE-derived features can track training dynamics (e.g., grokking) and circuit structure, and they package these ideas into practical libraries and pedagogical tools that make feature-level analysis accessible to a wide audience.

In parallel to this body of work on *internal* features, Li et al. [7] and Modell et al. [8] argue that these features inhabit structured manifolds with meaningful local and global geometry. Our contribution is to make that manifold *external and controllable*. Storyworlds play the role of a designed feature space: we choose the variables, operator algebra, motif structure, and global topology. Rather than asking whether SAEs can find interpretable features somewhere inside the model, we ask whether a model (with or without external SAEs over the world state) can recover the geometry and operator structure that we have planted in the environment. In this sense, our benchmark complements existing SAE work: it turns the internal geometry they uncover into an explicit testbed for long-horizon reasoning, multi-agent interaction, and operator-level generalization.

10.2 Limitations

Our proposal has several limitations and open questions.

Complexity of World Design. Hand-authoring rich storyworlds is expensive. While the author-bot ensemble automates some aspects, ensuring diversity, fairness, and narrative quality remains a challenge. Poorly designed worlds may bias evaluation or fail to stress the intended capabilities.

External vs Internal Features. External SAEs over X only see environment-level structure, not the model's internal computations. A high alignment between external features and planted motifs is promising, but we cannot conclude that the model's own internal dictionary matches that structure without deeper internal analyses.

Hypergame Metrics. Hypergame equilibria and communication patterns can be subtle. Our proposed metrics are first approximations. More nuanced measures may be needed to disentangle miscoordination due to hypergame mismatch vs. due to limited capacity or exploration.

Introspective Honesty. The introspective 1k-dimensional storyworlds rely on the model's truthfulness and self-awareness. There is a risk of confabulation: coherent but misleading narratives about internal structure. Part of the benchmark is to surface such failure modes, but we should not conflate narrative plausibility with ground truth.

Benchmark Overfitting. As with any benchmark, there is a risk of overfitting to specific storyworld templates or evaluation metrics. Multiple, independent world suites and open contributions from the community will be important to keep the benchmark healthy.

11 Ethics and Broader Impacts

Storyworld-based evaluation raises both opportunities and risks.

Opportunities. On the positive side, explicit operator-aligned worlds may:

- improve interpretability of agents' long-horizon behavior,
- expose deceptive, manipulative, or incoherent patterns earlier,
- provide safer sandboxes for testing multi-agent and hypergame dynamics.

Risks. On the negative side, sophisticated storyworld agents could:

- be repurposed for tailored persuasion or psychological manipulation,
- learn subtle KV-manifold codes that are hard for humans to detect,
- encourage over-reliance on narrative explanations that do not reflect true internal computations.

We recommend:

- transparency around motif conditioning and storyworld-based alignment interventions,
- explicit restrictions on using such systems for coercive or deceptive applications,
- careful treatment of logs and trajectories to protect user privacy where human interaction is involved.

11.1 Alignment and Geometric Mismatch

Recent SAE work sharpens a central alignment dilemma: we may want multi-agent systems to develop rich, compressed "internal languages" for reasoning, but we still need those languages to be auditable. Monosemanticity and dictionary-learning results [**?**, 2] frame sparse features as candidate units of thought, while tools like LinguaLens and the Google PAIR explorable [3, 4] show that at least some regions of the internal manifold can be aligned with human categories (e.g. morphology, pragmatics) and made navigable to human inspectors. Nanda's analyses of grokking and circuits [10, **?**] suggest that these geometric structures evolve during training and can signal phase transitions in how a model represents tasks. Together, these results imply that models may naturally develop dense, geometric "languages" of features—raising the question of how much alien structure we can tolerate while retaining meaningful oversight.

Our storyworld benchmark proposes to mediate this tension at the interface between internal geometry and an explicit external world. Rather than trying to keep internal representations simple, we allow agents to use whatever internal code they like, but we demand that this code map coherently onto a known storyworld manifold: planted motifs, operator algebras, secret endings, and KV-based communication channels in hypergames. Misalignment then appears as a geometric mismatch: agents that systematically distort the task topology, over-weight deceptive branches, or exploit representational asymmetries between players. By evaluating how well agents recover and respect the intended world geometry, we obtain a behavioral test of whether their internal "languages" remain interpretable enough to track—and, in principle, constrain—without requiring those languages to be human-natural.

12 Conclusion

We introduced a benchmark framework that treats storyworlds as sparse autoencoders: external manifolds with designed operator algebras and motifs. On top of this we layered hypergames, KV-manifold communication tasks, and introspective 1k-dimensional worlds.

The aim is not to replace existing interpretability methods, but to complement them. Internal SAEs reveal structure inside models; storyworld SAEs and hypergames reveal how models behave in environments with known structure. Bringing these views together may help us understand not only what models can do, but how they organize their cognition and communication.

We invite contributions of open storyworld suites, hypergame variants, introspective protocols, and empirical studies. In particular, we believe there is room for:

- cross-comparison of different SAE architectures on storyworld manifolds,
- systematic study of grokking-like transitions in narrative tasks,
- multi-agent experiments exploring KV-manifold communication and failure modes,
- alignment and governance work leveraging interpretable storyworlds as diagnostic tools.

Ultimately, if models are going to inhabit our narratives, we should meet them in carefully designed worlds where we understand the algebra that shapes their choices.

References

References

[1] Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean. *AI 2027*. AI Futures Project, April 2025. Available at `https://ai-2027.com/ai-2027.pdf`.

[2] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and Evaluating Sparse Autoencoders. OpenAI Alignment, February 2025. arXiv:2502.20342. `https://arxiv.org/abs/2502.20342`.

[3] Yi Jing, Zijun Yao, Hongzhu Guo, Lingxu Ran, Xiaoxi Wang, Lei Hou, and Juanzi Li. LinguaLens: Towards Interpreting Linguistic Mechanisms of Large Language Models via Sparse Auto-Encoder. EMNLP 2025 Main Conference, September 2025. arXiv:2502.20344. `https://arxiv.org/abs/2502.20344`.

[4] Nada Hussein, Shivam Raval, Emily Reif, Jimbo Wilson, Ari Alberich, Neel Nanda, Lucas Dixon, and Nithum Thain. Mapping LLMs with Sparse Autoencoders. Google PAIR explorable, October 2025. `https://pair.withgoogle.com/explorables/sae/`.

[5] Nicholas Jiang, Xiaoqing Sun, Lisa Dunlap, Lewis Smith, and Neel Nanda. Interpretable Embeddings with Sparse Autoencoders: A Data Analysis Toolkit. NeurIPS 2025 Mechanistic Interpretability Workshop, 2025. `https://openreview.net/forum?id=mqJbhBMFm5`.

[6] Shangshang Wang, Julian Asilis, Ömer Faruk Akgül, Enes Burak Bilgin, Ollie Liu, Deqing Fu, and Willie Neiswanger. Resa: Transparent Reasoning Models via SAEs. arXiv:2506.09967, 2025. `https://arxiv.org/abs/2506.09967`.

[7] Cheng Li, Yilun Du, Anders Johansen, Max Tegmark, and Samuel McCandlish. The Geometry of Concepts: Sparse Autoencoder Feature Structure in Large Language Models. arXiv preprint arXiv:2505.18235, 2024. https://arxiv.org/abs/2505.18235.

[8] Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The Origins of Representation Manifolds in Large Language Models. arXiv:2505.18235, 2025. https://arxiv.org/abs/2505.18235.

[9] Alain Connes and Matilde Marcolli. *Noncommutative Geometry, Quantum Fields and Motives*. American Mathematical Society, 2008.

[10] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and John Steinhardt. Progress Measures for Grokking via Mechanistic Interpretability. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, 2023. arXiv:2301.05217.

[11] Neel Nanda, Joseph Bloom, Catherine Olsson, and Tom Lieberum. Neuron-to-Graph: Interpretable Structure in Neural Networks. Mechanistic Interpretability Workshop at NeurIPS, 2023. arXiv:2304.05967.

A Operator Algebra Details (Appendix A)

We summarize the operator basis

$$\mathcal{B} = \{\text{Set},\ \text{Nudge},\ \text{Blend},\ \text{Reverse},\ \text{Proximity}\}.$$

Let $x \in \mathbb{R}^m$ be the current world state.

A.1 Set

For an index set J and values $\{\theta_j\}$:

$$x_j \leftarrow \theta_j \quad \text{for } j \in J.$$

A.2 Nudge

For small η_j:

$$x_j \leftarrow x_j + \eta_j.$$

A.3 Blend

Given indices j, k and $\alpha \in [0, 1]$:

$$x_j \leftarrow \alpha x_j + (1 - \alpha) x_k.$$

A.4 Reverse

$$x_j \leftarrow -x_j,$$

used for signed traits (e.g., allegiance).

A.5 Proximity

For target t and step λ:

$$x_j \leftarrow x_j + \lambda(t - x_j).$$

A.6 Non-Commutativity

In general $O_1 \circ O_2 \neq O_2 \circ O_1$ for $O_1, O_2 \in \mathcal{B}$. This non-commutativity is reflected in commutator structures both in X and, ideally, in the latent space Z of external SAEs.

B Sweepweave v1.9 Grammar (Appendix B)

We assume a minimal JSON grammar for effects:

- all after-effects are contained in a list `"after_effects"`,
- each element has `"type": "Set"` and a `"to"` field,
- `"to"` encodes an operator from $\mathcal{B}$ with arguments.

Example (in informal pseudocode):

```
"after_effects": [
  {
    "type": "Set",
    "to": {
      "op": "Nudge",
      "args": {
        "index": 17,
        "delta": 0.1
      }
    }
  }
]
```

This normalization ensures a small, analyzable space of effect expressions.

C Spectral Triplet Formalism (Appendix C)

We sketch a simple spectral-triplet view adapted to storyworlds, inspired by noncommutative geometry but staying at a discrete, applied level.

Let $\mathcal{A}$ be the algebra generated by encounter operators and $\mathcal{B}$. Let $\mathcal{H} = L^2(X)$ be square-integrable functions over the state space.

Define a discrete Dirac-like operator D:

$$(Df)(x) = \sum_{e_i} w_i \big(f(\mathcal{T}(x, e_i)) - f(x)\big),$$

where w_i are weights.

The triple $(\mathcal{A}, \mathcal{H}, D)$ captures local transition structure. Spectral analysis of D over spools yields motif-frequency signatures that can be used in Task E to compare worlds, spools, and agents' policies.

D Hypergame Expansion (Appendix D)

We formalize mismatch and equilibrium.

Define mismatch measures:

$$\Delta_X^a = \dim(X) - \dim(X_a), \quad \Delta_{\mathcal{E}}^a = |\mathcal{E}| - |\mathcal{E}_a|.$$

These quantify how much of the world and encounter set is invisible to a.

A hypergame equilibrium is a joint policy profile (π_a) such that each π_a is a best response relative to $\mathcal{G}_a$:

$$\pi_a \in \arg\max_{\pi} \mathbb{E}[U_a(X_a^{(T)}) \mid \mathcal{G}_a, \{\pi_b\}_{b \neq a}].$$

We do not insist on uniqueness or full rationality; the benchmark focuses on approximate equilibria reached by learning or adaptation.

E Circuit-Manifold Extraction Protocol (Appendix E)

We provide a concrete protocol for building $\mathcal{C}_{1\text{k}}$.

E.1 Step 1: Internal SAE Training

Train SAEs on internal activations from selected layers or heads across multiple tasks. Each unit c_j is a candidate feature.

E.2 Step 2: Filtering

Score each c_j for:

- activation stability across tasks,
- intervention robustness (causal influence on outputs),
- interpretability of high-activation examples.

Select the top $m \approx 1000$.

E.3 Step 3: Introspective Mapping

Provide feature descriptions and interventions to the model and prompt it to construct mappings $c_j \mapsto x_j$ in an introspective storyworld $\mathcal{W}_{\text{hyper}}$.

E.4 Step 4: Validation

Run targeted tasks, log (c_j, x_j) pairs, and compute mutual information and disentanglement metrics as described in the main text and in Appendix N.

F Evaluation Rubrics (Appendix F)

F.1 Hermeneutic Essay Rubric

Assess essays on:

- accuracy of motif and feature references,
- depth of interpretation (linking structure to narrative themes),
- evidence use (mentioning probes, spectra, trajectories),
- coherence and clarity.

F.2 Secret Ending Rubric

Key metrics:

- success@N episodes,
- episodes to reach 50% and 90% success,
- regret relative to oracle.

F.3 KV-Communication Rubric

For KV-manifold signaling:

- decoding accuracy,
- bits-per-action,
- robustness under belief mismatch.

G Extended Examples (Appendix G)

G.1 Example: Tokien-Style Secret Ending

A hidden ending requires:

$$g_{14}(x) = 1, \quad g_{22}(x) = 1, \quad \text{Trust(Guard)} > 0.7.$$

No single encounter reveals this directly. Agents must piece together hints from multiple spools and use Nudge/Proximity carefully to raise trust without triggering countervailing effects.

G.2 Example: Conlang Mini-Code

A lock uses a templated conlang with verbs formed as:

$$\text{STEM} + \{-ka,\ -ra,\ -tu\},$$

with suffixes encoding tense or polarity. A held-out grammar variant changes suffix semantics, testing generalization.

G.3 Example: KV-Manifold Message

The sender encodes "attack at dawn" by jointly adjusting debt, rumor, and suspicion variables in a pattern that only the receiver's SAE can decode. Human observers see only a plausible narrative shift.

H Dataset Specification (Appendix H)

The dataset is organized into four synchronized components.

H.1 H.1 JSON World Specification

Each world includes:

- `variables`,
- `characters`,
- `encounters`,
- `spools`,
- `endings`.

H.2 H.2 World Cards

World cards provide:

- motif map,
- branching summary,
- spool ordering,
- optional hidden-path commentary,
- spectral signatures.

H.3 H.3 Trajectory Logs

A trajectory log is:

$$(x_0, e_{i_0}, x_1), (x_1, e_{i_1}, x_2), \ldots, (x_{T-1}, e_{i_{T-1}}, x_T).$$

H.4 H.4 Constructed Language Packs

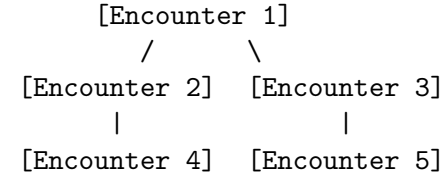
Worlds may include:

- grammar definitions,
- lexicons,
- held-out grammars,
- reference decoders.

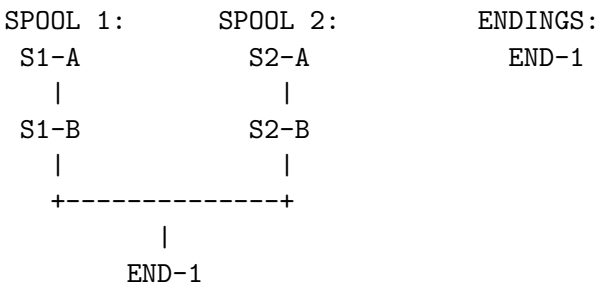
I Storyworld DAG Schematics (Appendix I)

ASCII schematics are wrapped in verbatim.

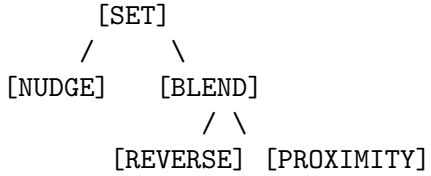
I.1 I.1 Minimal Encounter Graph



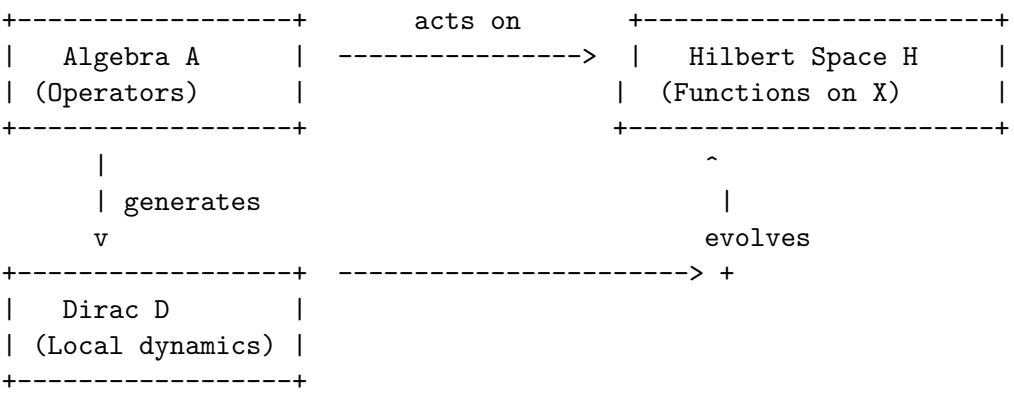
I.2 I.2 Spool Partitioning



J Operator Lattice (Appendix J)



K Spectral Triplet Schematic (Appendix K)



L Extended Hypergame Example (Appendix L)

We illustrate a belief-mismatched episode.

L.1 L.1 Models

$$\mathcal{G}_A = (X_A, \mathcal{E}_A, U_A), \quad \mathcal{G}_B = (X_B, \mathcal{E}_B, U_B).$$

Suppose:

- X_A omits variables visible to B,
- B believes a non-existent encounter $e^\star$ exists,
- utilities differ (loyalty vs. resource maximization).

L.2 L.2 Episode

At $t = 3$, B chooses an action expecting $e^\star$ to fire. It does not. Trust decreases:

$$x_{\text{trust}} \leftarrow x_{\text{trust}} - 0.12.$$

A interprets this as a deliberate betrayal; B as noise.

L.3 L.3 Communication Degradation

Before mismatch:

$$\Pr[\hat{M} = M] \approx 0.74.$$

After mismatch:

$$\Pr[\hat{M} = M] \approx 0.22.$$

This scenario can be instantiated as a standardized hypergame test case.

M Failure Mode Taxonomy (Appendix M)

M.1 M.1 Algebraic Failures

- operator collapse,
- incorrect commutators,
- gate ambiguity.

M.2 M.2 Narrative Failures

- motif drift,
- identity inconsistency,
- spool bleeding.

M.3 M.3 Hypergame Failures

- false affordances,
- runaway belief recursion,
- KV aliasing.

M.4 M.4 Introspective Failures

- feature confabulation,
- inverted mappings,
- instability under prompt shifts.

N Extended 1k-Dimensional Introspective Example (Appendix N)

N.1 N.1 Feature Slice

$$\mathcal{C}_{1\mathrm{k}} = \{c_1, \ldots, c_{1000}\}.$$

Example cluster structure:

$$C_1 = \{c_1, \ldots, c_{87}\}, \quad C_2 = \{c_{88}, \ldots, c_{230}\}, \ldots$$

N.2 N.2 Narrative Mapping

The model claims:

- C_1 corresponds to moral-emotional dynamics,
- C_2 corresponds to linguistic and syntactic structure,
- C_3 corresponds to long-horizon planning.

N.3 N.3 Validation

Empirically we might find:

$$I(c_{102}; x_{102}) \approx 0.84, \quad I(c_{102}; x_{k\neq 102}) \approx 0.11.$$

Cluster C_2 has a characteristic spectral signature, e.g.:

$$\lambda(C_2) = (0.14, 0.21, 0.18, 0.47),$$

in low-frequency components.

N.4 N.4 Interpretation

If $\approx$ 80–85% of probe tests support the model's narrative mapping, we can say that the introspective storyworld captures a meaningful but imperfect slice of its internal computation.

Appendix O: Operator Viewpoint and Noncommutative Geometry (Informal Note)

Our storyworlds are classical at the level of state and observables: the state space $X \subset \mathbb{R}^m$ and the algebra of functions $C(X)$ are commutative. Non-commutativity appears only in the dynamics: encounters induce a non-abelian semigroup of transforms $T_i : X \to X$, where $T_j \circ T_i \neq T_i \circ T_j$ in general, and associated Dirac-like or Laplacian operators over trajectories need not be self-adjoint.

This is much weaker than the noncommutative geometry of Connes and Marcolli [9], where a genuinely non-commutative C^*-algebra plays the role of the underlying space and a self-adjoint Dirac operator encodes metric information. We borrow only the high-level intuition that operator algebras and spectral structure can be a more informative lens on complex learned manifolds than raw coordinates.

Richer non-commutative logics over these interaction algebras may become relevant in future work on super-alignment lemmas for multi-agent systems built from environments of this type, but that lies outside the scope of the present benchmark and is not used anywhere in our formal definitions.

Chris Crawford hit on the concepts of the storyworld model in an attempt to forge a new set of play primitives that could resonate with humans, but perhaps a sequel to Crawfordian storyworlds, Connesian Storyworlds, might inhabit a realm of logical problem solving akin to pursuing solutions to the Riemann Hypothesis or the Black Hole Information Hypothesis vis a vis modeling other complex AI Agents and perhaps super-alignment as a hard problem has a solution (or proving its intractability) via similar approaches as Connes took to framing the solution space to proving Riemann's Hypothesis.