Manifold Steering Reveals the Shared Geometry of Neural Network Representation and Behavior

Anonymous Author(s)
Affiliation
Address
email

Neural representations carry rich geometric structure; but does that structure causally shape behavior? To address this question, we intervene along paths through activation space defined by different geometries, and measure the behavioral trajectories they induce. In particular, we test whether interventions that respect the geometry of activation space will yield behaviors close to those the model exhibits naturally. Concretely, we first fit an activation manifold $\mathcal{M}_h$ to representations and a behavior manifold $\mathcal{M}_y$ to output probability distributions. We then test the link $\mathcal{M}_h \leftrightarrow \mathcal{M}_y$ via interventions: we find that steering along $\mathcal{M}_h$, which we term *manifold steering*, yields behavioral trajectories that follow $\mathcal{M}_y$, while linear steering—which assumes a Euclidean geometry—cuts through off-manifold regions and hence produces unnatural outputs. Moreover, optimizing interventions in activation space to produce paths along $\mathcal{M}_y$ recovers activation trajectories that trace the curvature of $\mathcal{M}_h$. We demonstrate this bidirectional relationship between the geometry of representation and behavior across tasks and modalities. In language models, we use reasoning tasks with cyclic and sequential geometries as well as in-context learning tasks with more complex graph geometries. In a video world model, we use a task with geometry corresponding to physical dynamics. Overall, our work shows that geometry in neural representation is not merely incidental, but is in fact the proper object for enabling principled control via intervention on internals. This recasts the core problem of steering from finding the right *direction* to finding the right *geometry*.

1 Introduction

A plethora of geometric structures have been documented in neural network representations (Modell et al., 2025b; Park et al., 2025b; Kozlowski et al., 2025; Shai et al., 2024b; Pearce et al., 2025; Gurnee et al., 2026). Recent literature has begun to identify the origins of these structures by attributing them back to data statistics shaped by conceptual structure (Karkada et al., 2026; Prieto et al., 2026; Park et al., 2025b; Merullo et al., 2025). However, we have barely begun to understand what causal role these geometric structures play in a model's computation (cf. Engels et al. 2024; Kantamneni & Tegmark 2025; Csordás et al. 2024; Sarfati et al. 2026). We address this question by intervening on model activations under different geometric assumptions and measuring the effect on behavior.

Currently, it is common for activation-based intervention methods to assume a Euclidean geometry for activation space, where steering is performed by adding a *steering vector* to model activations with a scalar that modulates intervention strength (Bau et al., 2019; Subramani et al., 2022; Marks & Tegmark, 2024; Panickssery et al., 2024; Turner et al., 2024; Li et al., 2023; Rimsky et al., 2024; Chen et al., 2025). This approach is motivated by the *linear representation hypothesis* (LRH), which posits that neural activations can be decomposed into atomic concepts encoded along single (approximately) orthogonal directions (Smolensky, 1986; Park et al., 2023; Elhage et al., 2022b). However, linear steering often produces degraded fluency, diversity collapse, and unstable off-target behavior (Wu et al., 2025; Da Silva et al., 2025; Bigelow et al., 2025; Tan et al., 2024; Hao et al., 2025; Bhalla et al., 2024; Pres et al., 2024), which suggests the assumed Euclidean geometry is inappropriate.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

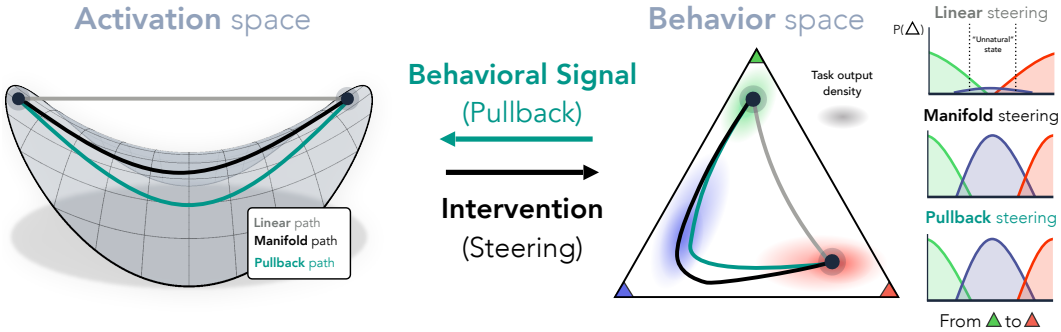

Figure 1: **How do different geometries of activation space modulate behavior?** We illustrate paths through activation space (left), each defined by a different geometry. Interventions along paths in activation space induce paths in behavior space (right, illustrated on a three-concept probability simplex). **Euclidean:** the standard approach of linear steering assumes a flat geometry and interventions follow a straight line. Such paths may cut across the activation manifold, yielding *unnatural* behavioral trajectories that pass through off-manifold regions of behavior space. **Density geometry:** a density-based metric whose geodesics follow the intrinsic geometry of a fitted activation manifold, yielding more natural transitions in behavior space. **Pullback geometry:** a behavior-aware metric obtained by "pulling back" behavior-space geometry into activation space, yielding paths that follow the manifold of natural (unintervened) output distributions. Overall, we argue that geometric structure in neural representations encodes the conceptual space a model is reasoning over, which in turn constrains its output behavior. Hence, manifolds in activation and behavior space are two images of the same underlying structure, and so we expect the density and pullback geometries to coincide.

In this work, we advance the hypothesis that *representation geometry* provides a blueprint for effective steering that will overcome the limitations of the linear approach. Steering is fundamentally about how internal representations control behavior, so to test this hypothesis we must study not only paths through activation space, but also the behavioral trajectories induced by interventions along these paths. Successful steering will produce trajectories that are in line with the model's *natural* (unintervened) output distribution. If we are right, then interventions that respect the geometry of internal representations and interventions that respect the geometry of behavior will be one and the same. Motivated by this, we make the following contributions in this work.

- **Uncovering isometric geometries in neural network representation and behavior.** We use tasks where models output a distribution over a set of concepts with known structure. In each task, we fit an *activation manifold* $\mathcal{M}_h$ to internal representations and a *behavior manifold* $\mathcal{M}_y$ to model outputs (probability distributions over task-relevant concepts). We show the two geometries are tightly interlinked via a scaled isometry relation: geodesic distances on $\mathcal{M}_h$ align closely with those on $\mathcal{M}_y$, and neither match Euclidean distances.
- **Validating the causal role of representation geometry.** We perform geometry-aware steering experiments and compare against the baseline of linear steering (see Fig. 1). We show linear steering cuts through low-density regions of behavior space and passes through unnatural intermediate distributions; meanwhile, steering along the activation manifold $\mathcal{M}_h$ yields behavioral trajectories that follow $\mathcal{M}_y$ closely. In fact, optimizing for paths along $\mathcal{M}_y$ recovers activation trajectories that trace the curvature of $\mathcal{M}_h$, further tightening the link between activation geometry and behavior.
- **A theoretical framework for geometry-aware steering.** Building on the results above, we formulate steering as a problem of choosing the right *geometry* for activation space, rather than the right *direction*. In particular, we argue steering can be defined as the problem of finding a geodesic connecting two points under different activation-space metrics: linear steering assumes a flat metric (Euclidean geometry), steering along the activation manifold uses a metric derived from natural activations, and steering optimized to follow the behavior manifold uses a metric derived from natural behaviors.

We demonstrate these findings hold across modalities and tasks. In large language models, we test geometries from cyclic concepts (weekdays, months; Engels et al. 2024; Modell et al. 2025b), sequential concepts (ages, letters), and multi-dimensional graph structures learned in context (Park et al., 2025b). In a video world model, we test a geometry of physical position in a simulated

environment (mountain car; Moore 1990; Towers et al. 2024). Together, these findings provide evidence for the posited account, and support steering along neural manifolds as the principled form of activation-based intervention.

2 The Geometry of Representation and Behavior

2.1 Setup

Running example. We will explicate our framework and empirical methods using a running example where a language model is required to reason about the days of the week (Engels et al., 2024). Specifically, we consider prompts of the form: `What day is` k `days after` z`?` with $z \in \mathcal{Z} = \{\text{Mon}, \text{Tue}, \ldots, \text{Sun}\}$ and $k \in \{1, \ldots, 7\}$. Given such a prompt, the LM outputs a probability distribution over all possible tokens.

Concept geometry. We draw inspiration from work on *conceptual spaces* in cognitive science, where conceptual domains, e.g., days of the week, are geometrically enriched with a metric such that distances between points encode similarity and guide patterns of inference (Shepard 1987; Gärdenfors 2000; Tenenbaum & Griffiths 2001; Bellmund et al. 2018; see Fel et al. 2025b; Lubana et al. 2025; Yocum et al. 2025; Modell et al. 2025b for related work in interpretability). For example, the days of the week $\mathcal{Z}$ may be organized in a cyclic structure that is captured by a *metric* $d_{\mathcal{Z}}$ measuring temporal distance between days, e.g., neighboring days are closer together. Indeed, when humans mistakenly report the current day, they most often confuse it with its neighboring days (Ellis et al., 2015).

Karkada et al. (2026) and Prieto et al. (2026) show that similarity structure between days of the week is reflected in the statistics of training data, which in turn shape the geometry of internal representations (Engels et al., 2024; Park et al., 2025a; Modell et al., 2025a; Prieto et al., 2026). We hypothesize that a model's output distributions over $\mathcal{Z}$ are similarly shaped by $d_{\mathcal{Z}}$: e.g., when asked `What day is four days after Monday?`, the model concentrates mass on Friday and spreads the remainder onto nearby days like Thursday and Saturday.

Notation. We work with two spaces: the activation space $\mathcal{A} = \mathbb{R}^n$, and the behavior space $\mathcal{Y} = \Delta^{|\mathcal{Z}|}$, which is the open probability simplex[1] over the conceptual domain $\mathcal{Z}$, with an additional 'other' class for off-concept probability mass. For an input x, let $\boldsymbol{p}(x) \in \mathcal{Y}$ denote the model's output distribution over $\mathcal{Z}$, given by restricting the full vocabulary distribution of the model to the tokens in $\mathcal{Z}$, in addition to the 'other' class for remaining probability mass. Let $\boldsymbol{h}(x) \in \mathcal{A}$ denote an activation vector of interest for input x.

For a class of input queries that share the same answer, e.g., `What is two days after Monday?` and `What is three days after Sunday?`, we average the hidden activations and output distributions to produce "activation centroids" and "behavior centroids", respectively.

Experimental tasks. We perform language model experiments on four tasks, two with cyclic conceptual structure and two with sequential conceptual structure. The cyclic tasks require reasoning about days of the week and months of the year, e.g., `What is four months after January?`. The sequential tasks require reasoning about letters and ages, e.g., `What is four letters after m?` or `Alice is 7, Bob is 5 years older. How old is Bob?`.

2.2 Fitting the Manifolds

We fit a smooth manifold within each space to the model's unintervened activations or outputs for a task: $\mathcal{M}_h \subseteq \mathcal{A}$, the *activation manifold*, and $\mathcal{M}_y \subseteq \mathcal{Y}$, the *behavior manifold*. To fit the activation manifold $\mathcal{M}_h$, we reduce activation vectors $\boldsymbol{h}(x)$ to 64 dimensions via PCA, compute "concept centroids" (e.g., averaging all activations where the correct answer is Wednesday), and fit cubic splines (Reinsch, 1967) through the centroids (see App. F.3 for further spline fitting details). To fit the behavior manifold $\mathcal{M}_y$, we follow a similar procedure but first map each centroid from the probability simplex onto Hellinger space via $p \mapsto \sqrt{p}$. This linearizes the geometry of the simplex: the Hellinger

[1] We require the open simplex $\{\boldsymbol{p} \in \mathbb{R}^{|\mathcal{Z}|}_{>0} : \sum_i p_i = 1\}$, i.e., strictly positive entries. The closed simplex, which includes faces where some $p_i = 0$, has boundary and corners and is not a smooth manifold.

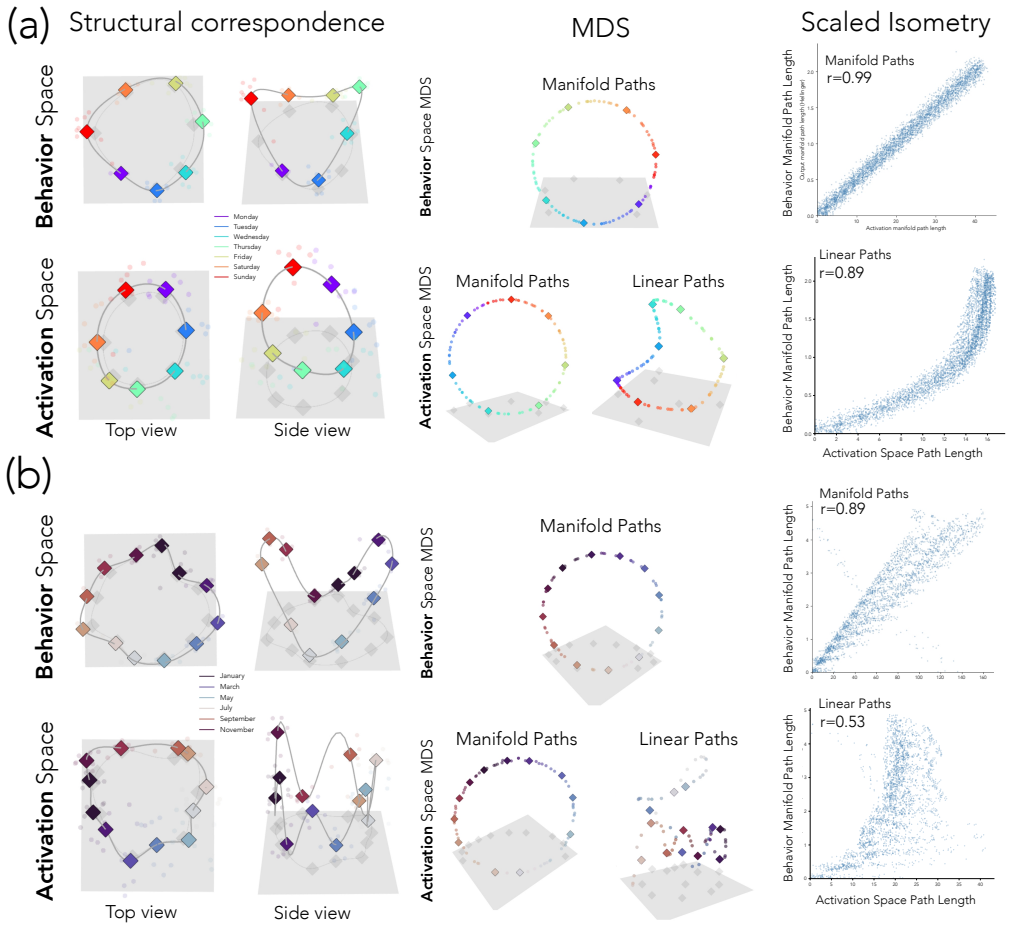

Figure 2: **Approximate isometry between activation and behavior manifolds for cyclic concepts**. Manifolds (cubic splines) fit to activation and behavior (i.e., output distributions over concept tokens) spaces of Llama 3.1 8B. The weekdays **(a)** and months **(b)** tasks consist of simple addition questions such as: `What is four days after Monday?`. Both activation and behavior manifolds show cyclic structure (PCA visualization shown in left column). Furthermore, on-manifold distances in activation space show strong correlation with on-manifold distances in behavior space (right column), as well as a clear structural match via a multidimensional scaling (MDS) embedding (middle column). In contrast, linear distances in activation space show weaker correlations and warped structures. These results demonstrate an approximate isometry between the activation and behavior space manifolds. See more domains in Figure 8.

distance between distributions becomes an ordinary Euclidean distance, $d_H(p,q) = \frac{1}{\sqrt{2}}\|\sqrt{p} - \sqrt{q}\|$, so we can fit splines and compare distributions with standard Euclidean tools while still respecting the underlying probabilistic geometry (Amari & Nagaoka, 2000). Decoded points are squared back to recover valid distributions (further details, including how we keep the fit on the sphere, are in App. F.4). Unless stated otherwise, we use Llama 3.1 8B (Touvron et al., 2023) with activations from layer 28, and visualize manifolds via 3D PCA.

2.3 Conceptual Structure Appears in Behavior and Activation Space

We will now begin our investigation into the connection between the activation and behavior manifolds. Before we perform any interventions on internal representations, we examine the structural correspondence between these spaces, and measure whether distances along the two manifolds are proportional (a scaled isometry).

We find that both the activation and behavior manifolds recapitulate conceptual structure (see Figs. 2, 8 for visualizations). For example, in the case of days of the week, the activations and output distri-

butions are arranged in order around a loop, with Monday adjacent to Tuesday and Sunday, and Thursday on the opposite side (Fig. 2). The circle representing days of the week in activation space is already known to exist (Engels et al., 2024; Modell et al., 2025b; Karkada et al., 2026). However, the circle in behavior space is a novel discovery, and results from sharply-peaked output distributions placing most mass on the target concept, with the remainder concentrated on its neighbors. The correspondence is striking; output distributions and internal activations recover the same cyclic ordering. In contrast, the conceptual structure for the ages and letters task is sequential rather than cyclic, and so both the activations and output distributions for these tasks lie on an open curve (Fig. 8).

Going beyond qualitative structural correspondence, we wish to examine the mapping between distances along each manifold. We test this by computing pairwise distances between points in both spaces: geodesic distances $d_{\mathcal{M}_h}(m_i, m_j)$ on the activation manifold, and geodesic distances $d_{\mathcal{M}_y}(p_i, p_j)$ on the behavior manifold. We compute geodesic distance on $\mathcal{M}_h$ using cumulative Euclidean distance between points along a geodesic path, and follow the same procedure for geodesic distance on $\mathcal{M}_y$, but using cumulative Hellinger distance (see App. F.5 for further details). The two distances are highly correlated ($r = 0.99$ weekdays, $r = 0.89$ months, $r = .999$ letters, $r = .999$ ages) indicating that $\mathcal{M}_h$ and $\mathcal{M}_y$ are approximately isometric. Meanwhile, linear paths between the same activation-space points correlate less well with $\mathcal{M}_y$ geodesics, with the relationship showing clear non-linear patterns ($r = 0.89$ weekdays, $r = 0.53$ months, $r = 0.71$ letters, $r = 0.36$ ages).

This correspondence leads to an intuitive hypothesis. $\mathcal{M}_y$ was fit to unintervened task behavior, so it traces a path through *natural* output distributions for the model. If the $\mathcal{M}_h \leftrightarrow \mathcal{M}_y$ mapping holds, paths along one manifold should track paths along the other. Interventions in activation space that follow $\mathcal{M}_h$ should produce natural trajectories along $\mathcal{M}_y$. Conversely, activation-space paths that are optimized to produce trajectories on $\mathcal{M}_y$ should recover $\mathcal{M}_h$. Next, we test both directions via intervention: representation to behavior ($\mathcal{M}_h \rightarrow \mathcal{M}_y$) and behavior to representation ($\mathcal{M}_h \leftarrow \mathcal{M}_y$).

3 Connecting Representation and Behavior via Intervention

Now that we have established a correlational correspondence between the activation and behavior manifolds across four tasks, we turn to steering interventions for causal evidence. First, we steer along $\mathcal{M}_h$ and measure whether output trajectories follow $\mathcal{M}_y$ (§3.2). Second, we optimize interventions on internal representations to produce output distributions that follow $\mathcal{M}_y$ and measure whether the optimized activation trajectory follows $\mathcal{M}_h$ (§3.3).

3.1 Steering Intervention Notation

The basic intervention operation entails replacing the model's activation at a chosen layer with a target activation, and continuing the forward pass. Given a base input x and a target $\boldsymbol{h}^\star \in \mathcal{A}$, we write $\boldsymbol{p}_{\boldsymbol{h} \leftarrow \boldsymbol{h}^\star}(x)$ for the resulting output distribution. A *steering path* is a curve $\boldsymbol{\pi} : [0, 1] \rightarrow \mathcal{A}$ between endpoints $\boldsymbol{h}_0^\star$ and $\boldsymbol{h}_1^\star$, inducing a trajectory $\boldsymbol{p}_{\boldsymbol{h} \leftarrow \boldsymbol{\pi}(t)}(x)$ through behavior space $\mathcal{Y}$. The behavioral trajectory will be non-stationary only if the target $\mathbf{h}$ mediates the causal effect from input to output (Pearl, 2001; Vig et al., 2020; Mueller et al., 2024).

We consider two strategies, both constructed by interpolation between the endpoints; the strategies differ only in the coordinate system in which the interpolation is taken (Fig. 1):

$$\boldsymbol{\pi}_{\text{lin}}(t) = (1-t)\,\boldsymbol{h}_0^\star + t\,\boldsymbol{h}_1^\star \qquad \text{(linear steering);} \tag{1}$$

$$\boldsymbol{\pi}_{\text{m}}(t) = \boldsymbol{s}\big((1-t)\,\boldsymbol{u}_0 + t\,\boldsymbol{u}_1\big), \quad \boldsymbol{u}_i = \boldsymbol{s}^{-1}(\boldsymbol{h}_i^\star) \qquad \text{(manifold steering).} \tag{2}$$

In the above, $\boldsymbol{s} : \mathbb{R}^k \rightarrow \mathcal{A}$ is a *parameterization* of $\mathcal{M}_h$—the map sending k-dimensional intrinsic coordinates to the corresponding point on the manifold in the activation space $\mathcal{A}$. Linear steering (also known as 'diff-in-means steering') (Bau et al., 2018; Subramani et al., 2022; Turner et al., 2023) interpolates in $\mathcal{A}$ directly—the standard additive-vector baseline. Manifold steering interpolates in the intrinsic coordinates of $\mathcal{M}_h$ and maps the result back through $\boldsymbol{s}$, so $\boldsymbol{\pi}_{\text{m}}$ stays on the activation manifold $\mathcal{M}_h$ throughout. Each strategy thus corresponds to a different choice of geometry on activation space, which we concretize in §3.4.

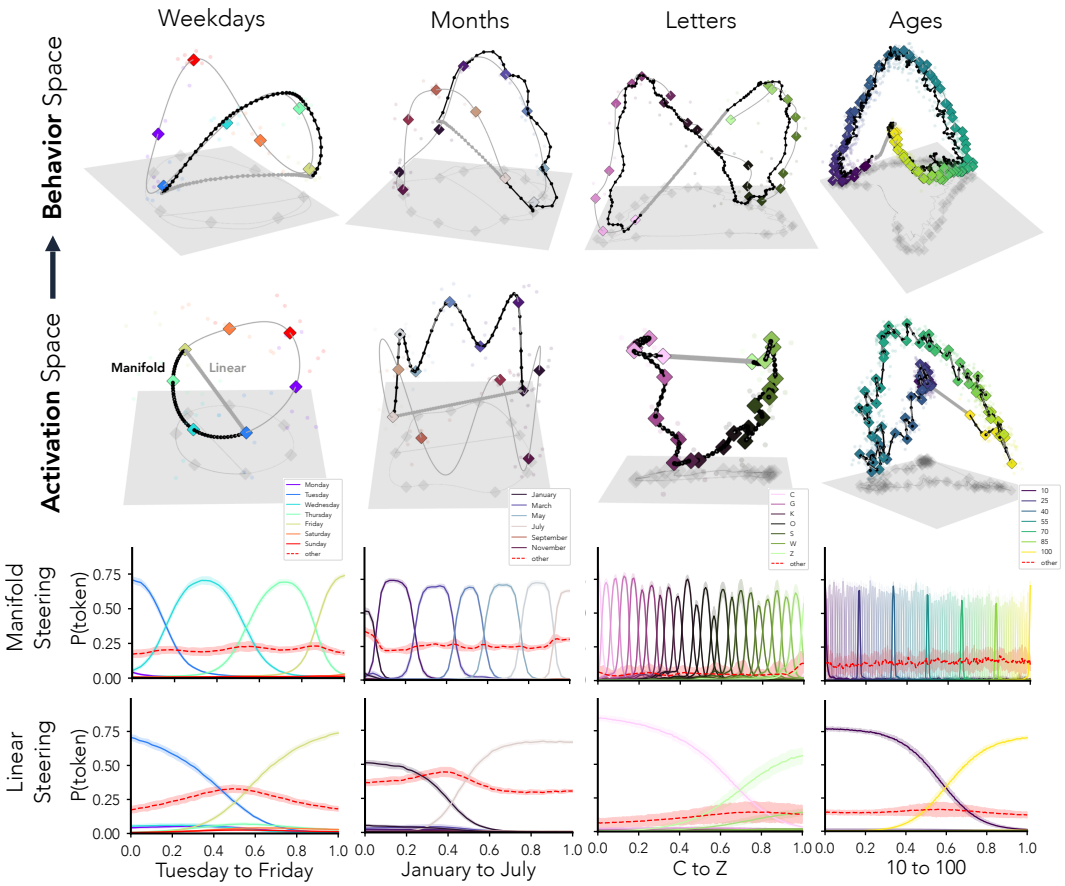

Figure 3: **Manifold steering yields smooth and ordered behavioral transitions.** Using simple addition tasks which require reasoning over structured concepts (e.g., `What is four days after Monday?`), we compare two steering strategies in activation space: standard linear steering, which takes direct paths, and *manifold steering*, which takes paths along a fitted activation manifold. The bottom panel shows example output paths given by each method. Across four settings, manifold steering produces smooth and ordered output transitions between adjacent concepts. In contrast, linear steering leads to 'teleportation' of probability between non-adjacent concepts, and at times results in probability on non-related tokens surpassing any individual concept near the path midpoint. The top panel shows output trajectories in behavior space resulting from manifold steering. We find that steering along the activation-space manifold yields paths that follow the behavior manifold, while linear steering traces paths far from the manifold. Thus, the outputs produced under manifold steering more closely resemble *natural* outputs produced without intervention.

3.2 Steering Along the Activation Manifold Follows the Behavior Manifold

For every pair of start and end values, e.g., from Tuesday to Friday, we steer from the start centroid to the end centroid in activation space using manifold and linear steering with $K = 50$ intervention points along each path. We report the average trajectory in behavior space over a set of 16 prompts sampled randomly from the task's input distribution (Fig. 3, see App. F.6 for further experimental details). We find that manifold steering produces *smooth* and *ordered* behavioral transitions: probability mass shifts steadily through adjacent values of a concept—from Monday to Tuesday to Wednesday to Thursday—while linear steering instead exhibits *'teleportation'*: mass jumps between non-adjacent concepts as the straight line cuts through the manifold's interior.

This qualitative evidence is encouraging, but we have yet to examine our key hypothesis that interventions along the activation manifold $\mathcal{M}_h$ produce *natural* output trajectories that follow $\mathcal{M}_y$. This would mean that outputs produced under manifold steering resemble those produced without

intervention. We quantify this via an "energy function", as described next, under which a natural trajectory is one of low cumulative energy as defined by $\mathcal{M}_y$.

An Energy-based View of Naturalness. Energy functions have a long history in machine learning as a way to measure plausibility under a model (Hopfield, 1982; LeCun et al., 2006). These functions assign low values for likely states and high values for unlikely ones, with the standard correspondence $E(\boldsymbol{x}) \propto -\log p(\boldsymbol{x})$ giving energy the interpretation of an unnormalized log-density (Hopfield, 1982; LeCun et al., 2006; Grathwohl et al., 2019; Song & Kingma, 2021; Béthune et al., 2025). We adopt the same view here. The model's output distributions on unintervened forward passes trace out a low-energy region of behavior space (approximately captured by the manifold $\mathcal{M}_y$) and a steering trajectory is natural to the extent it stays within that region. Concretely, given a steering path $\boldsymbol{\pi} : [0, 1] \to \mathcal{A}$, let $\boldsymbol{\gamma}(t) = \boldsymbol{p}_{\boldsymbol{h} \leftarrow \boldsymbol{\pi}(t)}(\boldsymbol{x})$ be the behavioral trajectory it induces. We define its cumulative output energy:

$$E_{\text{BC}}(\boldsymbol{\gamma}) \;=\; \int_0^1 d_{\text{BC}}\big(\boldsymbol{\gamma}(t),\, \mathcal{M}_y\big)\, dt, \tag{3}$$

where $d_{\text{BC}}(\boldsymbol{p}, \mathcal{M}_y) = \inf_{\sqrt{\boldsymbol{q}} \in \mathcal{M}_y} d_{\text{BC}}(\boldsymbol{p}, \boldsymbol{q}) = -\log(\sum_i \sqrt{\boldsymbol{p}_i}\sqrt{\boldsymbol{q}_i})$ is the Bhattacharyya distance to the nearest point on $\mathcal{M}_y$, a natural choice given that it is simply the negative log of the dot product in Hellinger space (in which $\mathcal{M}_y$ is fit). We note that the formulation above is a tractable proxy we use to estimate distance from the model's natural output distribution, yet it is but one instantiation of a more general framework we develop in §3.4. Applying this measure, we find that manifold steering (weekdays $E_{\text{BC}} = 0.34 \pm 0.03$; months $E_{\text{BC}} = 0.36 \pm 0.01$; letters 2.42 ± 0.07; ages $E_{\text{BC}} = 5.21 \pm 0.09$) produces significantly more natural paths, i.e., lower cumulative energy, than linear steering (weekdays $E_{\text{BC}} = 0.93 \pm 0.11$; months $E_{\text{BC}} = 1.09 \pm 0.06$; letters 6.95 ± 0.27; ages $E_{\text{BC}} = 13.49 \pm 0.29$); on average, we see an improvement of a factor of $2.8\times$, with all statistical comparisons yielding $p < 0.001$.

We find further verification of the claim above by visualizing output trajectories in behavior space (Fig. 3). Manifold steering consistently traces paths close to $\mathcal{M}_y$, while linear steering cuts through regions far from the behavior manifold, yielding less natural outputs. This result provides causal support for the correspondence between activation and output geometry, and establishes manifold steering as a principled form of steering that yields natural behavioral trajectories ($\mathcal{M}_h \to \mathcal{M}_y$). Next, we explore whether we can find evidence from the opposite direction ($\mathcal{M}_y \to \mathcal{M}_h$).

3.3 Behavior Space Geometry Recovers the Activation Manifold

In this section, we aim to uncover whether steering intervention paths optimized to follow the behavior manifold $\mathcal{M}_y$ recover the activation manifold $\mathcal{M}_h$. To do so, we first take a path π_y^* along $\mathcal{M}_y$ in behavior space. We work within the layer and the first 32 dimensions of the subspace in which we fit the activation manifold (64 dimensional PCA), and optimize via L-BFGS for a path in activation space which, upon intervention, induces the behavioral path π_y^* (see App. F.8 for further details regarding the optimization procedure). We call the resulting path in activation space the *pullback*.

To quantify how faithfully a pullback path in activation space π_h^{pullback} recapitulates the manifold steering path π_h^*, we report an *intrinsic* R^2. Both paths are projected into a common subspace given by the singular directions explaining 99% of the variance in π_h^*, restricting the comparison to directions where the path actually extends. We then compute the R^2 in this subspace, defining the residual at each point of π_h^{pullback} as its orthogonal closest-point distance to π_h^*. We compute this score for each optimized pullback path π_h^{pullback} and compare with a linear path baseline (see App. F.9 for more details).

Results are shown in Fig. 4. We find that the pullback activation paths follow the activation manifold $\mathcal{M}_h$ more closely than the linear steering path, and resemble the shape of the manifold steering paths of §3.2 (weekdays $R^2_{\text{pullback}} = 0.77 \pm 0.03$ vs. $R^2_{\text{linear}} = 0.42 \pm 0.07$; months $R^2_{\text{pullback}} = 0.75 \pm 0.04$ vs. $R^2_{\text{linear}} = 0.32 \pm 0.05$; ages $R^2_{\text{pullback}} = 0.47 \pm 0.05$ vs. $R^2_{\text{linear}} = 0.24 \pm 0.01$; letters $R^2_{\text{pullback}} = 0.78 \pm 0.04$ vs. $R^2_{\text{linear}} = 0.23 \pm 0.03$. All statistical comparisons yield $p < 0.001$). Again we see a striking correspondence between representation and behavior; despite being derived from different sources—manifold steering from the density of activations and pullback from the structure of outputs—the two geometries are tightly connected.

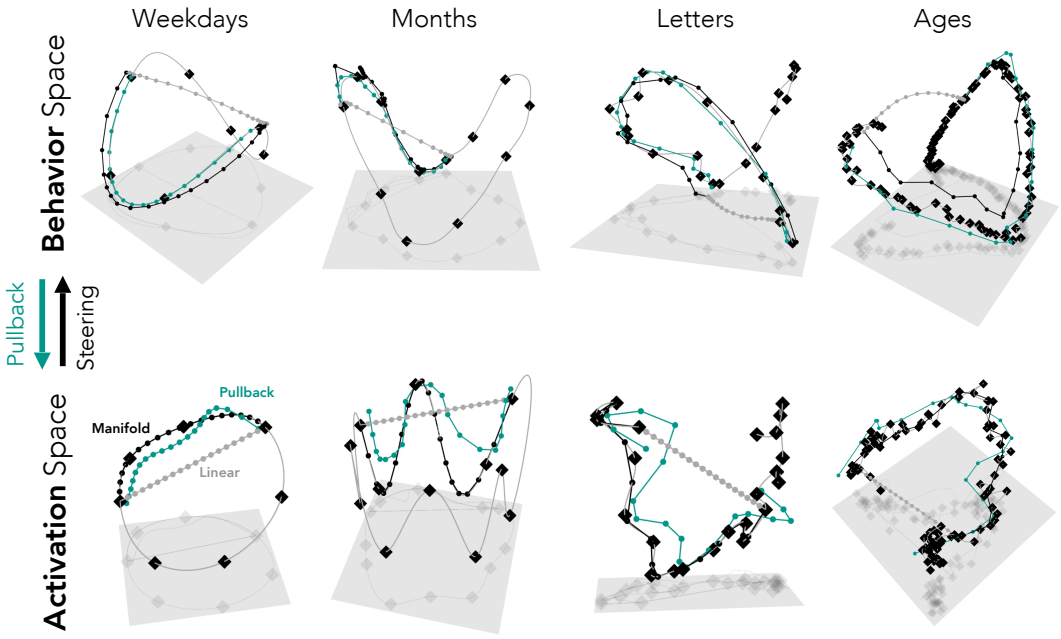

Figure 4: **Manifold steering and pullback yield coinciding trajectories in activation and behavior space.** Going in the **Activations**→**Behavior** direction, we find that steering along the activation manifold $\mathcal{M}_h$ (black) produces paths that lie close to the behavior manifold $\mathcal{M}_y$. We then examine the reverse direction, **Activations**←**Behavior**: We start with paths along the behavior manifold and optimize for corresponding paths in activation space (i.e., a set of activations that yields the path on $\mathcal{M}_y$ upon intervention). This *pullback* procedure (teal) yields trajectories that resemble the activation manfiold $\mathcal{M}_h$. Thus, we offer bidirectional support for the connection between activation geometry and behavior, and their correspondence reflecting a shared underlying conceptual organization. Paths shown: weekdays 'Thursday' to 'Sunday'; Months 'August' to 'December'; Letters 'C' to 'Q'; Ages 36 to 91.

Taken together, these results, alongside those of §3.2, provide bidirectional support for the connection between activation geometry and behavior. This convergence indicates that $\mathcal{M}_h$ is a core object in the model's representation: the geometry of activation space and the geometry of behavior are alternate views of the same underlying conceptual organization.

3.4 Unifying Steering Strategies Through Geometry

In the sections above, we analyzed three methods for steering between two points in activation space that each assume a different geometry: linear steering, which assumes a flat Euclidean geometry; manifold steering, which derives a geometry from naturally occurring activations; and pullback steering, which derives a geometry from naturally occurring output distributions. We provided empirical support for our hypothesis that the geometries derived from internal activations and output behaviors are much more similar to each other than the standard Euclidean geometry. We now formalize the question of how to steer as *how to choose the right geometry for activation space*.

The Geometry of Steering: Consider a Riemannian metric $\boldsymbol{G}$, which assigns an inner product at each point of $\mathcal{A}$; together with a path $\boldsymbol{\pi} : [0, 1] \to \mathcal{A}$, this defines the notion of path length as follows.

$$L_{\boldsymbol{G}}(\boldsymbol{\pi}) \;=\; \int_0^1 \sqrt{\dot{\boldsymbol{\pi}}(t)^\top \, \boldsymbol{G}(\boldsymbol{\pi}(t)) \, \dot{\boldsymbol{\pi}}(t)}\, dt. \tag{4}$$

Then, a geodesic is defined as the path of minimum length between two endpoints, and each choice of geometry picks out a steering strategy. The strategies of linear steering and manifold steering (§3.2), written as interpolations in two different coordinate systems (Eqs. 1, 2), are two such choices; the pullback procedure of §3.3 is a third. Now, we make all three geometries explicit.

Definition 1 (Geometries of Steering)**.** *Let $E : \mathcal{A} \to \mathbb{R}$ be an energy function such that $E(\boldsymbol{h}) \propto -\log p(\boldsymbol{h})$, and let $\boldsymbol{g}_y$ be a chosen Riemannian metric on $\mathcal{M}_y$. We define:*

$$\boldsymbol{G}_I = \boldsymbol{I}_n, \qquad \textit{(linear steering)} \tag{5}$$

$$\boldsymbol{G}_E(\boldsymbol{h}) = \left(\alpha\, e^{-E(\boldsymbol{h})} + \beta\right)^{-1} \boldsymbol{I}_n, \qquad \textit{(manifold steering)} \tag{6}$$

$$\boldsymbol{G}_F(\boldsymbol{h}) = \boldsymbol{J}_{\boldsymbol{F}}(\boldsymbol{h})^\top\, \boldsymbol{g}_y\big(\boldsymbol{F}(\boldsymbol{h})\big)\, \boldsymbol{J}_{\boldsymbol{F}}(\boldsymbol{h}) + \epsilon\, \boldsymbol{I}_n, \qquad \textit{(pullback)} \tag{7}$$

where $\alpha, \beta > 0$ are calibration constants, $\epsilon > 0$ regularizes the pullback, $\mathbf{F} : \mathcal{A} \to \mathcal{Y}$ is the function from naturally occurring activations to naturally occurring behaviors, and $\boldsymbol{g}_y$ is any Riemannian metric on $\mathcal{M}_y$ (e.g., the induced Hellinger metric used in our experiments).

We discuss the intuitive interpretation of Defn. 1 below.

- **The Flat Geometry $\boldsymbol{G}_I$.** Linear steering treats activation space as Euclidean: all directions and regions are equally valid, with Geodesics as straight lines $\boldsymbol{\ell}(t) = (1-t)\,\boldsymbol{h}_0 + t\,\boldsymbol{h}_1$. This geometry thus encodes no knowledge of naturally occurring activation or outputs.
- **The Density Geometry $\boldsymbol{G}_E$.** Manifold steering derives a geometry for activation space from naturally occurring internal representations. Specifically, consider the geometry induced from an energy function $E(\boldsymbol{h}) \propto -\log p(\boldsymbol{h})$ by rescaling the identity according to local density. Here $e^{-E(\boldsymbol{h})}$ plays the role of an unnormalized density: large where activations concentrate (on $\mathcal{M}_h$) and small where they are sparse (off $\mathcal{M}_h$). The inverse makes off-manifold regions expensive and on-manifold movement cheap, with constants $\alpha, \beta > 0$ calibrating the dynamic range (Béthune et al., 2025). Geodesics under $\boldsymbol{G}_E$ thus follow $\mathcal{M}_h$, recovering manifold steering.
- **The Pullback Geometry $\boldsymbol{G}_F$.** The steering path given by pullback derives geometric structure from naturally occurring model outputs. Specifically, $\boldsymbol{G}_F$ is the pullback of a chosen geometry on $\mathcal{M}_y$ through the Jacobian of the map from activation space to behavior space $\boldsymbol{F} : \mathcal{A} \to \mathcal{Y}$. By construction, path length under $\boldsymbol{G}_F$ equals path length of the induced behavioral trajectory along $\mathcal{M}_y$ (up to a regularization term). Geodesics under $\boldsymbol{G}_F$ are therefore activation paths whose induced behavioral trajectories are geodesics on $\mathcal{M}_y$—exactly the pullback construction of §3.3. The regularization $\epsilon\, \boldsymbol{I}_n$ ensures positive definiteness, since $\boldsymbol{J}_{\boldsymbol{F}}$ has rank at most $|\mathcal{Z}| - 1 \ll n$; as ϵ tends to 0, the geometry approaches the pure pullback in the range of $\boldsymbol{J}_{\boldsymbol{F}}$ and remains Euclidean in its null space.

Overall, we claim that while the metrics $\boldsymbol{G}_E$ and $\boldsymbol{G}_F$ are derived from different sources (internal activations and outputs, respectively), they converge on approximately the same paths in activation space (§3.3). This suggests the manifolds $\mathcal{M}_h$ and $\mathcal{M}_y$ are two images of the same conceptual geometry, related by an approximate Riemannian isometry. Consequently, the question of optimally steering model behavior boils down to isolating the geometry of a concept and defining operators to navigate it.

4 Discussion

Geometry-aware steering reveals the shared structure of behavior and representation. We build out an empirical phenomenology that relates structure in activation space to the model output behavior. First, we show an isometry between representations and behavior manifolds, i.e., distance between two points on the activation manifold $\mathcal{M}_h$ aligns with distance between the distributions induced by those points on the behavior manifold $\mathcal{M}_y$. Second, we show that steering representations along geodesics on $\mathcal{M}_h$ induces smooth, coherent transitions in behavior that follow geodesics on $\mathcal{M}_y$. Third, we show that optimizing interventions to produce behaviors following geodesics on $\mathcal{M}_y$ recover trajectories in activation space that follow $\mathcal{M}_h$. Thus, we establish a causal bridge between representation and behavior that reveals shared structure reflecting underlying conceptual geometry.

Our results also suggest that pathologies of linear steering—brittleness, incoherence, off-target effects (Wu et al., 2025; Bigelow et al., 2025; Da Silva et al., 2025; Bhalla et al., 2024; Tan et al., 2024)—stem from the mismatch between assumed flat geometry and the true curved geometry of representation space, rather than an inherent challenge with representation-based intervention. This reframes the challenge of steering from "finding the right direction" to "finding the right geometry".

References

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14638–14663, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.785.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018. doi: 10.1162/tacl_a_00034. URL https://aclanthology.org/Q18-1034/.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Identifying and controlling important neurons in neural machine translation, 2018. URL https://arxiv.org/abs/1811.01157.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1z-PsR5KX.

Sander Beckers and Joseph Halpern. Abstracting causal models. In *AAAI Conference on Artificial Intelligence*, 2019.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Jacob LS Bellmund, Peter Gärdenfors, Edvard I Moser, and Christian F Doeller. Navigating cognition: Spatial codes for human thinking. *Science*, 362(6415):eaat6766, 2018.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: perfect linear concept erasure in closed form. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html.

Louis Béthune, David Vigouroux, Yilun Du, Rufin VanRullen, Thomas Serre, and Victor Boutin. Follow the energy, find the path: Riemannian metrics from energy-based models. *ArXiv e-print*, 2025.

Usha Bhalla, Suraj Srinivas, Asma Ghandeharioun, and Himabindu Lakkaraju. Towards unifying interpretability and control: Evaluation via intervention. *ArXiv e-print*, 2024.

Eric Bigelow, Daniel Wurgaft, YingQiao Wang, Noah Goodman, Tomer Ullman, Hidenori Tanaka, and Ekdeep Singh Lubana. Belief dynamics reveal the dual nature of in-context learning and activation steering, 2025. URL https://arxiv.org/abs/2511.00617.

Eric J Bigelow, Ekdeep Singh Lubana, Robert P Dick, Hidenori Tanaka, and Tomer D Ullman. In-context learning dynamics with random binary sequences. *arXiv preprint arXiv:2310.17639*, 2023.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Nick Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Kay Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for bert. *arXiv preprint arXiv:2104.07143*, 2021.

F.L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989. doi: 10.1109/34.24792.

Matthew Brand. Charting a manifold. *Advances in neural information processing systems*, 15, 2002.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. *ArXiv e-print*, 2024.

Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. Thread: Circuits. *Distill*, 2020. doi: 10.23915/distill.00024. https://distill.pub/2020/circuits.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Thomas H Costello, Kellin Pelrine, Matthew Kowal, Antonio A Arechar, Jean-François Godbout, Adam Gleave, David Rand, and Gordon Pennycook. Large language models can effectively convince people to believe conspiracies. *arXiv preprint arXiv:2601.05050*, 2026.

Róbert Csordás, Christopher Potts, Christopher D Manning, and Atticus Geiger. Recurrent neural networks learn to store and generate sequences using non-linear representations. *arXiv preprint arXiv:2408.10920*, 2024.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *ArXiv e-print*, 2023.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. Steering off course: Reliability challenges in steering language models. *arXiv preprint arXiv:2504.04635*, 2025.

Xander Davies, Max Nadeau, Nikhil Prakash, Tamar Rott Shaham, and David Bau. Discovering variable binding circuitry with desiderata. *CoRR*, abs/2307.03637, 2023. doi: 10.48550/ARXIV.2307.03637. URL https://doi.org/10.48550/arXiv.2307.03637.

Jean Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In Walter Schempp and Karl Zeller (eds.), *Constructive Theory of Functions of Several Variables*, pp. 85–100, Berlin, Heidelberg, 1977. Springer Berlin Heidelberg. ISBN 978-3-540-37496-1.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. In *Proceedings of the 2020 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, November 2020. doi: 10.18653/v1/W18-5426.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022a.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022b.

David A Ellis, Richard Wiseman, and Rob Jenkins. Mental representations of weekdays. *PloS one*, 10(8):e0134555, 2015.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. *arXiv preprint arXiv:2405.14860*, 2024.

Thomas Fel, Ekdeep Singh Lubana, Jacob S Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba Ba, and Talia Konkle. Archetypal sae: Adaptive and stable dictionary learning for concept extraction in large vision models. *Proceedings of the International Conference on Machine Learning (ICML)*, 2025a.

Thomas Fel, Binxu Wang, Michael A Lepori, Matthew Kowal, Andrew Lee, Randall Balestriero, Sonia Joseph, Ekdeep S Lubana, Talia Konkle, Demba Ba, et al. Into the rabbit hull: From task-relevant concepts in dino to minkowski geometry. *arXiv preprint arXiv:2510.08638*, 2025b.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zb3b6oKO77.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Atticus Geiger, Kyle Richardson, and Christopher Potts. Neural natural language inference models partially embed theories of lexical entailment and negation. In Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad (eds.), *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 163–173, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.16. URL https://aclanthology.org/2020.blackboxnlp-1.16/.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 2025a. URL http://jmlr.org/papers/v26/23-0058.html.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025b. URL http://jmlr.org/papers/v26/23-0058.html.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models, 2023. URL https://arxiv.org/abs/2304.14767.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 240–248, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5426. URL https://aclanthology.org/W18-5426/.

Satchel Grant, Noah D. Goodman, and James L. McClelland. Emergent symbol-like number variables in artificial neural networks, 2025. URL https://arxiv.org/abs/2501.06141.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks, 2026. URL https://arxiv.org/abs/2511.04638.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *ArXiv e-print*, 2019.

Clément Guerner, Anej Svete, Tianyu Liu, Alexander Warstadt, and Ryan Cotterell. A geometric notion of causal probing. *CoRR*, abs/2307.15054, 2023. doi: 10.48550/ARXIV.2307.15054. URL https://doi.org/10.48550/arXiv.2307.15054.

Yoav Gur-Arieh, Roy Mayan, Chen Agassy, Atticus Geiger, and Mor Geva. Enhancing automated interpretability with output-centric feature descriptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. URL https://aclanthology.org/2025.acl-long.288/.

Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task. *arXiv preprint arXiv:2601.04480*, 2026.

Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. The MIT Press, 03 2000. ISBN 9780262273558. doi: 10.7551/mitpress/2076.001.0001. URL https://doi.org/10.7551/mitpress/2076.001.0001.

David Ha and Jürgen Schmidhuber. World models. In *Advances in Neural Information Processing Systems*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Yixiong Hao, Ayush Panda, Stepan Shabalin, and Sheikh Abdur Raheem Ali. Patterns and mechanisms of contrastive activation engineering. *arXiv preprint arXiv:2505.03189*, 2025.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 1982.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. Ravel: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. URL https://aclanthology.org/2024.acl-long.470.

David Hume. *An Enquiry Concerning Human Understanding*. A. Millar, London, 1748.

Iolo Jones. Diffusion geometry. *arXiv preprint arXiv:2405.10858*, 2024.

Iolo Jones and David Lanners. Computing diffusion geometry. *arXiv preprint arXiv:2602.06006*, 2026.

Subhash Kantamneni and Max Tegmark. Language models use trigonometry to do addition, 2025. URL https://arxiv.org/abs/2502.00873.

Dhruva Karkada, Daniel J. Korchinski, Andres Nava, Matthieu Wyart, and Yasaman Bahri. Symmetry in language statistics shapes the geometry of model representations, 2026. URL https://arxiv.org/abs/2602.15029.

Daniel J Korchinski, Dhruva Karkada, Yasaman Bahri, and Matthieu Wyart. On the emergence of linear analogies in word embeddings. *arXiv preprint arXiv:2505.18651*, 2025.

Austin C. Kozlowski, Callin Dai, and Andrei Boutyline. Semantic structure in large language model embeddings, 2025. URL https://arxiv.org/abs/2508.10003.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 2006.

Michael A Lepori, Tal Linzen, Ann Yuan, and Katja Filippova. Language models struggle to use representations learned in-context. *arXiv preprint arXiv:2602.04212*, 2026.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure, 2017. URL https://arxiv.org/abs/1612.08220.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Ekdeep Singh Lubana, Can Rager, Sai Sumedh R Hindupur, Valerie Costa, Greta Tuckute, Oam Patel, Sonia Krishna Murthy, Thomas Fel, Daniel Wurgaft, Eric J Bigelow, et al. Priors in time: Missing inductive biases for language model interpretability. *arXiv preprint arXiv:2511.01836*, 2025.

Niru Maheswaranathan, Alex Williams, Matthew Golub, Surya Ganguli, and David Sussillo. Universality and individuality in neural dynamics across large populations of recurrent networks. *Advances in neural information processing systems*, 32, 2019.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2023.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Marina Meilă and Hanyu Zhang. Manifold learning: what, how, and why, 2023. URL https://arxiv.org/abs/2311.03757.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022. arXiv:2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=MkbcAHIYgyS.

Jack Merullo, Noah A. Smith, Sarah Wiegreffe, and Yanai Elazar. On linear representations and pretraining data frequency in language models, 2025. URL https://arxiv.org/abs/2504.12459.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models, 2025a. URL https://arxiv.org/abs/2505.18235.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models, 2025b. URL https://arxiv.org/abs/2505.18235.

Andrew William Moore. *Efficient Memory-Based Learning for Robot Control*. PhD thesis, University of Cambridge, 1990.

Depen Morwani, Benjamin L. Edelman, Cosmin Oncescu, Rosie Zhao, and Sham M. Kakade. Feature emergence via margin maximization: Case studies in algebraic tasks. In *The Twelfth International Conference on Learning Representations*, 2024.

Aaron Mueller, Jannik Brinkmann, Millicent L. Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability. *CoRR*, abs/2408.01416, 2024. doi: 10.48550/ARXIV.2408.01416. URL https://doi.org/10.48550/arXiv.2408.01416.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. Mib: A mechanistic interpretability benchmark, 2025. URL https://arxiv.org/abs/2504.13151.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: Surveying mechanistic interpretability for nlp through the lens of causal mediation analysis. *Computational Linguistics*, pp. 1–48, 02 2026. ISSN 0891-2017. doi: 10.1162/COLI.a.572. URL https://doi.org/10.1162/COLI.a.572.

Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task, 2024.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024. URL https://arxiv.org/abs/2312.06681.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. Iclr: In-context learning of representations. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 53258–53284, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/83fe5a77502e3d4cfab5960aed0ee6c3-Paper-Conference.pdf.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. Iclr: In-context learning of representations. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025c.

Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The information geometry of softmax: Probing and steering. *arXiv preprint arXiv:2602.15293*, 2026.

Michael Pearce, Elana Simon, Michael Byun, and Daniel Balsam. Finding the tree of life in evo 2. *Goodfire*, August 2025. Correspondence to michael@goodfire.ai.

Judea Pearl. Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, 121(1):93–149, 1999.

Judea Pearl. Direct and indirect effects, 2001. URL https://arxiv.org/abs/1301.2300.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs, 2025. URL https://arxiv.org/abs/2505.14685.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. Towards reliable evaluation of behavior steering interventions in llms. *arXiv preprint arXiv:2410.17245*, 2024.

Lucas Prieto, Edward Stevinson, Melih Barsbey, Tolga Birdal, and Pedro A. M. Mediano. Correlations in the data lead to semantically rich feature geometry under superposition. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7akSRQS5Xh.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. From causal to concept-based representation learning. *Advances in Neural Information Processing Systems*, 37:101250–101296, 2024a.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning interpretable concepts: Unifying causal representation learning and foundation models. *arXiv preprint arXiv:2402.09236*, 2024b.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure, 2022.

Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. Log-linear guardedness and its implications. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 9413–9431. Association for Computational Linguistics, 2023a. doi: 10.18653/V1/2023.ACL-LONG.523. URL https://doi.org/10.18653/v1/2023.acl-long.523.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Kernelized concept erasure, 2023b.

Christian H Reinsch. Smoothing by spline functions. *Numerische mathematik*, 10(3):177–183, 1967.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Juan Diego Rodriguez, Aaron Mueller, and Kanishka Misra. Characterizing the role of similarity in the property inferences of language models. *CoRR*, abs/2410.22590, 2024. doi: 10.48550/ARXIV.2410.22590. URL https://doi.org/10.48550/arXiv.2410.22590.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=l2zFn6TIQi.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Naomi Saphra and Sarah Wiegreffe. Mechanistic? *CoRR*, abs/2410.09087, 2024. doi: 10.48550/ARXIV.2410.09087. URL https://doi.org/10.48550/arXiv.2410.09087.

Raphaël Sarfati, Eric Bigelow, Daniel Wurgaft, Jack Merullo, Atticus Geiger, Owen Lewis, Tom McGrath, and Ekdeep Singh Lubana. The shape of beliefs: Geometry, dynamics, and interventions along representation manifolds of language models' posteriors, 2026. URL https://arxiv.org/abs/2602.02315.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pp. 583–588. Springer, 1997.

Adam Shai, Loren Amdahl-Culleton, Casper L. Christensen, Henry R. Bigelow, Fernando E. Rosas, Alexander B. Boyd, Eric A. Alt, Kyle J. Ray, and Paul M. Riechers. Transformers learn factored representations, 2026. URL https://arxiv.org/abs/2602.02385.

Adam S Shai, Sarah E Marzen, Lucas Teixeira, Alexander G Oldenziel, and Paul M Riechers. Transformers represent belief state geometry in their residual stream. *Advances in Neural Information Processing Systems*, 37:75012–75034, 2024a.

Adam S. Shai, Sarah E. Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream, 2024b. URL https://arxiv.org/abs/2405.15943.

Roger N Shepard. Toward a universal law of generalization for psychological science. *Science*, 237 (4820):1317–1323, 1987.

P. Smolensky. *Neural and conceptual interpretation of PDP models*, pp. 390–431. MIT Press, Cambridge, MA, USA, 1986. ISBN 0262631105.

Jiajun Song and Yiqiao Zhong. Uncovering hidden geometry in transformers via disentangling position and context. *arXiv preprint arXiv:2310.04861*, 2023.

Yang Song and Diederik P Kingma. How to train your energy-based models. *ArXiv e-print*, 2021.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 7035–7052. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.435. URL https://doi.org/10.18653/v1/2023.emnlp-main.435.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48/.

Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability?, 2025. URL https://arxiv.org/abs/2507.08802.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. *Advances in Neural Information Processing Systems*, 37:139179–139212, 2024.

Generalist AI Team. Gen-0: Embodied foundation models that scale with physical interaction. *Generalist AI Blog*, 2025. https://generalistai.com/blog/nov-04-2025-GEN-0.

Joshua B Tenenbaum and Thomas L Griffiths. Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, 24(4):629–640, 2001.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization, 2023.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Daniel Vennemeyer, Phan Anh Duong, Tiffany Zhan, and Tianyu Jiang. Sycophancy is not one thing: Causal separation of sycophantic behaviors in llms. *arXiv preprint arXiv:2509.21305*, 2025.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for score-based conditional model. In *ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2023b.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders, 2025. URL https://arxiv.org/abs/2501.17148.

Daniel Wurgaft, Ekdeep Singh Lubana, Core Francisco Park, Hidenori Tanaka, Gautam Reddy, and Noah D. Goodman. In-context learning strategies emerge rationally, 2025. URL https://arxiv.org/abs/2506.17859.

Yongyi Yang, Hidenori Tanaka, and Wei Hu. Provable low-frequency bias of in-context learning of representations. *arXiv preprint arXiv:2507.13540*, 2025.

Julian Yocum, Cameron Allen, Bruno Olshausen, and Stuart Russell. Neural manifold geometry encodes feature fields. In *NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, 2025.

Carolina Zheng, Nicolas Beltran-Velez, Sweta Karlekar, Claudia Shi, Achille Nazaret, Asif Mallik, Amir Feder, and David M Blei. Model directions, not words: Mechanistic topic models using sparse autoencoders. *arXiv preprint arXiv:2507.23220*, 2025.

Tianyi Zhou, Deqing Fu, Mahdi Soltanolkotabi, Robin Jia, and Vatsal Sharan. Fone: Precise single-token number embeddings via fourier features. *arXiv preprint arXiv:2502.09741*, 2025.

A Manifold Steering Yields Factored Control in Multi-Dimensional Spaces

Our experiments thus far have been limited to one dimensional conceptual spaces arising from training data imbued with real-world structure, i.e., days, months, ages, and letters. In turn, the manifolds we found have been one dimensional curves with a single intrinsic coordinate. Now, we extend our results to a setting with two dimensional conceptual spaces whose geometry are defined via in-context learning. We fit manifolds and show there is a two dimensional intrinsic coordinate system for the manifold, where steering along each coordinate controls an independent dimension of the conceptual space.

In-context learning tasks with synthetic conceptual spaces. Park et al. (2025b) introduce a family of tasks to study the in-context learning of representations (ICLR). For each task, arbitrary tokens are assigned to a discrete graphical structure and language models are supplied with sequences of tokens derived from a random walk on that graph. They show that the statistical patterns in the random walk of tokens induce a reorganization of representations that recapitulates the graphical structure used to generate data. This in turn enables the language model to match the next token distribution, i.e., predict tokens adjacent to the current location of the random walk on the grid. For our two experiments, we assign arbitrary tokens to grid and cylinder graph structures. Thus, the conceptual domain $\mathcal{Z}$ is the set of tokens and the distance metric $d_{\mathcal{Z}}$ is distance on the graph used to generate the random walk. Fig. 5 (a) shows an example grid and an input prompt generated by a random walk.

Manifold fitting. The ICLR grid manifold $\mathcal{M}_h$ is topologically described by a two dimensional surface with no holes or tears. The activation geometry of $\mathcal{M}_h$ is more complex. Its semi-spherical shape shown in Fig. 5 (c) is induced by task statistics: the random walk visits inner sites more frequently than peripheral sites, leading to slight distortions with respect to the ground truth geometry (Park et al., 2025b; Yang et al., 2025; Karkada et al., 2026). We fit two-dimensional sheets to internal activations and output distributions via thin plate splines (TPS; Duchon 1977; Bookstein 1989), which can be seen as the 2D analog of the cubic splines used previously. In this case, we use activations corresponding to the last token in the context, and compute centroids according to graph location at a given timestep. Then, TPS finds the smoothest surface interpolating through the centroids (see App. F.3 for further details).

Isometry results. For each ICLR domain, we compute pairwise distance matrices over graph-node centroids under three metrics: Euclidean (linear) distance in the activation subspace, geodesic distance along the fitted activation manifold $\mathcal{M}_h$, and geodesic distance along the behavior manifold $\mathcal{M}_y$. We find very high correlations between geodesic paths on the activation and behavior manifolds ($r = .99$ for both the 5×5 grid and 9×9 cylinder domains) and reduced correlations for linear paths (5×5 grid $r = 0.90$;9×9 cylinder $r = 0.81$). To further examine these results, we embed each distance matrix with multidimensional scaling (MDS). Fig. 5(b) shows a clean grid structure in the activation manifold and behavior manifold embeddings, while the linear paths in activation space yield a warped surface. Again, we see the conceptual space recapitulated in both representation and behavior.

Manifold vs. Linear steering. We next test whether the fitted two-dimensional activation manifold affords coherent, factored control over the graph geometry used to generate the ICLR inputs. In particular, we assess whether we can control the position of the random walk input via intervention, and, moreover, whether there is an intrinsic coordinate system where steering along each coordinate independently controls the horizontal and vertical position on the grid. For each ordered pair of nodes, we use manifold steering and linear steering to interpolate between the start and end centroid and average results over 5 input prompts.

The top panel of Fig. 5(c) shows that manifold steering produces smooth transitions vertically along the steered graph dimension while remaining at the same horizontal position. The bottom panel of Fig. 5(c) shows similar smooth transitions but along a horizontal dimension while keeping the same vertical position. This demonstrates the manifold has an intrinsic coordinate system corresponding to the two dimensions of the grid, enabling factored control. Furthermore, this shows the smooth and ordered transitions of manifold steering generalize to multi-dimensional spaces. In contrast, linear steering again fails to provide ordered transitions through grid locations, and shows very clear 'teleportation' behavior between the endpoint locations along its path.

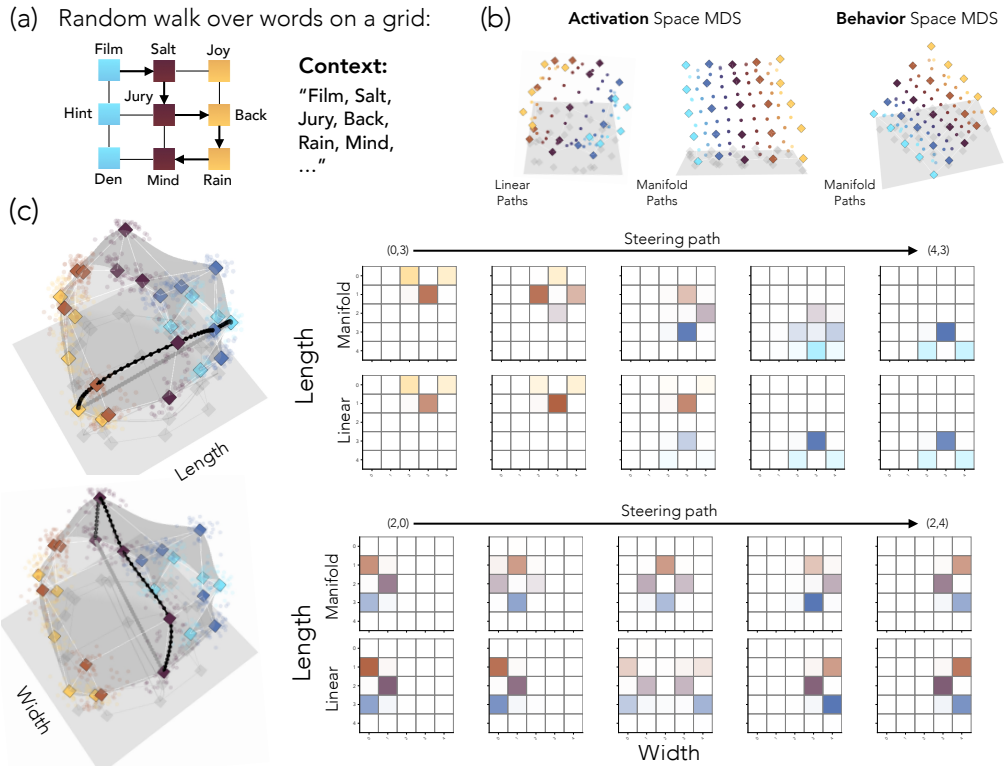

Figure 5: **Manifold steering enables factored control in multi-dimensional conceptual spaces.** **(a)** We examine manifold steering on multidimensional spaces using Park et al. (2025b)'s in-context learning of representations (ICLR) task. In an ICLR task, arbitrary tokens are assigned to nodes along a graph, and a language model is prompted with tokens from a random walk along the graph. Park et al. (2025b) showed that with sufficient context, models encode the structure of the latent graph in their activations. In this work, we study two graph structures learned in-context (5×5 grid shown above, 9×9 cylinder in App. H). We fit manifolds to activations and output behaviors and intervene on activations using linear and manifold steering. **(b)** We examine the mapping between the activation and behavior manifolds by computing on-manifold and linear distances in activation space and comparing them to on-manifold distances in behavior space via a multidimensional scaling (MDS) embedding. We find a clear structural match of both the activation and output manifolds with the latent graph, providing direct evidence for these two manifolds encoding a similar underlying conceptual space. In contrast, linear distances in activation space yield a warped structure. **(c)** We find that manifold steering maintains the quality of smooth and ordered transitions beyond one dimension, and in conceptual spaces learned in-context. Furthermore, we find that steering along one dimension leads to minimal off-target impact, thus affording the appealing quality of factored control. In contrast, linear steering maintains its teleportation behavior.

B Manifold Steering on a Visual World Model: Mountain Car Task

We now ask whether the same principles of geometry-aware steering extend to the *visual* domain of world models. This question is practically motivated: learned world models that predict future observations from past frames and actions are central to model-based reinforcement learning and robotic planning (Ha & Schmidhuber, 2018; Hafner et al., 2020; Team, 2025; Black et al., 2024). If the internal representations of such models admit geometric structure, manifold-based steering could provide a principled mechanism for intervening on a model's behavior through changing its beliefs about the state of the world.

Environment and model architecture. We train a recurrent world model on the Mountain Car environment (Moore, 1990; Sutton & Barto, 2018), a classical control task in which a car must escape a valley by building momentum. The environment has continuous position $p \in [-1.2, 0.6]$,

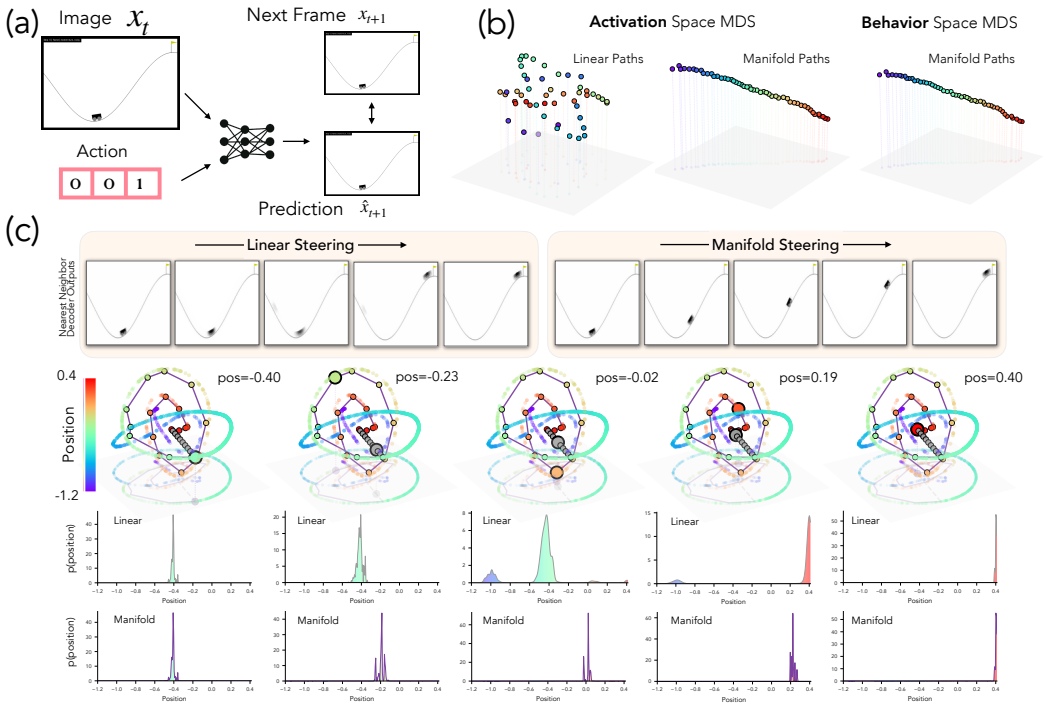

Figure 6: **Manifold steering on a visual world model produces smooth movement. (a)**. We examine whether manifold steering can generalize to a visual modality by training a recurrent network on the Mountain Car environment (Moore, 1990; Sutton & Barto, 2018) to predict the next frame x_{t+1} given the previous frame and an action. **(b)** We test the mapping between the activation and behavior manifolds by computing on-manifold and linear distances in activation space and comparing them to on-manifold distances in behavior space via an MDS embedding. On-manifold paths in activation and behavior space both recover a clean sequential ordering corresponding to location, while the embedding of linear distances scrambles it. **(c) Middle:** PCA visualization of the activation manifold and five waypoints along a path between $p_A = -0.40$ and $p_B = 0.40$ for both linear (red) and manifold (blue) steering. **Top**: At intermediate positions along the linear steering path, decoding shows the car as blurred or ambiguously placed, reflecting an incoherent superposition of positional beliefs (**bottom**) as the path departs from the activation manifold $\mathcal{M}_h$. In contrast, manifold steering along $\mathcal{M}_h$ yields smooth movement of the car up the hill.

continuous velocity $v \in [-0.07, 0.07]$, and three discrete actions (left, no-op, right). The model predicts the next frame x_{t+1} given the previous frame x_t and action a_t (see Fig. 6(a) for an illustration). The full architecture is shown in Fig. 7: A convolutional encoder maps each $128 \times 128 \times 3$ RGB frame to a latent vector, v_t, which is concatenated with a learned action embedding $e(a_t) \in \mathbb{R}^{16}$ and fed to a Gated Recurrent Unit (GRU; Cho et al. 2014):

$$\mathbf{h}_t = \mathrm{GRU}\big([v_t;\, e(a_t)],\ \mathbf{h}_{t-1}\big) \in \mathbb{R}^n; \quad v_t = \mathrm{LayerNorm}\big(f_{\mathrm{enc}}(x_t)\big) \in \mathbb{R}^n, \tag{8}$$

where $n = 64$. A convolutional decoder produces a residual image from the hidden state, yielding the prediction $\hat{x}_{t+1} = x_t + f_{\mathrm{dec}}(\mathbf{h}_t)$.

Activation and behavior manifold fitting. For this setting, we consider *position* to play the role of the conceptual domain $\mathcal{Z} = [p_{\min}, p_{\max}]$ and aim to capture the manifold structure in both activation and behavior space of the vision encoder. To start, we first collect encoder activations from 100 rollouts in the environment (see §G.1 for details) and observe they occupy a curved, low-dimensional manifold $\mathcal{M}_h \subset \mathbb{R}^n$ (Fig. 6(c)). We parameterize this manifold by partitioning the position range into bins and fit a smooth spline through the means $\{\mu_b\}_{b=1}^B \subset \mathbb{R}^n$. For a given input, x, we compute the output distribution over positions, $\mathbf{p}(x)$ using the distance of the activations $v(x)$ to the centroid for each position:

$$\mathbf{p}(x) \;=\; \mathrm{softmax}\left(-\frac{\|v - \mu_b\|_2}{\tau}\right)_{b=1}^{B} \;\in\; \Delta^{B-1}, \tag{9}$$

with temperature $\tau = 0.5$. We follow §3.2 to parameterize the behavior manifold $\mathcal{M}_y$ by embedding each bin in Hellinger coordinates on the unit sphere and fit a 1D smoothing spline $\gamma_{\mathcal{M}_y} : \mathcal{Z} \to \mathbb{R}^B$ parameterized by position (full details in App. G). Note that both $\mathcal{M}_h$ and $\mathcal{M}_y$ are 1D structures parameterized by the conceptual coordinate p; under PCA visualization (Fig. 12), both trace closed curves in their respective ambient spaces. The curves are closed because visually distinctive states at the wall, $p \approx -1.2$, and goal, $p \approx 0.4$, are mapped to neighboring activations.

Geometry-aware steering in activation space. Fig. 6 compares linear (Eq. 1) and manifold (Eq. 2) steering through $K = 20$ waypoints between encoder states corresponding to positions $p_A = -0.4$ and $p_B = 0.4$, projected into the first three principal components of encoder space. The geodesic path closely tracks $\mathcal{M}_h$, and the corresponding decoded frames display a smooth, coherent progression of the car through intermediate positions. The linear path departs from $\mathcal{M}_h$ at intermediate points and decoded frames exhibit blurred or ambiguous car placement, reflecting an incoherent superposition. Then, a 'teleportation' to the endpoint is observed, analogous to the behavior of linear steering in the language model experiments. Moreover, linear steering causes the probability distribution over position to show greater spread compared to on-manifold paths, yielding the ambiguous car placement seen in intermediate points along the path. Finally, we reproduce the pullback procedure in §H.4 and show that optimizing paths in the output distribution over possible car positions results in activation space paths that closely track the $\mathcal{M}_h$ (Fig. 12).

The geometry assumed by linear steering is not faithful to the conceptual ordering. To make the difference between the two steering metrics visually concrete, we apply multidimensional scaling (MDS) to the pairwise distance matrices induced by three different distance functions over $W = 50$ anchor positions evenly spaced along $[p_{\text{min}}, p_{\text{max}}]$ (Fig. 6(b)). Both activation space and behavior space on-manifold distance embeddings recover a clean one-dimensional rainbow ordering of positions, while the linear distance embedding produces a scrambled three-dimensional structure whose colors are visibly out of order. This is because $\mathcal{M}_h$ folds back on itself in the encoder's ambient space, so two activations whose underlying positions are far apart can sit arbitrarily close in ambient space. Quantitatively, the Pearson correlation between the two arc-length distance matrices is $r = 0.99$, while correlation between activation-space linear paths and behavior manifold arc-length falls to $r = 0.06$, confirming that the linear-steering metric is not a faithful proxy for the conceptual ordering that the encoder has learned.

C Related Work

Activation Steering and the Linear Representation Hypothesis. Activation steering protocols (Bau et al., 2018; Subramani et al., 2022; Marks & Tegmark, 2023; Panickssery et al., 2024; Turner et al., 2024) are often motivated by the linear representation hypothesis (LRH)—a geometric assumption on model representations (Smolensky, 1986; Elhage et al., 2022a; Park et al., 2023; Costa et al., 2025; Zheng et al., 2025). In particular, LRH argues neural networks encode concepts, i.e., latent variables underlying the data distribution (Wang et al., 2023b; Rajendran et al., 2024a,b; Okawa et al., 2024), along directions. This motivates tools like linear probing (Belinkov, 2022; Guerner et al., 2023) and sparse autoencoders (Cunningham et al., 2023; Bricken et al., 2023; Gao et al., 2025; Bussmann et al., 2024; Fel et al., 2025a). In the case where the representation geometry for a concept truly aligns with LRH, Bigelow et al. (2025) showed the effects of activation steering on model behavior can be accurately captured by a linear increase in concept log-probability. However, in the general scenario where geometry of representations does not abide by LRH, the effects of linear steering protocols are less clear. Recent work has started to fill this gap: e.g., work by Rodriguez et al. (2025); Ravfogel et al. (2022) has shown that linear steering protocols match the first moments of the current output distribution produced by the model with the target distribution; however, the effects on higher order moments can be unconstrained and adversarial (see results by Sarfati et al. (2026)), possibly explaining why it produces incoherent outputs.

In contrast, when the representation geometry is fully respected, our work shows that steering smoothly interpolates the source and the target distributions. Prior work has shown similar results to this effect in narrow domains, e.g., Engels et al. (2024) ablate representations and write the days-of-the-week circle directly and Kantamneni & Tegmark (2025) follow a similar protocol for a helix representing numbers, but they lack a more general account of how representation geometry and output behavior map on to each other. The closest work to ours is the contemporary paper by

Park et al. (2026), who study a toy model where the representation-to-output distribution mapping is described via a simple softmax operation.

Activation Geometry and its Origins. A large body of recent work has shown neural networks encode concepts along nonlinear, curved geometries embedded in low-dimensional subspaces across both modalities and architectures (Fel et al., 2025b; Pearce et al., 2025; Yocum et al., 2025; Modell et al., 2025a; Lubana et al., 2025; Costa et al., 2025; Park et al., 2025b; Engels et al., 2024; Karkada et al., 2026; Shai et al., 2024a, 2026; Saxe et al., 2019; Park et al., 2025c; Morwani et al., 2024; Kantamneni & Tegmark, 2025; Song & Zhong, 2023; Zhou et al., 2025; Maheswaranathan et al., 2019). While earlier work (Saxe et al., 2019; Arora et al., 2018; Park et al., 2023; Yocum et al., 2025) concretized, in toy settings, how structure in the data-generating process imposes geometric constraints on a neural network's representations, only recently have such accounts been extended to make predictions about the geometry of neural representations at scale (e.g., Merullo et al. (2025)). Karkada et al. (2026) and Korchinski et al. (2025) argue that symmetries in data statistics enforce geometries best suited for reflecting the uncertainty of the distribution in a model's representation (cf. Prieto et al. 2026), and offer plausible accounts for the formation of representations in-context, as shown by works such as Park et al. (2025a) and Lepori et al. (2026).

Causal Analysis of Neural Networks. Several works have convincingly argued that tools like probing or visualization of representations are insufficient to make claims about model behavior, i.e., artifacts produced via these tools can yield misleading explanations for why a model behaves the way it does (Geiger et al., 2020; Belinkov, 2022; Bolukbasi et al., 2021; Saphra & Wiegreffe, 2024). As such, a vast array of research has used intervention on activations to study model internals (Li et al., 2017; Giulianelli et al., 2018; Cammarata et al., 2020; Elazar et al., 2020; Ravfogel et al., 2022, 2023a,b; Belrose et al., 2023; Geva et al., 2023; Meng et al., 2022, 2023; Vig et al., 2020; Geiger et al., 2020; Davies et al., 2023; Stolfo et al., 2023; Guerner et al., 2023; Wang et al., 2023a; Todd et al., 2024; Arora et al., 2024; Huang et al., 2024; Feng & Steinhardt, 2024; Mueller et al., 2025; Prakash et al., 2025; Gur-Arieh et al., 2025; Grant et al., 2025; Rodriguez et al., 2024). This interpretability research leverages the frameworks of causal mediation (Pearl, 2001; Vig et al., 2020; Mueller et al., 2026) and causal abstraction (Rubenstein et al., 2017; Beckers & Halpern, 2019; Geiger et al., 2021, 2025a,b) to ground understanding of model internals in the theory of causality (Hume, 1748; Pearl, 1999; Spirtes et al., 2000).

D Extended Discussion

Where does the shared geometry of behavior and representation come from? While we do not study the origins of the shared geometry between behavior and representation, our experimental results are consistent with the hypothesis that conceptual structure constrains the geometry of both representation and behavior. While data statistics shape the geometry of neural representations (Merullo et al., 2025; Karkada et al., 2026; Prieto et al., 2026), this fails to explain how geometric structure is formed for out-of-distribution inputs. For example, our in-context learning tasks (Sec. A) have synthetically defined geometries that imbue tokens with contextual meaning that is wildly different from the meaning learned during training. As such, the model must form novel representations and produce novel behaviors. The fact that we are able to establish a shared geometry for representation and behavior in these novel in-context learning tasks suggests that regardless of how training data statistics inform the geometries seen in the model, the output behavior is now computationally constrained by the activation geometry (see the contemporary work by Yocum et al. (2025) for a formalization of this claim).

Intrinsic coordinates of representation manifolds as units of causal analysis. Mueller et al. (2024) frames the field of mechanistic interpretability as being on a quest to discover a primitive unit of representation best suited for the causal analysis of neural network internals. Causal abstraction provides a theoretical framework for defining such units of analysis (Geiger et al., 2025a,b), however Sutter et al. (2025) point out that allowing arbitrarily complex units admits degenerate solutions. Our work suggests a path toward both answering Mueller et al. (2024) and addressing the problem identified by Sutter et al. (2025): the appropriate units of causal analysis are intrinsic coordinates on manifolds in activation space, and fitting these manifolds to naturally occurring activations provides a constraint that helps rule out degenerate solutions (cf. Grant et al. 2026).

E Future Work and Limitations

The goal of our paper was to understand the role of geometry in neural networks and, subsequently, use this understanding to concretize what it means to steer model behavior via representations. We have shown that the geometry of neural network representations provides a blueprint for effective control. When interventions respect the geometry of activation space, the change in behavior is smooth and coherent; when they ignore it, they risk producing states with no natural behavioral counterpart. While we believe our results have enabled significant progress towards the motivating goals, there are remaining limitations that need to be addressed in future work.

- **Expanding experimental validation to more complex domains.** To illustrate our arguments, we focused on simple settings for which the concept of interest had a well-defined domain (e.g., weekdays), and the expected task outputs are the concepts themselves, therefore the conceptual geometry is directly displayed in the outputs. To further validate the claims posited in this paper, future work is needed to explore more abstract concepts, e.g., refusals (Arditi et al., 2024), sycophancy (Vennemeyer et al., 2025), and persuasion (Costello et al., 2026). For such concepts, output behavior will likely reflect conceptual structure in subtler ways, and it remains to be tested whether the conceptual geometry of such concepts can be inferred from behavior and related to representations as we did in this work. Moreover, in these more complex cases, it is unclear what are the right primitives for a representational account; we may need a notion of dynamics over manifolds, a view of representation as an aggregation of several geometric structures (similar to results seen by Fel et al. (2025b) in a vision context), or perhaps an altogether different object. Even if geometry is the right substrate to work with, we emphasize the simplicity of our domains allowed us to easily isolate the target concept's geometry via synthetic, template-based text. Moving to more complex scenarios will require isolating the geometry of concepts from in-the-wild data.
- **Moving from token to sequence-level outputs.** Another way in which our tasks are simplified is our focus on the next-token distribution. This makes analysis feasible and helps avoid the combinatorial complexity involved in studying multi-token sequences. The obvious way to expand from our work's token-level focus to sequence-level focus involves formalizing arguments in the language of "beliefs", i.e., latent variables underlying the posterior predictive induced by a model in response to an input (Bigelow et al., 2023, 2025; Wurgaft et al., 2025). Correspondingly, what we expect to see via geometry-aware interventions is the nature of output sequences produced by a model will change as we perform steering: e.g., navigating the geometry of sycophancy (were it to exist) should allow us to alter the extent or type of sycophancy exhibited in model outputs; however, this property will be latent, rather than a concrete token-level change.
- **Fitting the geometry.** While we used a specific protocol to fit the observed geometries (Bookstein, 1989), we note there is a rich literature on fitting low-dimensional manifolds (Coifman & Lafon, 2006; Brand, 2002; Schölkopf et al., 1997; Roweis & Saul, 2000; Jones, 2024; Jones & Lanners, 2026; Meilă & Zhang, 2023). Critically, beyond just fitting the manifold, what we seek is an operator that allows us to navigate the manifold. For the domains analyzed in this work, we have ground-truth knowledge about how different states of the concept relate to each other, which allows us to define intrinsic coordinates for spline fitting. An unsupervised protocol would however significantly broaden the applicability of our methods.
- **Manipulating intermediate algorithmic variables.** All of our experiments are about manipulating the output behavior of neural networks directly. However, the most interesting control protocols will require manipulating intermediate quantities that mediate the flow of information from input to output, e.g., an image model determining the shape of an object in service of predicting its weight.

F Experimental Details for Language Tasks

This Appendix describes the procedures behind the experiments of Sections §2, §3, and §A. The following section (Appendix §G) will provide details for the mountain-car experiment in §B.

F.1 Tasks and Datasets

Natural domain tasks. We use four natural-domain addition tasks: weekdays and months (cyclic), and letters and ages (sequential). The full templates and entity sets are listed in Table 1. For each

task, we enumerate every (entity, increment) pair whose result lies in the task's target set, dropping pairs whose result would fall outside it (e.g. letters past Z, or ages outside $[10, 100]$). The reported activations and output distributions are computed at the answer-token position, and concept centroids are obtained by averaging across all prompts whose ground-truth result is the same value of $\mathcal{Z}$.

Task	Template	Entities	Increments	Structure	$\lvert\mathcal{D}\rvert$
Weekdays	Q: What day is {k} days after {entity}?\nA:	*Monday*, ..., *Sunday* (7)	*one*, ..., *seven*	cyclic	49
Months	Q: What month is {k} months after {entity}?\nA:	*January*, ..., *December* (12)	*one*, ..., *seven*	cyclic	84
Letters	Consider letters in the alphabet. Starting at letter {entity}, we increment by {k}. The result is letter	*C*, ..., *Z* (24)	*one*, *two*	sequential	48
Ages	Alice is {entity} years old. Bob is {k} years older than Alice. Q: How old is Bob?\nA: Bob is	*1*, ..., *99*	*1*, ..., *10*	sequential	909

Table 1: Natural-domain arithmetic tasks. For cyclic domains, results wrap around the modulus (7 for weekdays, 12 for months); for sequential domains, (entity, increment) pairs whose result falls outside the target set are filtered. The last column reports the dataset size $\lvert\mathcal{D}\rvert$ after this filter.

In-context learning of representations. For the multi-dimensional setting of §A, we use the in-context learning of representations (ICLR) family of Park et al. (2025b): arbitrary tokens corresponding to nouns (e.g., "film", "rain") are assigned to the nodes of a graph, and prompts are random walks on that graph. We study a 5×5 grid and a 9×9 cylinder, with random walks of 2048 entity tokens. As in Park et al. (2025b)'s setup, the random walks we sample do not allow backtracking, which we find aids models in learning the underlying structure.

F.2 Model, Intervention Site, and Output Distribution

We investigate Llama 3.1 8B (Touvron et al., 2023) activations at layer 28 in bfloat16 for all tasks. All interventions are performed on the residual stream at the last-token position. We chose to examine a late layer of the model to ensure that concept geometries are fully computed.

For an input x, the output distribution $\boldsymbol{p}(x) \in \mathcal{Y}$ used throughout the paper is constructed as follows: we softmax over the full vocabulary logit distribution and aggregate probability mass over each concept value's variant token spellings (e.g. the tokens `' Monday'`, `'Monday'`, and `'monday'` are all summed into the `Monday` entry). The remaining probability mass on tokens not associated with any concept value is collected into a single *'other'* bin, yielding a distribution on the open simplex $\Delta^{|\mathcal{Z}|}$ over $|\mathcal{Z}| + 1$ classes.

F.3 Fitting the Activation Manifold $\mathcal{M}_h$

To identify the activation manifold $\mathcal{M}_h$, we first obtain points in full activation space and transform them into a 64-dimensional subspace obtained via PCA over the activations $\boldsymbol{h}(x)$ across all prompts in the task. The manifold lives entirely in the 64-dimensional PCA subspace; the orthogonal complement is preserved during all subsequent interventions (§F.6).

We compute concept centroids c_i as the mean of the projected activations across all prompts whose ground-truth result equals the i-th concept value, and fit a smooth interpolant through them. For the four natural-domain tasks the interpolant is a one-dimensional cubic spline (Reinsch, 1967): a natural cubic spline (with vanishing second derivatives at the endpoints) for the sequential tasks (letters, ages), and a periodic cubic spline for the cyclic tasks (weekdays, months) so that the curve closes smoothly. For the sequential tasks we use the ground-truth ordinal index of each concept as its intrinsic coordinate. For the cyclic tasks the centroids form a near-circular loop in the top two principal components of the activation subspace, so we instead derive the intrinsic coordinate $\theta = \mathrm{atan2}(\mathrm{PC}_2, \mathrm{PC}_1)$ in an unsupervised manner.

The interpolant for the ICLR tasks is a thin-plate spline (TPS; Duchon (1977); Bookstein (1989)), a multi-dimensional generalisation of the cubic spline which minimizes the bending energy $\int \|\nabla^2 f\|^2$. Thin-plate splines map points in a lower-dimensional intrinsic space to the full ambient space. The TPS parameterisation requires a choice of intrinsic coordinates for the centroids. We use the ground-truth graph coordinates of each node in the ICLR task as intrinsic coordinates. Both the grid and cylinder tasks use the standard TPS kernel $r^2 \log r$; for the cylinder, which has both a linear and a periodic dimension, we additionally apply a ghost-point procedure where each control point is duplicated at one period above and below its θ value (and we drop the linear-in-θ polynomial column) to enforce closure across the periodic dimension. In every case the spline interpolates the centroids exactly, so $\mathcal{M}_h$ passes through every c_i.

F.4 Fitting the Behavior Manifold $\mathcal{M}_y$

Behavior centroids $b_i = \bar{\boldsymbol{p}}_i$ are computed analogously to the activation centroids, by averaging the model's output distributions across all prompts whose ground-truth result equals the i-th concept value. Because the probability simplex is not a proper metric space, we map each centroid into Hellinger coordinates via $b_i \mapsto \sqrt{b_i}$, placing it on the non-negative orthant of the unit ℓ_2 sphere in $\mathbb{R}^{|\mathcal{Z}|+1}$. We then fit the same family of splines used for $\mathcal{M}_h$ to the Hellinger-embedded centroids: a 1D cubic spline (natural or periodic) for the natural-domain tasks, and a thin-plate spline for the ICLR tasks. Analogous to the activation manifold, the fit passes exactly through every $\sqrt{b_i}$ and we do not apply a smoothing penalty.

The spline is fit in Euclidean space, but valid $\sqrt{b_i}$ points lie on a curved sphere. A naive fit to their ambient coordinates would leave the sphere between centroids, and an off-sphere vector does not square to a valid distribution. We therefore fit the spline in the *tangent plane* of the sphere at a base point b_* – a flat space that touches the sphere at b_* – and lift back to the sphere at decode time.

We take b_* to be the Euclidean mean of $\{\sqrt{b_i}\}$, re-normalized to unit length; because every $\sqrt{b_i}$ lies in the non-negative orthant, so does b_*. The *log-map* $t_i = \log_{b_*}(\sqrt{b_i})$ projects each centroid onto the tangent plane: it returns a vector whose direction points from b_* along the geodesic to $\sqrt{b_i}$ and whose length equals that geodesic distance. We fit the spline to the tangent vectors $\{t_i\}$. To decode at a query coordinate u, we evaluate the spline to obtain a tangent vector t and apply the *exponential map* $\exp_{b_*}(t)$, the inverse of the log-map: it walks distance $\|t\|$ along the geodesic on the sphere starting at b_* in direction t. The result is unit-norm by construction, so $\mathcal{M}_y$ stays on the sphere everywhere, and because $\exp_{b_*} \circ \log_{b_*}$ is the identity on the sphere, the decoded curve passes through every $\sqrt{b_i}$ exactly.

F.5 Geodesic Distances and the Isometry Test

To compute the geodesic distance $d_{\mathcal{M}_h}(c_i, c_j)$ between two concept centroids, we discretize the line segment between $\boldsymbol{s}^{-1}(c_i)$ and $\boldsymbol{s}^{-1}(c_j)$ in intrinsic coordinates into 150 equal sub-intervals, decode each waypoint through $\boldsymbol{s}$, and accumulate consecutive ambient distances. Each waypoint therefore lies on $\mathcal{M}_h$ by construction, and the resulting arc length is measured in the 64-dimensional PCA subspace in which the manifold lives, with the Euclidean norm as the ambient norm. For the behavior manifold $\mathcal{M}_y$ we follow the same procedure but compute distances in the full sqrt-probability ambient space $\mathbb{R}^{|\mathcal{Z}|+1}$ rather than a PCA-reduced subspace, using the Hellinger distance $d_H(p, q) = \frac{1}{\sqrt{2}}\|\sqrt{p} - \sqrt{q}\|_2$ directly on the sqrt-embedded waypoints.

The isometry score reported in §2 is the Pearson correlation between the upper-triangular entries of the resulting pairwise distance matrices. We augment the W centroid vertices with K interior points sampled at equally spaced fractions of the u-space geodesic between each centroid pair, decoded onto the manifold via $\boldsymbol{s}$ so that every vertex lives on $\mathcal{M}_h$ or $\mathcal{M}_y$. We choose K so that the vertex set is dense enough to probe the geometry between centroids: $K = 4$ for weekdays ($W = 7$); $K = 1$ for months ($W = 12$), alphabet ($W = 24$, letters C–Z), and the grid 5×5 task ($W = 25$); and $K = 0$ for age ($W = 91$, ages 10–100) and the cylinder 9×9 task ($W = 81$), whose centroids are already dense. We then correlate every off-diagonal pair in the full vertex set except those whose two vertices lie on a common centroid-pair geodesic, since those distances are sub-arcs of the same geodesic and would inflate the correlation by construction. To visualise the resulting pairwise structure, we embed each distance matrix into three dimensions via classical multidimensional scaling, as shown in Figs. 2, 8, and 5.

F.6 Steering Interventions

For each pair of concept values (z_a, z_b), we steer the model from the centroid c_a to the centroid c_b via a path of $K = 50$ waypoints. We use a fixed set of base prompts sampled randomly from the task's input distribution (the prompts' ground-truth results vary, and the same set is reused across all pairs): 16 prompts for the natural-domain tasks and 5 for the ICLR tasks. At each waypoint $\boldsymbol{\pi}(t)$, we intervene at the last-token residual-stream activation of the target layer and continue the forward pass to obtain $\boldsymbol{p}_{\boldsymbol{h} \leftarrow \boldsymbol{\pi}(t)}(x)$. Every reported behavioral trajectory is the pointwise mean over the base prompts. We use up to 50 randomly-sampled pairs per task. On the smaller-domain tasks where $W \cdot (W-1) < 50$, all pairs are used.

The two steering strategies differ both in how the waypoints $\boldsymbol{\pi}(t)$ are constructed and in what the intervention replaces (Eqs. 1, 2). For *manifold* steering, c_a, c_b live in intrinsic coordinates and the path is the manifold geodesic between them; at each waypoint we decode $\boldsymbol{\pi}(t)$ onto $\mathcal{M}_h$ in the 64-dimensional PCA subspace, lift it back to the residual-stream basis via the PCA inverse, and combine it with the prompt's unchanged off-subspace residual – so the steered activation differs from the base only in its top-64 PCA components. For *linear* steering, c_a, c_b are the raw activation centroids in the full residual stream, the path is the straight line between them, and the entire residual-stream activation is replaced by $\boldsymbol{\pi}(t)$ at each step.

F.7 Naturalness Metric

The cumulative output energy E_{BC} of §3.2 (Eq. 3) is computed as the sum of the Bhattacharyya distances $D_{\text{BC}}(\boldsymbol{\gamma}(t), \mathcal{M}_y) = -\log \sum_i \sqrt{\gamma_i(t)\, q_i(t)}$ between the induced output distribution at each of the $K = 50$ waypoints along steering paths and the closest point $q(t)$ on $\mathcal{M}_y$. We use the Bhattacharyya distance because $\mathcal{M}_y$ is fit in Hellinger geometry and the two are tightly related, $D_{\text{BC}} = -\log(1 - d_H^2)$, so D_{BC} stays inside the same geometry the manifold was constructed in. For each of the up to 50 sampled centroid pairs, we average the per-waypoint cumulative sum across the base prompts to obtain one scalar per pair, and report the mean and standard error of these per-pair scalars.

F.8 Pullback Optimization

The pullback procedure of §3.3 consists of two stages. First, we fix a behavioral target by evaluating the spline geodesic on $\mathcal{M}_y$ between the two behavior centroids b_a and b_b at $K = 20$ uniform fractions, yielding a sequence of target distributions $\hat{\boldsymbol{p}}_t \in \mathcal{M}_y$. Second, we optimize an activation-space path π_h^{pullback} which, when used to intervene at each waypoint, induces a behavioral trajectory matching $\hat{\boldsymbol{p}}_{0:K}$.

Path Parameterization. We parameterise π_h^{pullback} as a one-dimensional natural cubic spline through 10 control vectors at uniform t-positions, all of which are optimisation variables. The path is evaluated at the same $K = 20$ uniform fractions used to generate the target. Each control vector is restricted to the first 32 PCA components of the 64-dimensional subspace; the remaining 32 components, together with the orthogonal residual, are held at the base prompt's activation values during the intervention. We note that the linear and manifold-steering paths used as comparisons span the full 64-dimensional subspace, so the pullback optimization is operating within a strictly smaller search space.

Loss and optimizer. The loss at each waypoint t is the squared Hellinger distance $d_H^2(\boldsymbol{p}_{\boldsymbol{h} \leftarrow \boldsymbol{\pi}(t)}(x_n), \hat{\boldsymbol{p}}_t)$ between the induced output distribution and the target, averaged over 16 base prompts $\{x_n\}$ sampled freshly per pair, each conditioned on ground-truth z_a (in contrast to the steering setup of §F.6, where a fixed unfiltered set is reused across pairs). We minimize the sum of these per-waypoint losses with L-BFGS using strong-Wolfe line search, running 50 outer steps with up to 5 inner iterations each. The optimization is initialized by linearly interpolating between the two centroids in the 32-dimensional subspace and then sampling the resulting line at the 10 control-t positions. We stop early when the relative change in loss between two consecutive outer steps falls below 10^{-3}. We disable the path-norm regularizer for weekdays; on the other three natural-domain tasks we add a small regularizer that penalizes deviations of $\|\boldsymbol{\pi}(t)\|$ from the linear interpolation

between the endpoint centroid norms $|c_a|, |c_b|$. We use weight 10^{-3} for age and 5×10^{-4} for months and alphabet. This discourages the optimizer from drifting into a high-norm shortcut basin.

F.9 Pullback Recovery R^2

To compare the optimised pullback path π_h^{pullback} to the manifold-steering path π_h^* along $\mathcal{M}_h$, we project both into the SVD basis of π_h^* that captures at least 99% of π_h^*'s variance. In this basis we define the residual at each pullback waypoint as its orthogonal closest-point distance to π_h^*, and report

$$R^2 \;=\; 1 - \frac{\sum_t \|\pi_h^{\text{pullback}}(t) - \text{proj}_{\pi_h^*} \pi_h^{\text{pullback}}(t)\|^2}{\sum_t \|\pi_h^{\text{pullback}}(t) - \bar{\pi}_h^{\text{pullback}}\|^2}.$$

The linear baseline used in the same comparison is the straight chord between c_a and c_b in the 64-dimensional PCA subspace—not the linear-steering trajectory after intervention. As in §F.7, the values reported in §3.3 are mean $\pm$ standard error across the per-pair scalars, with p-values from paired t-tests against the linear baseline.

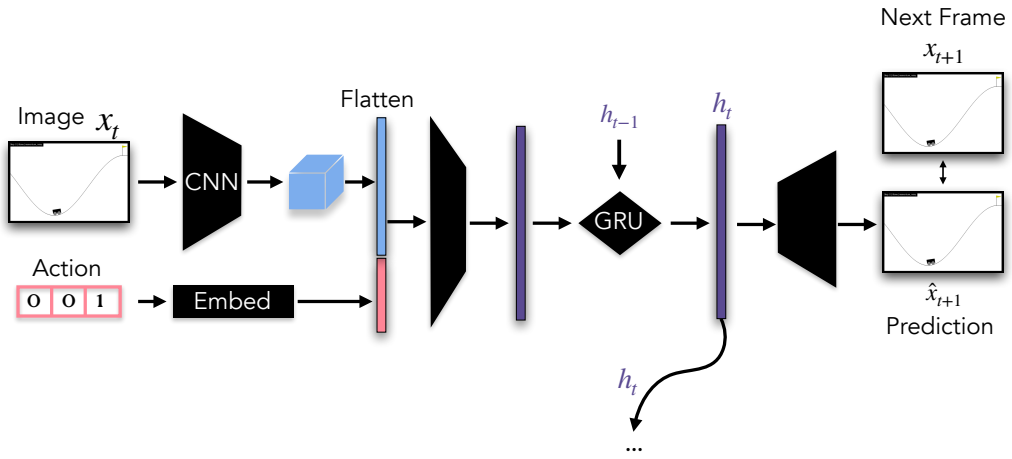

Figure 7: **Recurrent visual world-model architecture.** A convolutional encoder f_{enc} maps each $128 \times 128 \times 3$ frame x_t to a layer-normalized latent $z_t \in \mathbb{R}^n$ with $n = 64$. The discrete action $a_t \in \{0, 1, 2\}$ is mapped to a learned embedding $e(a_t) \in \mathbb{R}^{16}$, concatenated with z_t, and fed to a GRU together with the previous hidden state $\mathbf{h}_{t-1}$. A convolutional decoder f_{dec} produces a residual image from the resulting hidden state $\mathbf{h}_t$, yielding the next-frame prediction $\hat{x}_{t+1} = x_t + f_{\text{dec}}(\mathbf{h}_t)$, supervised against the ground-truth frame x_{t+1}.

G Experimental Details for the Vision Task

This section contains additional details on the experiment from §B.

G.1 Mountain Car.

Data collection. To recover the encoder's position manifold we harvest activations on 100 rollouts collected in `MountainCar-v0` (max 200 steps per episode) under a mixed stochastic policy chosen to give broad coverage of the position-velocity state space. At the start of each episode we sample one of two policies: with probability 0.7, a *noisy momentum* policy that pushes in the direction of the current velocity but, at each step, replaces the action with a uniform random action with probability 0.4; with probability 0.3, an *oscillating square-wave* policy that alternates between full-left and full-right thrust on a fixed period sampled uniformly from $\{5, \ldots, 25\}$ steps. We then pass each rendered frame through the trained encoder, label the resulting activation with the underlying ground-truth position, and fit the manifold to this collection of position-labelled activations.

Manifold fitting. To parameterize the activation manifold, $\mathcal{M}$, we partition the position range into $B = 100$ bins, compute the mean encoder output per occupied bin, and fit a smoothing spline $\gamma_{\mathcal{M}} : [0, 1] \to \mathcal{M}$ through these means (one univariate spline per coordinate, weighted by the square root of bin counts to regularize sparse regions). We additionally verify via linear probing that the encoder representations z_t encode the ground-truth physics: a Ridge regression probe recovers position with $R^2 \approx 0.95$ and velocity with $R^2 \approx 0.90$. A three-component PCA of the encoder outputs reveals the spline $\gamma_{\mathcal{M}}$ as a curve that closely tracks the data manifold, while the chord ℓ between the same endpoints cuts through its interior.

To parameterize the behavior manifold, $\mathcal{M}_y$, discretize $\mathcal{Z}$ into B bins with centers $\{\mu_b\}_{b=1}^{B} \subset \mathbb{R}^n$ obtained by evaluating the activation-manifold spline at B evenly spaced positions, $\mu_b = \gamma_{\mathcal{M}}(p_b)$. The mapping to behavior is

$$F(z) \;=\; \operatorname{softmax}\!\left(-\frac{\|z - \mu_b\|_2}{\tau}\right)_{b=1}^{B} \;\in\; \Delta^{B-1}, \tag{10}$$

with temperature $\tau = 0.5$. F is a smooth, deterministic map from activations to position distributions. We use $B = 128$, which makes F's Jacobian full column-rank, ensuring the inverse problem has a locally unique solution. For each bin i, the natural centroid on the behavior manifold is the model's

average output distribution conditioned on samples in that bin:

$$b_i \;=\; \mathbb{E}[F(z) \mid \mathrm{bin}(z) = i] \;\in\; \Delta^{B-1}. \tag{11}$$

Because the bin grid is dense relative to the data manifold's intrinsic dimension, b_i is well-defined for all bins. We embed each b_i in Hellinger coordinates $h_i = \sqrt{b_i}$ on the unit sphere of $\mathbb{R}^B$ and fit a 1D smoothing spline $\gamma_{\mathcal{M}_y} : \mathcal{Z} \to \mathbb{R}^B$ through $\{h_i\}$ parameterized by position. This is the behavior manifold $\mathcal{M}_y$.

H Additional Results

H.1 Natural Domain Arithmetic: Ages and letters

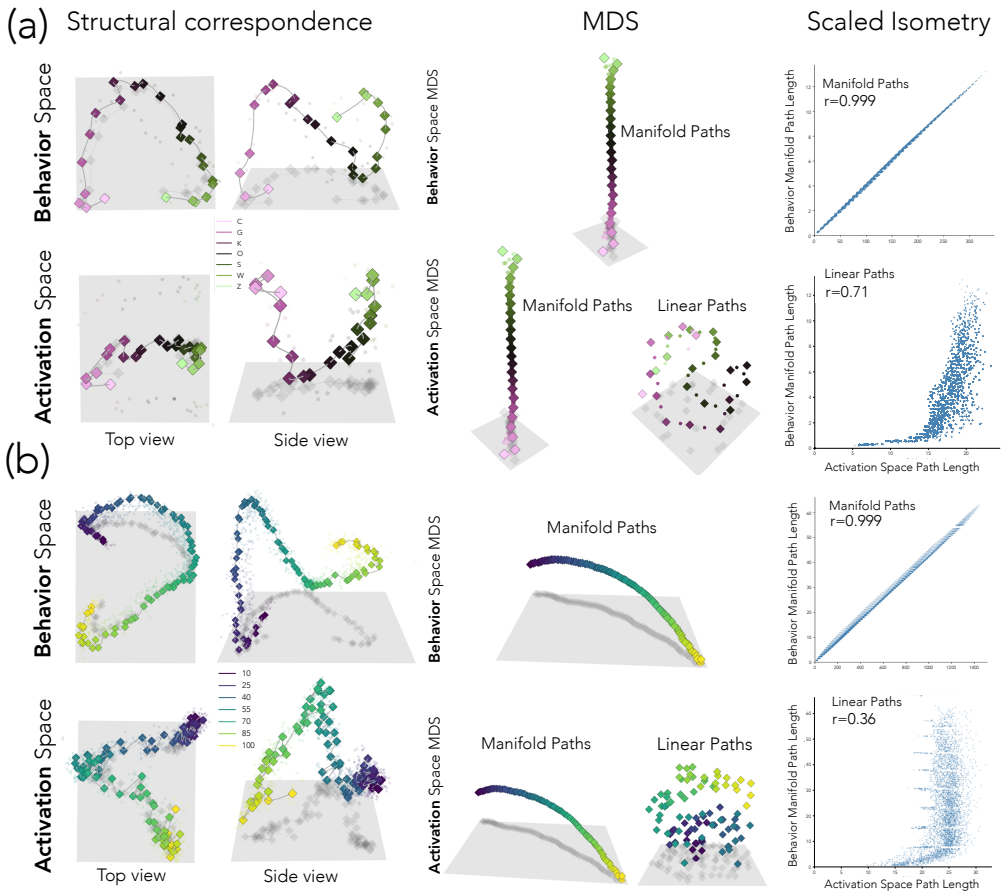

Figure 8: **Approximate isometry between activation and behavior manifolds for sequential concepts**. Manifolds (cubic splines) fit to activation and behavior (i.e., output distributions over concept tokens) spaces of Llama 3.1 8B. The letters **(a)** and ages **(b)** tasks consist of simple addition questions such as: `What letter comes four letters after M?`. Both activation and behavior manifolds show sequential structure (PCA visualization shown in left column). Furthermore, on-manifold distances in activation space show strong correlation with on-manifold distances in behavior space (right column), as well as a clear match via a multidimensional scaling (MDS) embedding (middle column). In contrast, linear distances in activation space show weaker correlations and warped or incoherent structure. These results demonstrate an approximate isometry between the activation and behavior space manifolds.

H.2 In-Context Learning of Representations.

In addition to results provided in the main text, we test a 9×9 cylinder in the ICLR domains, and find that despite the added complexity of a periodic dimension and substantially more graph nodes, when Llama 3.1 8B is provided sufficient context (2048 tokens in this case), it reaches above 80% neighborhood accuracy (probability mass on valid neighbors). We fit a manifold and steer along this domain, finding that the result of factored control generalizes beyond the graph domain.

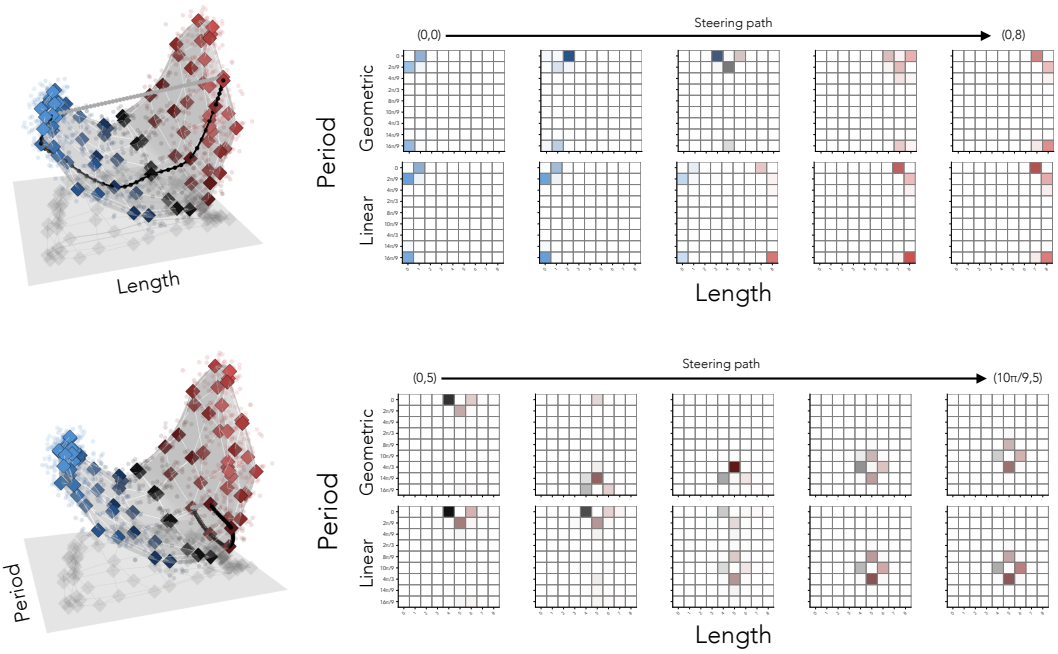

Figure 9: Results for in-context learning of representations on a 9×9 cylinder domain. We find that, as in the grid domain, manifold steering achieves factored control: coherent steering of independent dimensions, while linear steering once again shows 'teleportation' behavior.

H.3 Manifold steering allows manipulation of uncertainty without loss of structure.

To examine multi-dimensional concepts in known domains, we partition weekday addition centroids by addition value (1–5, 6–10, 11–15, 16–20), revealing concentric circles along a second manifold dimension forming a cylinder-like structure (Fig. 11). Manifold steering along the circular dimension maintains ordered weekday transitions with increasing entropy per group. This suggests manifold geometry can serve as a handle for calibrating model confidence in a controlled fashion. The experiment was conducted with Llama 3.1 70B layer 70.

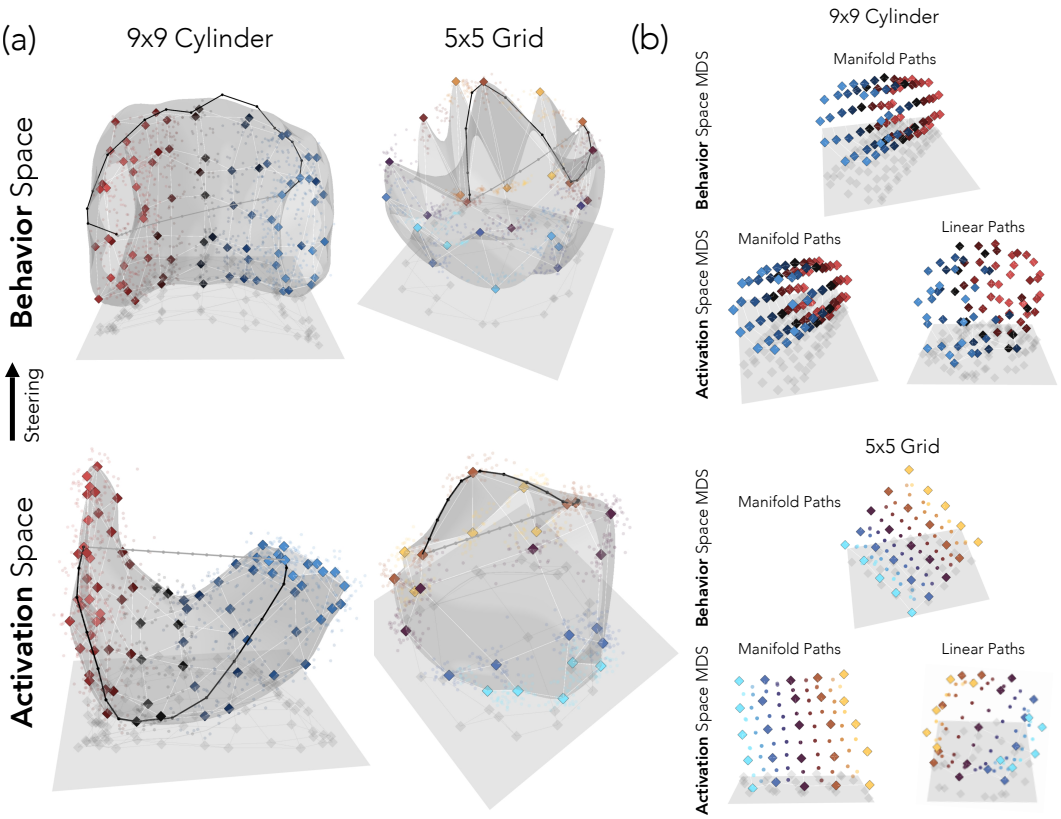

Figure 10: **(a)** Activation and behavior space paths for the 5×5 Grid task and 9×9 Cylinder. Similarly to the addition tasks with known concepts, we find that the manifold steering paths closely follow the behavior manifold $\mathcal{M}_y$. **(b)** Multidimensional scaling (MDS) embedding for linear and manifold distances in activation space and manifold distances in behavior space. As with the addition tasks with known concepts, manifold distances in activation space show a clear structural match to behavior space, whereas linear distances warp the structure.

H.4 Mountain Car

Structural Correspondence between $\mathcal{M}_h$ and $\mathcal{M}_y$. If $\mathcal{M}_h$ encodes the model's predictive distributions over $\mathcal{Z}$, then $\mathcal{M}_h$ and $\mathcal{M}_y$ should be approximately isometric—distances along one manifold should correlate with distances along the other. We test this by sampling $W = 50$ anchor positions inside the shared parameter range and computing pairwise arc lengths along each manifold:

$$d_{\mathcal{M}}(p_i, p_j) = \int_{p_i}^{p_j} \|\gamma'_{\mathcal{M}_h}(p)\|_2 \, dp, \quad d_{\mathcal{M}_y}(p_i, p_j) = \frac{1}{\sqrt{2}} \int_{p_i}^{p_j} \|\gamma'_{\mathcal{M}_y}(p)\|_2 \, dp, \tag{12}$$

where the $1/\sqrt{2}$ on the behavior side converts the Euclidean integral in Hellinger ambient space to Hellinger units. The Pearson correlation between $\{d_{\mathcal{M}_h}(p_i, p_j)\}$ and $\{d_{\mathcal{M}_y}(p_i, p_j)\}$ over all $\binom{50}{2} = 1225$ pairs is $r = \mathbf{0.996}$; the chord distances used by linear steering correlate far less ($r = 0.06$ between activation chord and behavior arc length), since chords cut across the encoder loop and are structurally divorced from the encoded conceptual geometry.

Pullback: Behavior space steering. Having established the bottom-up direction in Fig. 6, i.e., paths along $\mathcal{M}_h$ produce behavior trajectories on $\mathcal{M}_y$, We now test the top-down direction: starting from a behaviorally-natural trajectory in $\mathcal{M}_y$, do we naturally recover an activation path that traces $\mathcal{M}_h$? For each endpoint pair (p_a, p_b) we construct the conformal behavior target $\hat{\gamma}_\alpha$ via the procedure of §3.3 (geodesic on the simplex under cost $c(p) = \exp(\alpha \cdot d_H(p, \mathcal{M}_y))$). We then optimize an

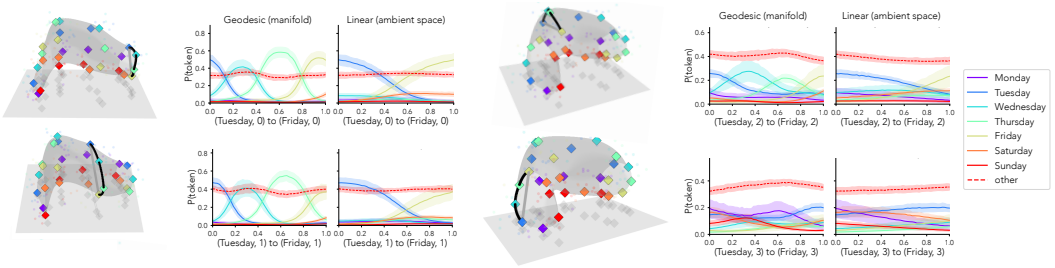

Figure 11: **Inducing greater uncertainty over a conceptual space and steering along it.** By increasing the addition value, we induce greater uncertainty in the model with respect to the right answer. Instead of grouping across addition values (as we do in Fig. 3), we visualize centroids by addition value, and find that these groups yield a series of circles organized into a curved cylinder-like shape. Manifold steering along the circle in the first three groups maintains ordered transitions, yet with increasing entropy in each group.

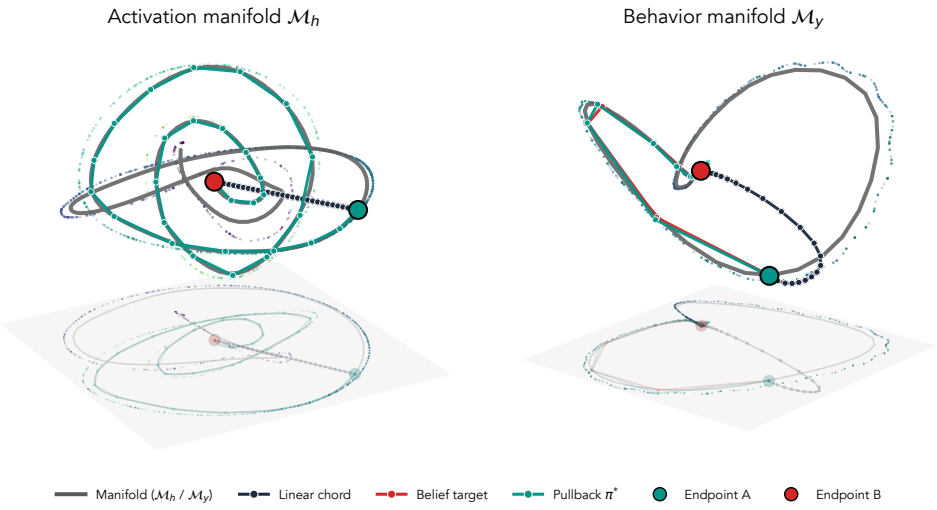

Figure 12: **Pullback from $\mathcal{M}_y$ recovers $\mathcal{M}_h$ in the visual world model. Left (Activation Space):** PCA visualization of the encoder representations, showing the geometric path along $\mathcal{M}_h$, the linear chord, and the pullback-optimized path $\pi^\star$ between endpoints p_A and p_B. Although initialized at the chord, $\pi^\star$ converges onto $\mathcal{M}_h$, closing the spiral loop traced by the encoder geometry and becoming nearly indistinguishable from the activation reference. **Right (Behavior Space):** PCA visualization of the corresponding trajectories pushed through the operator F (Eq. 10), shown in Hellinger coordinates. The conformal target $\hat{\gamma}_\alpha$ tracks the behavior manifold $\mathcal{M}_y$, and the pushforward $F(\pi^\star)$ closely matches it, while the pushforward of the linear chord $F(\ell)$ departs sharply, cutting across the simplex interior rather than following $\mathcal{M}_y$. Together, the two panels show that optimizing an activation path to match a behavior-manifold target recovers $\mathcal{M}_h$ top-down: matching behavior along $\mathcal{M}_y$ is sufficient to pull activations back onto $\mathcal{M}_h$, mirroring the pullback result from the language-model experiments.

activation path $\pi_\alpha = (v_0, \ldots, v_K)$ in $\mathbb{R}^n$ to minimize

$$L(\pi) \;=\; \sum_{t=0}^{K} \left\| \sqrt{F(v_t)} - \sqrt{\hat{\gamma}_\alpha(t)} \right\|_2^2. \tag{13}$$

Following the language-model setup, all $K+1$ waypoints (including endpoints) are free parameters, initialized at the linear chord and optimized jointly via L-BFGS with strong-Wolfe line search. We use $K = 30$ waypoints and run independent optimizations for each of 30 endpoint pairs.

Across all 30 endpoint pairs, the pullback paths π_α closely trace $\mathcal{M}_h$ (Fig. 12): The mean Euclidean distance from π to $\mathcal{M}_h$, averaged over waypoints and pairs:

$$\text{linear chord}\colon\ 2.22, \qquad \text{geometric } (\mathcal{M}_h)\colon\ 0.20, \qquad \text{pullback}\colon\ 0.29.$$

The pullback path is at $\mathbf{95.4}\%$ of the chord-to-geometric recovery and dominates the chord baseline on $30/30$ pairs. The aggregate degradation is concentrated on pairs with one endpoint at the extreme wall position ($p \approx -1.2$), where the encoder geometry has tighter curvature; on the remaining $\sim$20 pairs, π_∞ is essentially indistinguishable from $\mathcal{M}_h$ itself. The α-sweep traces the same family of trajectories observed in the language-model experiments: at $\alpha = 0$ the conformal target is the unrestricted Hellinger geodesic on the simplex, and the recovered π_0 leaves $\mathcal{M}_h$ in order to match this off-manifold target; as α grows the target is pushed onto $\mathcal{M}_y$ and the recovered π_α correspondingly tracks $\mathcal{M}_h$.