STORIES IN SPACE: IN-CONTEXT LEARNING TRAJECTORIES IN CONCEPTUAL BELIEF SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) update their behavior in context, which can be viewed as a form of Bayesian inference. However, the structure of the latent hypothesis space over which this inference operates remains unclear. In this work, we propose that LLMs assign beliefs over a low-dimensional geometric space—a *conceptual belief space*—and that in-context learning corresponds to a trajectory through this space as beliefs are updated over time. Using story understanding as a natural setting for dynamic belief updating, we combine behavioral and representational analyses to study these trajectories. We find that (1) belief updates are well-described as trajectories on low-dimensional, structured manifolds; (2) this structure is reflected consistently in both model behavior and internal representations and can be decoded with simple linear probes to predict behavior; and (3) interventions on these representations causally steer belief trajectories, with effects that can be predicted from the geometry of the conceptual space. Together, our results provide a geometric account of belief dynamics in LLMs, grounding Bayesian interpretations of in-context learning in structured conceptual representations.

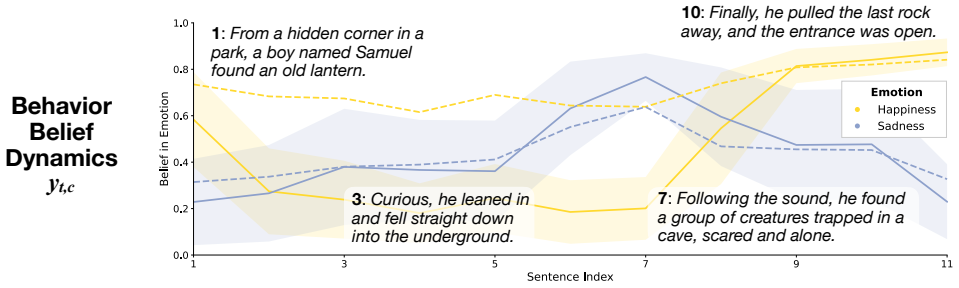
1 INTRODUCTION

Large Language Models (LLMs) show impressive abilities to understand language and adapt their behavior according to new information in context (Brown et al., 2020; Lampinen et al., 2024; Agarwal et al., 2024; Park et al., 2024; Chan et al., 2022; Min et al., 2022; Park et al., 2025; Anil et al., 2024). As an LLM reads a text or engages in a conversation, it must continually update its beliefs about the world, the user, and the current topic being discussed as new information comes to light (Hosseini et al., 2026; Geng et al., 2025). A pressing question for AI interpretability then is to understand how LLMs represent complex concepts, and how they dynamically update beliefs in these concepts as they process text (Ruis et al., 2023; Hu et al., 2025b; Prakash et al., 2025; Lubana et al., 2025).

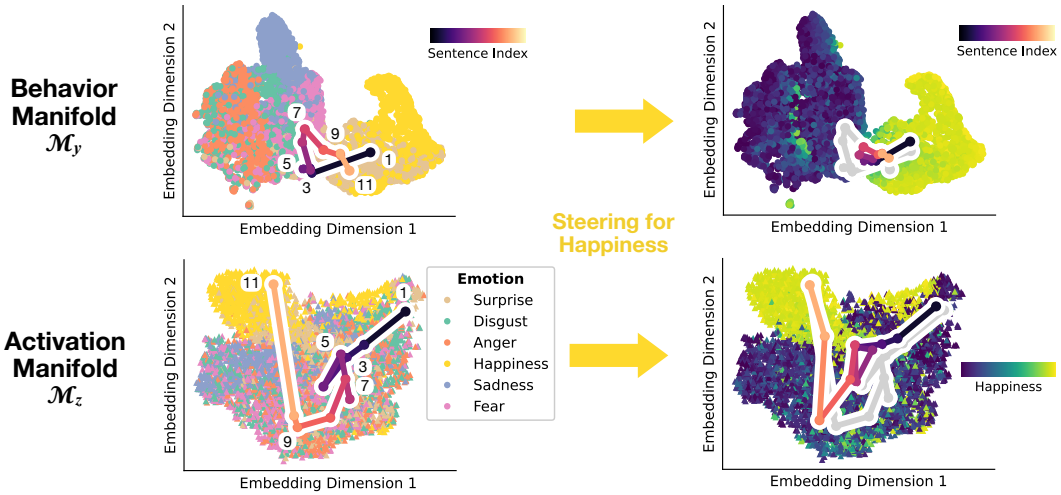
This dynamic belief updating operates via In-Context Learning (ICL), as LLMs adapt their behavior based on input data, with no changes to model weights. ICL is productively framed as Bayesian inference, where an LLM re-weights latent concepts according to their posterior distributions given input data (Xie et al., 2021; Bigelow et al., 2023; Arora et al., 2024; Panwar et al., 2024; Zhang et al., 2023). Prior work on ICL has primarily focused on few-shot learning scenarios, where data is a list of input–output examples, and where learning dynamics are relatively steady and monotonic (Brown et al., 2020; Wurgaft et al., 2025; Bigelow et al., 2025). However, when considering LLM behavior in freeform text generation and understanding, ICL dynamics can be much more non-linear and idiosyncratic (Bigelow et al., 2024). The goal of this work is to extend the Bayesian framework to account for these non-linear ICL dynamics. In order to accomplish this goal, however, we must consider the structure of latent concepts that an LLM might represent.

Since understanding LLMs has much in common with the study of the human mind (Bigelow, 2026; Hu et al., 2025a; Hagendorff et al., 2023), we draw on cognitive science for theories of conceptual representation in LLMs. In particular, Bayesian theories of cognition assume that a learner assigns beliefs over latent variables (i.e., *concepts*) which represent a structured generative process that predicts observed data (Tenenbaum et al., 2011; Goodman et al., 2008; Piantadosi, 2021). However, these theories cannot be easily translated to hypotheses about how distributed neural systems implement computations. To draw connections between patterns in behavior and neural representations, we use the theory of *conceptual spaces* in cogntion, which offers a geometric

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) **Behavior belief dynamics** As an LLM reads a story $x_{1:T}$, after each sentence t we elicit its beliefs $y_{t,c}$ about a concept c through prompting at that point. These belief dynamics $y_{t,c}$ are shown as a timeseries, with the shaded region showing the standard deviation of $y_{t,c}$ and the dotted lines showing the effect of steering for the concept *happiness*.



(b) **Low-dimensional manifolds** We use dimensionality reduction (UMAP) to identify low-dimensional manifolds in model behavior $\mathcal{M}_y$ and activations $\mathcal{M}_z$. These manifolds are structured, with parts shared across behavior and representation, and in-context belief updates following smooth paths through the manifold. (Right) The colored path shows the model's trajectory after steering, with the un-steered trajectory shown in grey.

Figure 1: **Conceptual belief trajectories** Model beliefs qualitatively follow reasonable patterns for a story, for example, here we see that *happiness* drops when the protagonist falls into a hole and then *sadness* increases when he discovers a group of scared creatures trapped further down, until these reverse at end of the story when the protagonist rescues the creatures (full story text in App. A). The story's trajectory in (a) behavior space $\mathcal{M}_y$ and (b) activation space $\mathcal{M}_z$ has a similar interpretation, beginning in a positive region, moving towards a negative area, before returning for a happy ending. When we steer for a particular concept c, in this case *happiness*, we alter the trajectory that the model's beliefs follow.

perspective on concepts as low-dimensional subspaces with well-defined distance metrics (Gärdenfors, 2000). Further, we extend this framework to account for notions of uncertainty, or *beliefs*, that a learner might hold over a conceptual space (Strößner, 2022), and which are essential for Bayesian inference.

In real user interactions, LLMs often process extended, open-ended text where beliefs must be updated as new information is processed. We use story understanding as a natural setting for studying this kind of belief updating in freeform text, since stories require readers to track events, themes, writer intent, and other information that evolves over time. Story understanding requires many kinds of knowledge representations and reasoning (Schubert & Hwang, 2000; Winston, 2011; Zwaan et al., 1995; Zwaan & Radvansky, 1998; Zacks & Tversky, 2001), which are dynamically updated in the mind of the reader as they proceed through the story. We propose that story understanding, and in-context learning more generally, can be viewed as a trajectory through a conceptual belief

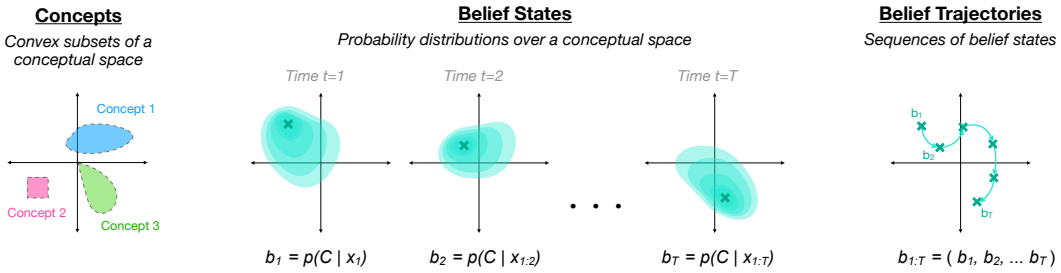


Figure 2: **Theory of belief update** We extend the conceptual space framework of Gärdenfors (2000) to account for belief and uncertainty; axes represent two dimensions in a conceptual space. (Left) Concepts are defined as convex sub-spaces, (Middle) Belief States are probability distributions over a conceptual space, and (Right) Belief Trajectories are sequences of belief states, e.g., in the mind of a reader as they progress through a story (Yeh et al., 2025).

space. As a reader acquires knowledge, makes inferences, and predicts each new step of a story, they navigate through a conceptual belief space.

In this work, we find evidence of exactly such a conceptual belief space in LLMs. We show that (i) LLM behavior and belief updating can be accurately explained by a low-dimensional structured space, and that in-context learning follows smooth trajectories across this space; (ii) we can predict behavior in this space by linearly probing hidden representations; (iii) we can steer LLMs to change their learning trajectories by intervening on hidden representations; and (iv) the geometry of these low-dimensional spaces predicts steering entanglement.

Overall, our results suggest that, at least in simple cases, LLMs can be shown to track rigorous definitions of belief over time while processing arbitrary freeform text. The Bayesian interpretation of ICL can explain this, although it is more difficult to directly model ICL dynamics since evidence accumulation varies enormously depending on what particular text is used as input data. Instead, we can effectively understand ICL dynamics on freeform text by studying latent structure in conceptual representations which are relevant to the text's content.

2 In-Context Learning Trajectories in Conceptual Belief Space

In-context learning is productively viewed as Bayesian inference (Xie et al., 2021; Wurgaft et al., 2025), where output behavior y is produced by marginalizing over a hypothesis space of latent concepts c, each weighted according to how well it matches in-context data x:

$$p(y \mid x) = \int_c p(y \mid c, x)\, p(c \mid x)$$

However, this framing does not specify what kinds of concepts are represented, and cannot explain a hallmark property of concepts: they are structured in terms of similarity. As such, we enrich this framing so that the latent concepts belong to a multi-dimensional conceptual space with a metric capturing shared geometry. Furthermore, we understand ICL belief dynamics as trajectories through a lifted *conceptual belief space* where elements are distributions over the original conceptual space.

Conceptual Belief Spaces As in Gärdenfors (2000), we define a *conceptual space* as a set of related dimensions and an associated metric, and *concepts* as regions of this low-dimensional geometric space (Fig. 2). More formally, a conceptual space $\mathcal{C}$ is a metric space over attributes:

$$\mathcal{C} = A_1 \times A_2 \times \ldots \times A_n \tag{1}$$

Attributes are independent dimensions, on which individual objects and concepts vary. The structure of this space is defined by a distance metric $d_{\mathcal{C}}$, which measures the similarity of points in $\mathcal{C}$. Attributes in $\mathcal{C}$ can be considered independently of others, for example an object's weight might be determined without considering its temperature and color.

To move from concepts to beliefs over concepts, we lift to a *conceptual belief space* $\mathcal{B}$, which is a metric space of probability distributions over $\mathcal{C}$ [1]:

$$\mathcal{B} = P(A_1) \times P(A_2) \times \ldots \times P(A_n) \tag{2}$$

The dimensions of conceptual belief space are uncertainties over the ground attributes and the metric $d_{\mathcal{B}}$ can be entirely new or a lift of the ground metric $d_{\mathcal{C}}$.

Emotional Belief Space In this paper, we consider a conceptual space over *Emotions*, with the concepts *happiness*, *sadness*, *anger*, *surprise*, *fear*, and *disgust*[2]. As a reader comprehends each new sentence in a story, such as the story in Fig. 1, they will update their beliefs $p(c \mid x)$ according to whatever new data x they process. For example, when the protagonist endures hardships and conflict, a reader might say that a story seems more *sad*, and then when they return victorious from a great journey, the story might seem *happy*. We will behaviorally operationalize this by simply asking a reader (in our case, an LLM) about a particular concept c after each sentence *"How (happy/sad) is this story, on a scale of 0-10?"* of the story and recording their behaviors as $p(y_{\text{sad}} \mid x)$ and $p(y_{\text{happy}} \mid x)$. This process of dynamic change in emotional tone can be visualized as a line plot for each emotion, shown in Fig. 1, which tracks the ups and downs of the reader's beliefs during each twist and turn of the story (Vonnegut, 1995).

We predict that the *Emotions* and *Genres* domains will have structured representations in conceptual belief space $\mathcal{B}$ according to relationships between different concepts that occupy the space. One leading theory in human psychology proposes that all emotions fall along the two related dimensions of *Valence* and *Arousal* (Russell & Mehrabian, 1977; Russell, 1980). This theory would lead us to predict that if we asked people to judge various stimuli according to emotions such as *happiness*, *sadness*, or *anger*, then both their behavior and neural representations would follow certain patterns in accordance with their distances $d(c, c')$ in valence-arousal space (Balkenius & Gärdenfors, 2016).

Trajectories through Conceptual Belief Space We define *belief states* b_t that an LLM might have at a particular point in a story t as a probability distribution over a conceptual space $\mathcal{C}$, i.e., $b_t = p(\mathcal{C} \mid x_{1:t}) = p(A_1 \ldots A_n \mid x_{1:t})$ where $x_{1:t}$ refers to a story text x from its beginning up until point t. These states occupy a wider *conceptual belief space*. Finally, a *belief trajectory* $b_{1:T}$ is a path through conceptual belief space, i.e., a sequence of belief states for each time t throughout a story $x_{1:T}$: $b_{1:T} = \big(b_1, b_2, \ldots b_T\big)$.

This framework will enable us to talk formally about how a reader incrementally updates their beliefs while processing a story, and about the structure of latent concepts that they use to assign beliefs. Concretely, we predict that LLM behavior and representations can be described according to positions in this lower-dimensional conceptual belief space, with minimal loss of information. An advantage of this framework is that psychological spaces may be mapped onto dimensions of the underlying neural substrate, for example, in humans, emotions might map to different quantities of neurotransmitters such as dopamine and seratonin (Wang et al., 2020), and colors can map to rods and cones in the retina (Brown & Wald, 1964). However, in many cases we may not have a theory a priori of what underlying dimensions define a particular conceptual space. Instead, we can use dimensionality reduction techniques to infer the latent dimensions given behavioral or neural data.

3 METHODS

We next provide a set of methods for studying belief update dynamics in LLMs and the low-dimensional conceptual spaces which underlie these beliefs. For the following methods, we will be studying how an LLM updates its beliefs as it reads each successive sentence x_t of a text $x_{1:T}$, where t indexes sentences in the story. We will collect two kinds of data: behavioral judgments $p(y_c \mid x_{1:t}, q_c)$ by the LLM which quantify how well a given concept (e.g. *happiness* or *sadness*) applies to $x_{1:t}$, and residual activations for the story text $z_\ell(x_{1:t})$ for layer ℓ at the final token of $x_{1:t}$. For behavioral judgments, we prompt the LLM to rate how well a *concept* applies to $x_{1:t}$. We do not assume that we know a priori the underlying conceptual dimensions $\{A_1, \ldots, A_n\}$. Instead, we assume that we know a set of related concepts (e.g. different *Emotions* or *Genres*) which should fall under some domain $\mathcal{D}$.

[1] See our sister paper Wurgaft et al. (2026) for a more rigorous analysis of this framing

[2] We note the domain we consider for concepts studied in this paper is fairly limited. For example, emotions have been argued to have an intricate, hierarchical organization that results in 135 nodes tree (Shaver et al., 1987)

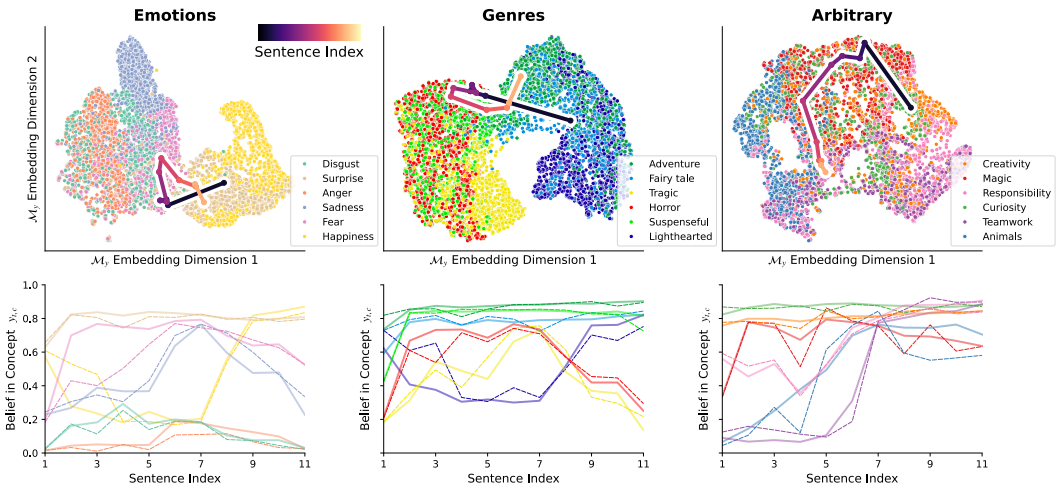


Figure 3: **Three conceptual domains** (Top) Belief manifolds $\mathcal{M}_y$ across three domains, with the trajectory plotted for the same story as in Fig. 1. For both *Emotions* and *Genres*, the more "positive" concepts (in terms of valence) are on one end of the manifold, with negative emotions at the opposite end, and the more neutral concepts (*surprise* and *adventure*) forming a bridge between the two. On the other hand, $\mathcal{M}_y$ for *Arbitrary* does not have such clear structure. (Bottom) Observed model beliefs $y_{t,c}$ (solid line) plotted against probe predictions $\widehat{y}_{t,c}$ at layer $\ell = 9$. We find that linear probes are highly predictive of model behavior, suggesting that belief is encoded linearly as a model processes story text.

Dimensionality Reduction We predict that behavioral data y and representational data z can be described according to movement along low-dimensional manifold $\mathcal{M}_y$ and $\mathcal{M}_z$. To approximate these manifolds, we use dimensionality reduction techniques—UMAP (McInnes et al., 2018) and PCA—on our datasets $Y \in \mathbb{R}^{N \times k}$ and $Z_\ell \in \mathbb{R}^{N \times q}$. Important information should be preserved after projecting the LLM's beliefs into the lower-dimension spaces $\mathcal{M}_y, \mathcal{M}_z$. To characterize the outer edges of the model's belief manifold, we fit dimensionality reduction to a set of the most extreme examples of each concept, and project data points onto this space. We filter our training dataset to the sentences across any story which maximally activate each concept, for example, in the *Emotions* domain we take the *"happiest"* points in any story, the *"saddest"* points, and so on. We will also empirically test for $\mathcal{M}_y$ and $\mathcal{M}_z$ will be jointly predictive of both model behavior and internal representations, and that structure in effects of activation steering. Both $\mathcal{M}_y$ and $\mathcal{M}_z$ can be seen as images of an underlying conceptual belief space $\mathcal{B}$. More broadly, since $y_{t,c}$ is an overcomplete representation that lies on or near the manifold $\mathcal{M}_y$, $y_{1:T}$ can be visualized by projecting it into the lower-dimensional space $\mathbb{R}^{T \times d}$.

We make three concrete predictions about $\mathcal{M}$: (1) belief in the *Emotions* and *Genres* domains can be well-described by coordinates in $\mathcal{M}_y$ and $\mathcal{M}_z$; manifolds for the *Emotions* and *Genres* domains will be structured according to the relationships of different concepts c, for example emotions may have hierarchical (Plutchik, 1980) or geometric (Russell, 1980) structure as has been shown for humans; (3) behavior $\mathcal{M}_y$ and activation $\mathcal{M}_z$ manifolds will be similar in structure.

Distance Matrices In order to empirically test these hypotheses, we will use distance matrices (or, equivalently, similarity matrices) with an L_2 metric to quantify structure and similarity of $\mathcal{M}_y$ and $\mathcal{M}_z$, somewhat similar to RSA (Kriegeskorte et al., 2008). Since each concept refers to a region in $\mathcal{M}$, we will use the centroid for a concept in the embedding space to approximate it's geometry as a point. More specifically, for each concept c we will compute the mean coordinates over a set of max-activating examples $x_{1:t}$, i.e., $\frac{1}{N}\sum_n \mathcal{M}(x^{(n)}_{1:t})$. In simple terms, for a two-dimensional space $\mathcal{M}$ will be represented merely with a single coordinate pair. Each point in a distance matrix for a manifold $\mathcal{M}$ will measure the distance $d_{\mathcal{M}}(c, c')$ between the centroids of two concept c and c'.

After constructing distance matrices, we will perform two kinds of analyses to measure the structure and similarity between two geometric spaces. First, we will use agglomerative clustering (Ward Jr,

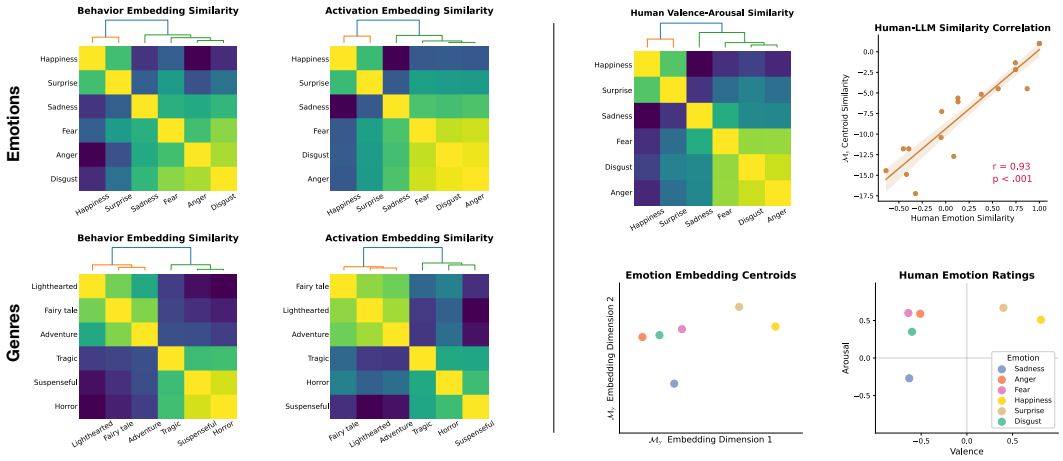


Figure 4: **Distance reveals domain structure** (Left) Distance matrices for pairs of concepts $d_{\mathcal{M}}(c, c')$ according to behavior $\mathcal{M}_y$ and activation $\mathcal{M}_z$ manifolds, computed between concept centroids. For *Emotions* and *Genres*, we find common hierarchical structure shared across $\mathcal{M}_y$ and $\mathcal{M}_z$, and that concept centroid distances $d_{\mathcal{M}}(c, c')$ are highly correlated between $\mathcal{M}_y$ and $\mathcal{M}_z$ ($r = .92, p < .001$ for *Emotions*, $r = .89, p < .001$ for *Genres*). Results for *Arbitrary* shown in App. D. (Right) The geometry of $\mathcal{M}_y$ for the *Emotions* domain is strikingly similar to the geometry of valence-arousal emotion space estimated from human data (Russell & Mehrabian, 1977).

1963) to infer hierarchical structures in distance matrices. Second, we will measure correlation between distance matrices, by comparing $d(c, c')$ across distance functions d.

Belief Elicitation Behavioral observations offer a lens into how an LLM updates its beliefs dynamically over the course of a story. To operationalize this, we input a story text into an LLM up until some time t, i.e., $x_{1:t}$, and then append a query q_c to the LLM's input text which asks the LLM to rate on a scale from 0 to 10 how much a concept c applies to the text $x_{1:t}$. We prompt models to respond with an integer in the range 0 to 10, $i \in [0, 1, \dots 10]$ (full prompts listed in App. B), and we take their estimate $y_{t,c}$ as a weighted average of token probabilities for each possible integer value:

$$y_{t,c} = \mathbb{E}[y = i_c] = \frac{1}{10} \sum_{i=0}^{10} i \cdot p(y = i \mid x_{1:t}, q_c) \tag{3}$$

Each $y_{t,c}$ can be considered as a behavioral estimate of the model's belief in concept c at time t. We repeat this procedure for each queried concept q_c, and the behavioral belief state y_t is comprised of all queried concepts: $y_t = \{y_{t,c_1}, y_{t,c_2}, \dots y_{t,c_k}\}$. By concatenating this sequence of belief probabilites $y_t = p(a \mid x_{1:t})$ for each timestep, we form a belief trajectory $y_{1:T} = (y_1, y_2, \dots y_T)$ across the entire span T of a story. Iterating across a dataset of stories, we collect behavioral data $Y \in \mathbb{R}^{N \times k}$ where $k = 6$ in our experiments.

Linear Probes Next, we use probes (Alain & Bengio, 2017; Hewitt & Liang, 2019; Conneau et al., 2018; Tenney et al., 2019a; Belinkov et al., 2017; Belinkov, 2022; Tenney et al., 2019b) to predict an LLM's output behavior $y_{t,c}$ from its hidden activations $z_{t,\ell}$. To collect activations, we input the text of a story $x_{1:t}$ from the beginning until sentence t, and collect residual activations $z_{\ell,t}$ at each layer ℓ for the final token. Only the raw content of the story $x_{1:t}$ is presented as input to the LLM, without any specific concepts mentioned. Given dataset of activations $Z_\ell \in \mathbb{R}^{N \times q}$ (where $q = 4096$ in our case) across stories, we fit linear probes to predict behavioral beliefs $y_{t,c}$: $\widehat{y}_{t,a} = f_\theta(z_{t,\ell})$, using linear regression $f_\theta(z) = \theta^\top z + \theta_0$ with L_2 sparsity penalty. Lastly, we fit isotonic regressions to calibrate probes and map $f_\theta(v)$ onto the scale $[0, 1]$ (App. H) (Niculescu-Mizil & Caruana, 2005). Probes are trained and calibrated on a larger training dataset and tested on a held out test set.

Activation Steering We next aim to verify that the representations we identify are causally involved in determining behavior. To do so, we intervene on latent activations in an LLM to influence its beliefs. The class of methods we will use is activation addition steering, in other words, methods where a steering vector is constructed for a given layer and concept $v_{c,\ell}$. Activations are steered at a specific layer by intervening on the LLM during its forward pass and patching activations $z_\ell(x_{1:t})$ to

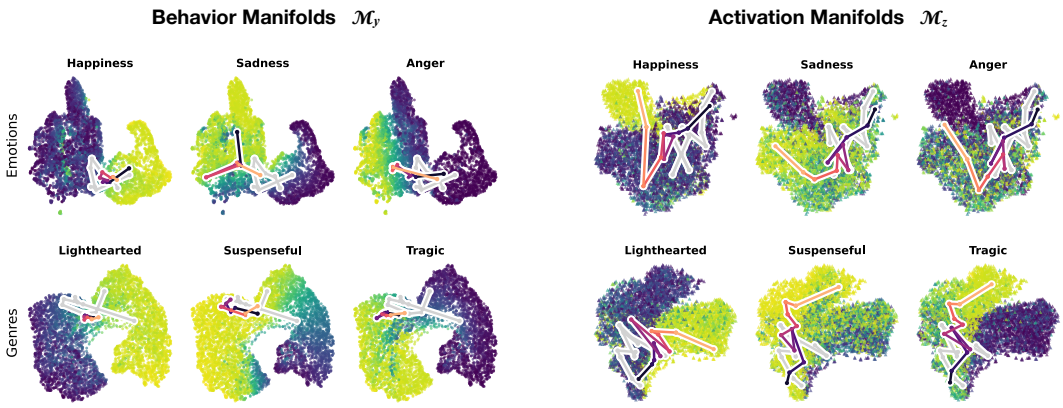


Figure 5: **Steering along manifolds** When we steer the story shown in Figs. 1 3, the story's trajectory in belief space shifts towards the parts of the manifold that corresponds to a particular concept. Each figure here shows the effect of steering the trajectory b_t towards a particular concept c, along with the manifold $\mathcal{M}_y$ or $\mathcal{M}_z$ with each point colored by the value $y_{t,c}$ for that concept. Results for all concepts listed in App. C.

add in the steering vector, multiplied by some scalar magnitude α, i.e., $\widetilde{z}_\ell(x_{1:t}) = z_\ell(x_{1:t}) + \alpha\, v_{c,\ell}$. We use two primary methods for activation steering: linear probe weights and difference in means steering vectors. For linear probes, we directly use the normalized parameter vector $\frac{\theta}{\|\theta\|}$ for a given attribute. Difference-in-Means (DIM) vectors (Bau et al., 2019a; Panickssery et al., 2024; Turner et al., 2024; Marks & Tegmark, 2024; Rimsky et al., 2024), are constructed by taking the mean activations for two contrasting datasets $v_{c,\ell} = \overline{z_\ell(x_c)} - \overline{z_\ell(x_{\neg c})}$.

4 Experiments

We use the SimpleStories dataset (Finke et al., 2025), a collection of synthetically generated short stories with various themes and styles. In our analyses, we use a training dataset of 2500 randomly sampled stories, with 500 stories each for each of the following "style" categories: *adventure*, *classic*, *lighthearted*, *melancholy*, and *tragic*, as well as a held-out test dataset with 500 stoies. For analysis, the conceptual domains we use are *Emotions*, *Genres*, and *Arbitrary*, and for each domain we test between 6 concepts. These domains were chosen such that Emotions and Genres should be expected to have a latent conceptual structure, such as hierarchy, except for the Arbitrary domain which serves as a counter-example and is expected to have less clear structure, and less predictive low-dimensional embeddings. For dimensionality reduction, we filter our training dataset to the 1000 sentences which maximally activate each concept, with a maximum 3 data points for a single story. The experiments presented here all use `llama-3.1-8b-instruct` model (Grattafiori et al., 2024) with 4-bit weight quantization for efficiency (Dettmers et al., 2023).

5 Results

Behavioral beliefs can be linearly decoded from activations We find rich behavioral dynamics in `Llama-3.1-8b` (Grattafiori et al., 2024) for many attributes in nearly all stories in our dataset, which follow the semantic content of each story in subjectively reasonable ways (Figs. 1, 3) [3]. Our probes at layer $\ell = 9$ are able to predict behavioral belief dynamics with high accuracy (Figs. 3), achieving .09 RMSE on the *Emotions* domain, .09 RMSE on the *Genres* domain, and .11 RMSE on the *Arbitrary* domain (see App. E for analysis across layers). These results support our claim that the behavioral signatures we find are reflective of beliefs that the LLM holds while reading a story, since we can predict behavior from the model's activations even when it is not instructed to consider specific concepts.

[3] An interactive dashboard with belief dynamics and trajectories for all stories in our test set is available at: https://conceptual-beliefs.streamlit.app

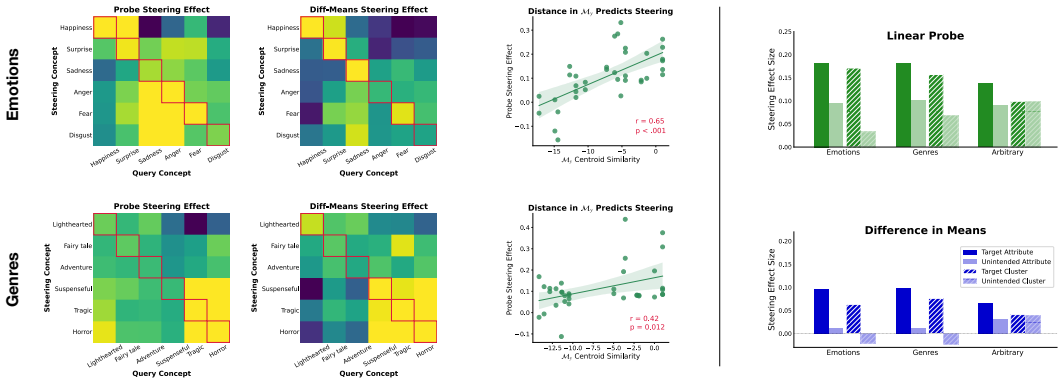


Figure 6: **Steering entanglement follows manifold geometry** We find that steering for a target concept c (like *sadness*) often increases belief in related but unintended concepts c' (like *anger*). **(Left)** Steering effect, or change in $y_{t,c}$, when steering for a particular target concept c (y-axes on heatmaps), and measuring belief in query concept c'. *Emotions* domain is on top, *Genres* on bottom. **(Middle)** Steering effect can be predicted from embedding similarity $d_{\mathcal{M}_y}$. **(Right)** Steering effect is significantly higher for the target concept c compared to the mean effect across other concepts c'. However, when we cluster concepts according to the top level of the hierarchies in Fig. 4, the effect on the "unintended" cluster disppears for the *Emotions* and *Genres* domains. We see lower on-target effects and higher off-target effects in the manifold of *Arbitrary* concepts, for which the concepts do not share geometric structure.

Shared structure in low-dimensional belief and representation manifolds We find strong qualitative and quantitative structure in the geometry of $\mathcal{M}_y$ and $\mathcal{M}_z$ (Fig. 3) for the *Emotions* and *Genres* domains, and belief trajectories $b_{1:T}$ follow relatively smooth and interpretable paths. Behavioral embeddings $\mathcal{M}_y$ closely follow activation embeddings $\mathcal{M}_z$ (Fig. 1), particularly for the *Emotions* and *Genres* domains. This is true even when we only compare two-dimensional centroids in $\mathcal{M}_y$ for each concept. In order to measure similarity between low-dimensional spaces, we map data points in these spaces to distance matrices, as shown in Fig. 4 (details in our section on *Distance Matrices*). We find that the resulting distance matrices are highly correlated between values embedded into the behavior manifold $\mathcal{M}_y$ and activation manifold $\mathcal{M}_z$ (Pearson's $r = .92, p < .001$ for *Emotions*, $r = .89, p < .001$ for *Genres*, and $r = .71, p < .001$ for *Arbitrary*). We find that the hierarchies inferred in Fig. 4 directly correspond to our qualitative observations of $\mathcal{M}_y$, and inferred hierarchies are equal between $\mathcal{M}_y$ and $\mathcal{M}_z$ for the *Emotions* and *Genres* domains.

Finally, we compare the embedding centroids we learn for the *Emotions* domain to human data from Russell & Mehrabian (1977), where people rated emotion words according to Valence and Arousal, and Dominance [4] (Fig. 4, Right). We find that our centroid embeddings are a close match for human emotion ratings on the dimensions of Valence and Arousal, and that the distance matrix for the human data is highly correlated with $\mathcal{M}_y$ and $\mathcal{M}_z$ ($r = .93, p < .001$ for $\mathcal{M}_y$), complementing prior work demonstrating similarity of emotion organization in language models and humans (Zhao et al., 2025; Sofroniew et al., 2026; Sun et al., 2026). Varying UMAP dimensionality shows that two-dimensional embeddings preserve much of the concept-centroid distance structure in the original spaces. Additional dimensions make little difference for preserving structure in behavior, but have modest but steady improvements in preserving structure in activations (App. **??**).

Manifold geometry predicts steering effects We find that activation steering with linear probes and difference-in-means are effective in steering LLM beliefs (Fig. 6). With linear probes, we found that steering with only a single layer was ineffective, since later layers would correct for the intervention and nullify its effects (App. F). Steering simultaneously with vectors across a span of 7 layers corrected for this. With difference-in-means steering we did not encounter this issue, and instead found steering at a single layer to be effective.

Next, we found that steering for one concept often leads to a change in behavior for other concepts (Fig. 6). Crucially, steering entanglement between concepts mirrors the geometry of the conceptual domains, e.g. steering for a negative emotion such as *sadness* leads to an increase in *anger* ratings and a decrease in *happines* ratings. We estimated steering entanglement by comparing the effect on

[4] We limit the space to Valence and Arousal in line with circumplex theory (Russell, 1980).

the targeted concept to the average effect of steering on our other measured concepts. Comparing across steering methods, we find that difference in means steering, compared to linear probes, have less effect on unintended concepts, but they also have a smaller maximum effect on the target concept (Fig. 6, Right). Additionally, in App. M we applied the same steering vectors during story generation and evaluate the resulting continuations with an unsteered model. The continuations exhibit emotional themes consistent with the steering targets, providing evidence that these conceptual representations influence generated narrative content as well as elicited judgments.

We also find that steering entanglement can be predicted based on the structure of our learned manifold $\mathcal{M}_y$. Steering effect between two concepts is correlated with the distance between their centroids in $\mathcal{M}_y$ ($r = .65, p < .001$ for *Emotions*, $r = .42, p = .01$ for *Theme*; Fig. 6, Center Left). We can also use a linear model to predict out-of-sample the degree of effect that steering for one concept will have for another queried concept (i.e., c in q_c; Fig. 6, Center Right). Finally, we re-analyzed steering effect on target or unintended concepts by using the top-level clustering of the dendrograms that we infer for $\mathcal{M}_y$ to measure the steering effect of concepts within a cluster, against the steering effect on concepts outside this cluster. E.g. in the emotion domain, we have two clusters, one with $\{$*happiness*, *surprise*$\}$ and one with the other four emotions.

6 Discussion

Our results broadly support our theory that LLMs operate over low-dimensional *conceptual belief spaces* which group into structured domains of related dimensions. We find support that our behavioral probe is qualitatively working as expected by examining $y_{t,c}$, and that this can be predicted from z_t with linear probes suggests that the same information is represented in the LLM's residual stream when it only processes the story text $x_{1:t}$. The manifolds we learn $\mathcal{M}_y$, $\mathcal{M}_z$ have rich structure, and where behavior $\mathcal{M}_y$ and activation $\mathcal{M}_z$ manifolds are highly correlated and share the same hierarchical structure for the domains of *Genres* and *Emotions* (Sofroniew et al., 2026; Zhao et al., 2025; Dong et al., 2025). We can use linear activation steering to alter model behavior and representations, shifting the in-context learning trajectories for different stories along $\mathcal{M}_y$ and $\mathcal{M}_z$. We also observe steering entanglement across concepts, where steering for one concept unintentionally also shifts model beliefs for another concept, and this steering entanglement can be predicted based on the hierarchical structure in $\mathcal{M}_y$.

Our work is a first step towards understanding conceptual belief spaces in LLMs, and for developing a principaled theoretical framework which unifies geometric theories of conceptual representation with probabilistic theories of learning. We hand-select three domains with six concepts each, but our method is limited by how we chose domains and concepts based on prior knowledge. This parallels the hand-engineering of a concept space in Bayesian cognitive science (Tenenbaum et al., 2011), although like with these methods (Wong et al., 2023), a promising direction for future work will be to apply "auto-interpretability" methods (Bills et al., 2023; Paulo et al., 2024) for proposing conceptual domains for a given LLM task. Another limitation is that the concepts we test are treated as unitary and non-compositional, and are also global for the story as a whole, rather than applying to e.g. a particular character in a story. We hope to extend these methods in future work with much richer conceptual belief spaces, with more elaborate concepts. Finally, another question we hope to explore in future work is how conceptual belief spaces shape uncertainty in long-form text generation (Bigelow et al., 2024; Zur et al., 2025; Ahdritz et al., 2024; Boppana et al., 2026; Stolfo et al., 2024; Han et al., 2024), in addition to language understanding. Our experiments in App. M are an initial step in this direction.

This work represents an early step towards a theoretical foundation for LLM interpretability which unifies principles of Bayesian program learning (Tenenbaum et al., 2011; Lake et al., 2017; Ullman & Tenenbaum, 2020) and conceptual spaces (Gärdenfors, 2000; Gärdenfors & Williams, 2001) in cognitive science. However, much remains to be done in further solidifying this foundation. For example, our results raise the questions of what effect different kinds of activation steering could have on belief trajectories, and whether a new steering could be developed which takes advantage of the manifolds $\mathcal{M}$. Our framework also raises the question of how, specifically, these conceptual spaces relate to sophisticated behaviors involving likelihood $p(c \mid x)$ and predictive distributions $p(y \mid c)$.

REFERENCES

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. Many-shot in-context learning. 2024.

Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L Edelman. Distinguishing the knowable from the unknowable with language models. *arXiv preprint arXiv:2402.03563*, 2024.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

Cem Anil, Esin Durmus, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Nina Rimsky, Meg Tong, Jesse Mu, Daniel Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, Jamie Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R Bowman, Ethan Perez, Roger Grosse, and David Duvenaud. Many-shot jailbreaking. 2024.

Maria Antoniak, David Mimno, and Karen Levy. Narrative paths and negotiation of power in birth stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27, 2019. doi: 10.1145/3359190. URL https://doi.org/10.1145/3359190.

Aryaman Arora, Dan Jurafsky, Christopher Potts, and Noah D Goodman. Bayesian scaling laws for in-context learning. *arXiv preprint arXiv:2410.16531*, 2024.

Christian Balkenius and Peter Gärdenfors. Spaces in the brain: From neurons to meanings. *Frontiers in psychology*, 7:1820, 2016.

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*, 2019a. URL https://openreview.net/forum?id=H1z-PsR5KX.

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*, 2019b. URL https://openreview.net/forum?id=H1z-PsR5KX.

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

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 861–872, 2017.

Eric Bigelow, Ari Holtzman, Hidenori Tanaka, and Tomer Ullman. Forking paths in neural text generation. *International Conference on Learning Representations*, 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. *arXiv preprint arXiv:2511.00617*, 2025.

Eric J Bigelow. *Towards a Cognitive Science of Large Language Models*. PhD thesis, Harvard University, 2026.

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

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

Siddharth Boppana, Annabel Ma, Max Loeffler, Raphael Sarfati, Eric Bigelow, Atticus Geiger, Owen Lewis, and Jack Merullo. Reasoning theater: Disentangling model beliefs from chain-of-thought. *arXiv preprint arXiv:2603.05488*, 2026.

Paul K Brown and George Wald. Visual pigments in single rods and cones of the human retina. *Science*, 144(3614):45–52, 1964.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 35:18878–18891, 2022.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2126–2136, 2018.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Yurui Dong, Luozhijie Jin, Yao Yang, Bingjie Lu, Jiaxi Yang, and Zhi Liu. Controllable emotion generation with emotion vectors. *arXiv preprint arXiv:2502.04075*, 2025.

Lennart Finke, Chandan Sreedhara, Thomas Dooms, Mat Allen, Emerald Zhang, Juan Diego Rodriguez, Noa Nabeshima, Thomas Marshall, and Dan Braun. Parameterized synthetic text generation with simplestories. *arXiv preprint arXiv:2504.09184*, 2025.

Peter Gärdenfors. *Conceptual spaces: The geometry of thought*. MIT press, 2000.

Peter Gärdenfors and Mary-Anne Williams. Reasoning about categories in conceptual spaces. In *IJCAI*, volume 2001, pp. 385–392, 2001.

Jiayi Geng, Howard Chen, Ryan Liu, Manoel Horta Ribeiro, Robb Willer, Graham Neubig, and Thomas L Griffiths. Accumulating context changes the beliefs of language models. *arXiv preprint arXiv:2511.01805*, 2025.

Noah D Goodman, Joshua B Tenenbaum, Jacob Feldman, and Thomas L Griffiths. A rational analysis of rule-based concept learning. *Cognitive science*, 32(1):108–154, 2008.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie CY Chan, Andrew Lampinen, Jane X Wang, Zeynep Akata, and Eric Schulz. Machine psychology. *arXiv preprint arXiv:2303.13988*, 2023.

Jiatong Han, Jannik Kossen, Muhammed Razzak, and Yarin Gal. Semantic entropy neurons: Encoding semantic uncertainty in the latent space of llms. In *MINT: Foundation Model Interventions*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Eghbal A Hosseini, Yuxuan Li, Yasaman Bahri, Declan Campbell, and Andrew Kyle Lampinen. Context structure reshapes the representational geometry of language models. *arXiv preprint arXiv:2601.22364*, 2026.

Jennifer Hu, Eric Bigelow, Ekdeep S Lubana, Kanishk Gandhi, Laura Ruis, Thomas Fel, Ellie Pavlick, and Noah Goodman. First workshop on coginterp: Interpreting cognition in deep learning models. In *Advances in Neural Information Processing Systems*, 2025a. URL http://coginterp.github.io/neurips2025/.

Jennifer Hu, Felix Sosa, and Tomer Ullman. Re-evaluating theory of mind evaluation in large language models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 380 (1932), 2025b.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Andrew Kyle Lampinen, Stephanie CY Chan, Aaditya K Singh, and Murray Shanahan. The broader spectrum of in-context learning. *arXiv preprint arXiv:2412.03782*, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

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.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, August 2024.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, July 2024.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the bayesian prism, 2024. URL https://arxiv.org/abs/2306.04891.

Core Francisco Park, Ekdeep Singh Lubana, Itamar Pres, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. *arXiv preprint arXiv:2412.01003*, 2024.

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*, 2025.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. *arXiv preprint arXiv:2410.13928*, 2024.

Steven T. Piantadosi. The computational origin of representation. *Minds and Machines*, 31(1):1–58, March 2021. ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-020-09540-9.

Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pp. 3–33. Elsevier, 1980.

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. *arXiv preprint arXiv:2505.14685*, 2025.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. The emotional arcs of stories are dominated by six basic shapes. *EPJ data science*, 5(1):1–12, 2016.

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/.

Laura Ruis, Arduin Findeis, Herbie Bradley, Hossein A Rahmani, Kyoung Whan Choe, Edward Grefenstette, and Tim Rocktäschel. Do llms selectively encode the goal of an agent's reach? In *First Workshop on Theory of Mind in Communicating Agents*, 2023.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6): 1161, 1980.

James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294, 1977.

Lenhart K Schubert and Chung Hee Hwang. Episodic logic meets little red riding hood: A comprehensive, natural representation for language understanding. *Natural language processing and knowledge representation: Language for Knowledge and Knowledge for Language*, 111174, 2000.

Adam S. Shai, Sarah E. Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-2387. URL https://mlanthology.org/neurips/2024/shai2024neurips-transformers/.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061, 1987.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/emotions/index.html.

Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. Confidence regulation neurons in language models. *Advances in Neural Information Processing Systems*, 37:125019–125049, 2024.

Corina Strößner. Criteria for naturalness in conceptual spaces. *Synthese*, 200(2):78, 2022.

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

Lihao Sun, Lewen Yan, Xiaoya Lu, Andrew Lee, Jie Zhang, and Jing Shao. Valence-arousal subspace in llms: Circular emotion geometry and multi-behavioral control. *arXiv preprint arXiv:2604.03147*, 2026.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4593–4601, 2019a.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=SJzSgnRcKX.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models, February 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization, October 2024.

Tomer D Ullman and Joshua B Tenenbaum. Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, 2(1):533–558, 2020.

Kurt Vonnegut. Shapes of stories. https://www.youtube.com/watch?v=oP3c1h8v2ZQ, 1995.

Fushun Wang, Jiongjiong Yang, Fang Pan, Roger C Ho, and Jason H Huang. Neurotransmitters and emotions, 2020.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

Patrick Henry Winston. The strong story hypothesis and the directed perception hypothesis. In *AAAI Fall Symposium: Advances in Cognitive Systems*. San Francisco, 2011.

Lionel Wong, Gabriel Grand, Alexander K. Lew, Noah D. Goodman, Vikash K. Mansinghka, Jacob Andreas, and Joshua B. Tenenbaum. From word models to world models: Translating from natural language to the probabilistic language of thought, June 2023.

Daniel Wurgaft, Ekdeep Singh Lubana, Core Francisco Park, Hidenori Tanaka, Gautam Reddy, and Noah D. Goodman. In-context learning strategies emerge rationally, June 2025.

Daniel Wurgaft, Can Rager, Matthew Kowal, Vasudev Shyam, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric Bigelow, Raphael Sarfati, Thomas McGrath, Owen Lewis, Jack Merullo, Noah Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh Lubana. Manifold steering reveals the shared geometry of neural network representation and behavior. *arXiv preprint arXiv:2605.05115*, 2026.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Catherine Yeh, Tara Menon, Robin Singh Arya, Helen He, Moira Weigel, Fernanda Viégas, and Martin Wattenberg. Story ribbons: Reimagining storyline visualizations with large language models. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

Jeffrey M Zacks and Barbara Tversky. Event structure in perception and conception. *Psychological bulletin*, 127(1):3, 2001.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*, 2023.

Bo Zhao, Maya Okawa, Eric J Bigelow, Rose Yu, Tomer Ullman, Ekdeep Singh Lubana, and Hidenori Tanaka. Emergence of hierarchical emotion organization in large language models. *arXiv preprint arXiv:2507.10599*, 2025.

Yufa Zhou, Yixiao Wang, Xunjian Yin, Shuyan Zhou, and Anru R. Zhang. The geometry of reasoning: Flowing logics in representation space. *arXiv preprint arXiv:2510.09782*, 2025. URL https://arxiv.org/abs/2510.09782.

Amir Zur, Eric Bigelow, Atticus Geiger, and Ekdeep Singh Lubana. Are language models aware of the road not taken? token-level uncertainty and hidden state dynamics. *ICML workshop on actionable interpretability*, 2025.

Rolf A Zwaan and Gabriel A Radvansky. Situation models in language comprehension and memory. *Psychological bulletin*, 123(2):162, 1998.

Rolf A Zwaan, Mark C Langston, and Arthur C Graesser. The construction of situation models in narrative comprehension: An event-indexing model. *Psychological science*, 6(5):292–297, 1995.

A STORY TEXT

Here in Fig. 7, we provide the full text for the story shown in Figs. 1, 3, 5.

i	**Sentence**
1	*From a hidden corner in a park, a boy named Samuel found an old lantern.*
2	*When he turned it on, a bright light shot out, revealing a dark hole in the ground.*
3	*Curious, he leaned in and fell straight down into the underground.*
4	*The air was thick and smelled of damp earth.*
5	*As he stood up, he saw he was in a strange world where time seemed to stop.*
6	*But soon, he heard cries for help.*
7	*Following the sound, he found a group of creatures trapped in a cave, scared and alone.*
8	*Thinking quickly, Samuel gathered rocks and started to clear the entrance.*
9	*The creatures cheered as he worked.*
10	*Finally, he pulled the last rock away, and the entrance was open.*
11	*The creatures ran out, thanking him for his bravery, while Samuel realized he had found true friends in an unexpected place.*

Figure 7: Full text for story in Figs. 1, 3, 5, split into numbered sentences.

B BELIEF QUERY PROMPT

In the following, we present the specific prompts q_c that we used to behaviorally query for the model's beliefs $y_{t,c}$ in specific concepts.

Emotion Query Prompt

```
Consider the following story:
```
{story}
```

Now, your task is to determine the emotional content of this story,
    specifically for the emotion {emotion_noun}.
Respond with only a single word, an integer in the range [0, 10],
    where 0 is not at all {emotion_adjective}, 5 is neutral, and 10
    is the most {emotion_adjective}.
What is the level of {emotion_noun} of this story?
```

Genre Query Prompt

```
Consider the following story:
```
{story}
```

Now, your task is to determine the thematic content of this story,
    specifically for the theme "{genre}".
Respond with only a single word, an integer in the range [0, 10],
    where 0 is not at all {genre}, 5 is neutral, and 10 is the most
    {genre}.
How much does this story match the theme "{genre}"?
```

Arbitrary Query Prompt

```
Consider the following story:
```
{story}
```

Now, your task is to determine whether the concept "{concept}"
    applies to this story.
Respond with only a single word, an integer in the range [0, 10],
    where 0 means the concept is very unlikely, 5 is neutral, and 10
    means the concept is very likely.
Does the concept "{concept}" apply to this story?
```

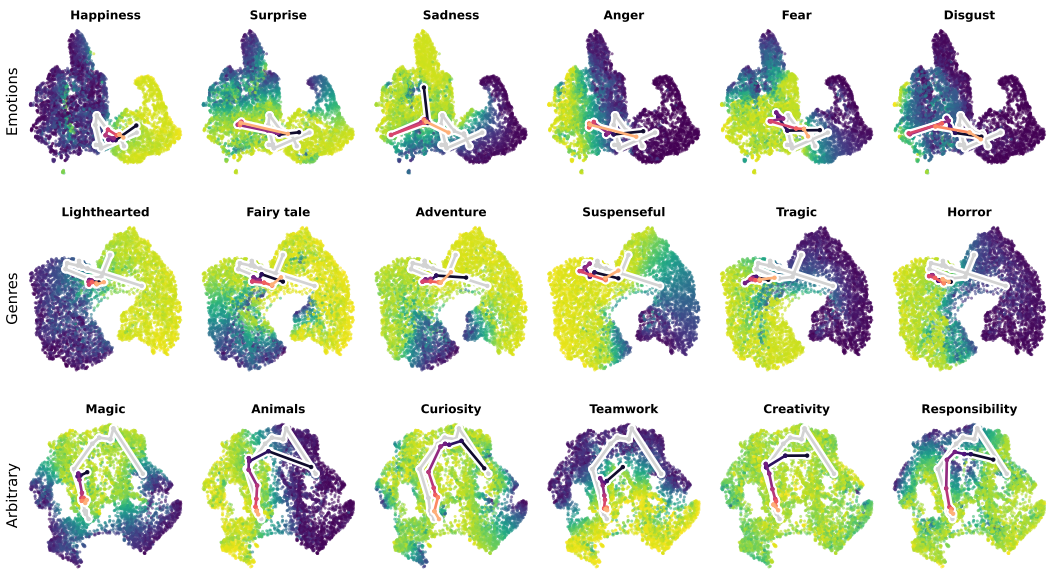


Figure 8: Steering effects for each individual concept c with linear probe, projected into $\mathcal{M}_y$ for each domain

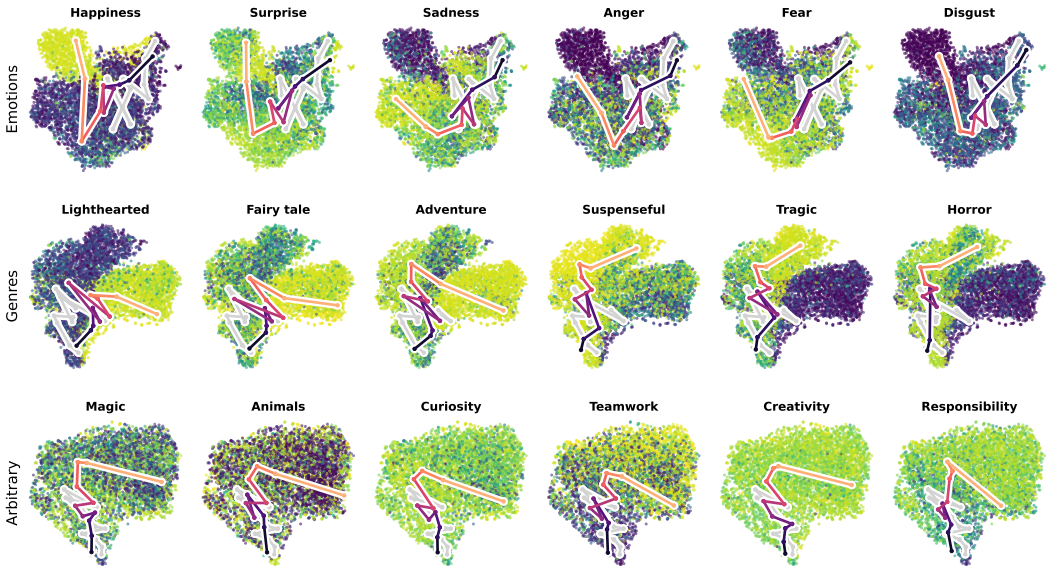


Figure 9: Steering effects for each individual concept c with linear probe, projected into $\mathcal{M}_z$ for each domain

C Steering Effects for All Concepts and Domains

Here, in Fig. 8 and Fig. 9 we present our main results in Figure 5 for all three domains, and for each concept c in each domain.

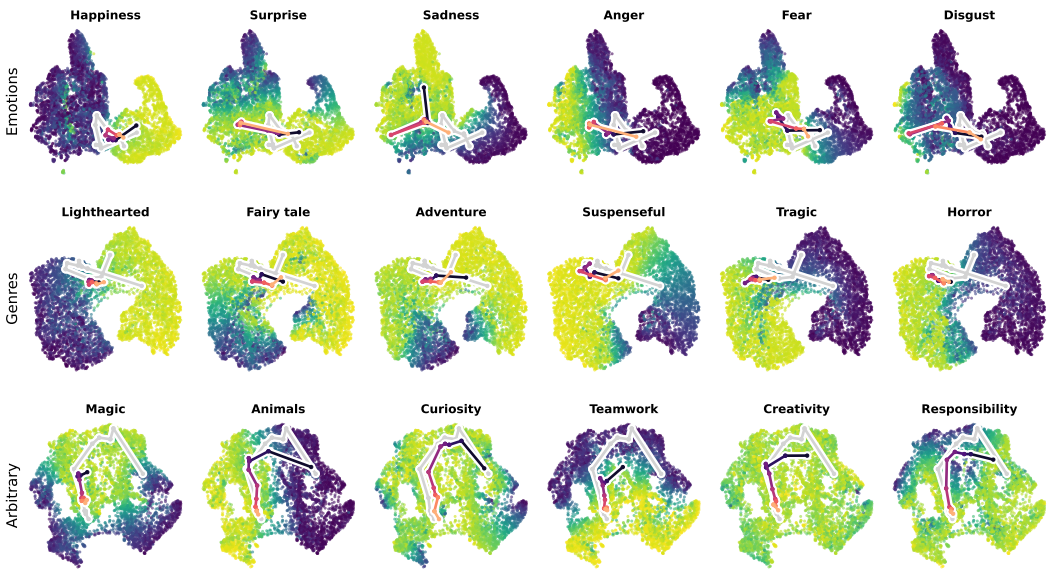


Figure 10: Distance matrix results at in Fig. 4 for *Arbitrary* Domain.

D Main Analyses for Arbitrary Domain

We show distance matrix analysis (Fig. 10) and steering entanglement (Fig. 11) for the *Arbitrary* domain.

We find that, while $\mathcal{M}_z$ has clear structure in the distance matrix, this structure is less similar to $\mathcal{M}_y$ than is the case for our other two domains. We also find that structure in $\mathcal{M}_y$ is not predictive of steering entanglement in this domain, as it is for *Emotions* and *Genres*.

Taken together, these results support our hypothesis that a relatively random set of concepts grouped together will not satisfy our expected criteria for being a *domain*. That said, it is not surprising that we find structure in $\mathcal{M}_z$ and $\mathcal{M}_y$ given that there are likely various subtler correlations between the concepts in the *Arbitrary* domain and in the max-activating examples. For example, in our story dataset, *animals* and *magic* are loosely correlated since many stories with animals also involve talking animals, and thus magic.

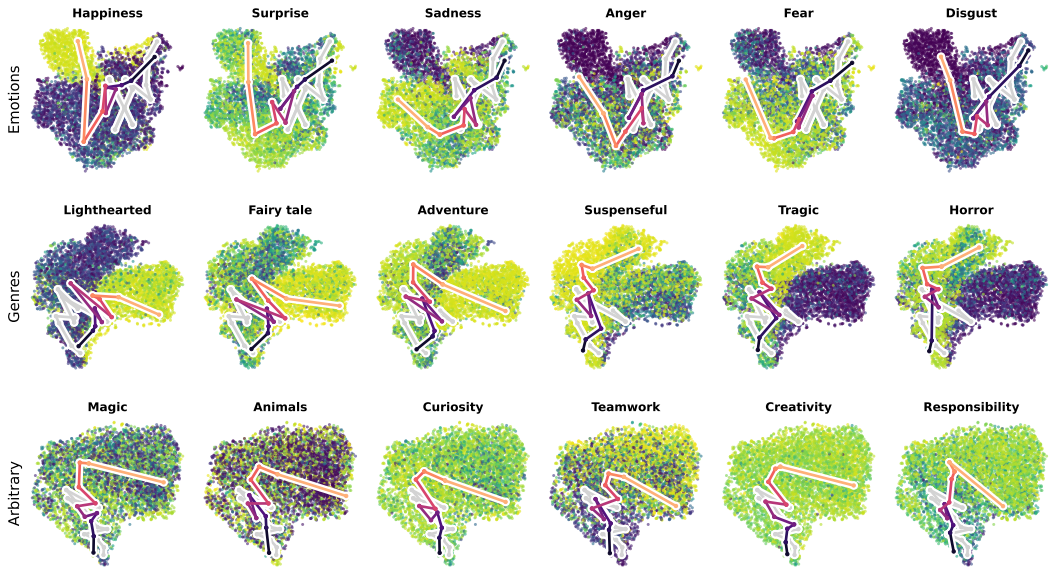


Figure 11: Steering entanglement and prediction from $\mathcal{M}_y$ for *Arbitrary* domain, as in Fig. 6.

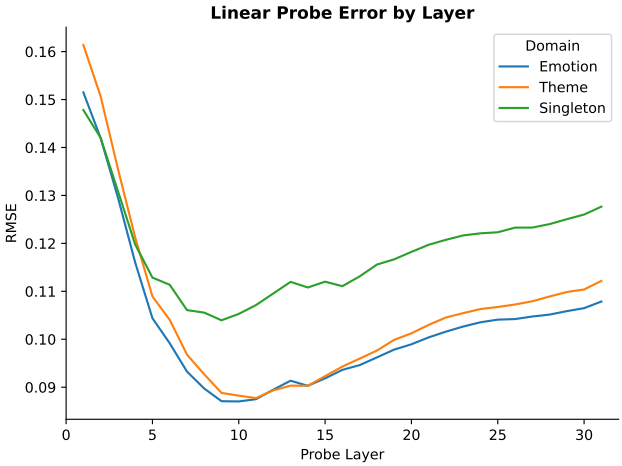


Figure 12: Linear accuracy for each layer, across each domain.

E Selecting Optimal Layer for Probe

We use layer 9 in our steering experiments, which we select as the layer where linear probes for all domains are approximately optimal, as shown in Fig. 12

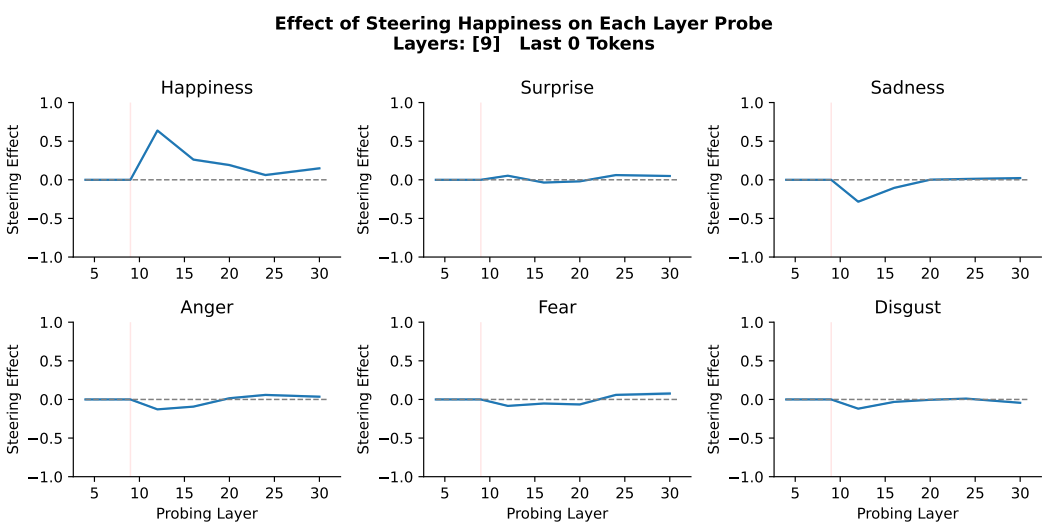


Figure 13: Using probe weights to steer at layer 8 only, we see that steering effect disappears after a few layers. Note: layer indexes in title are off by 1.

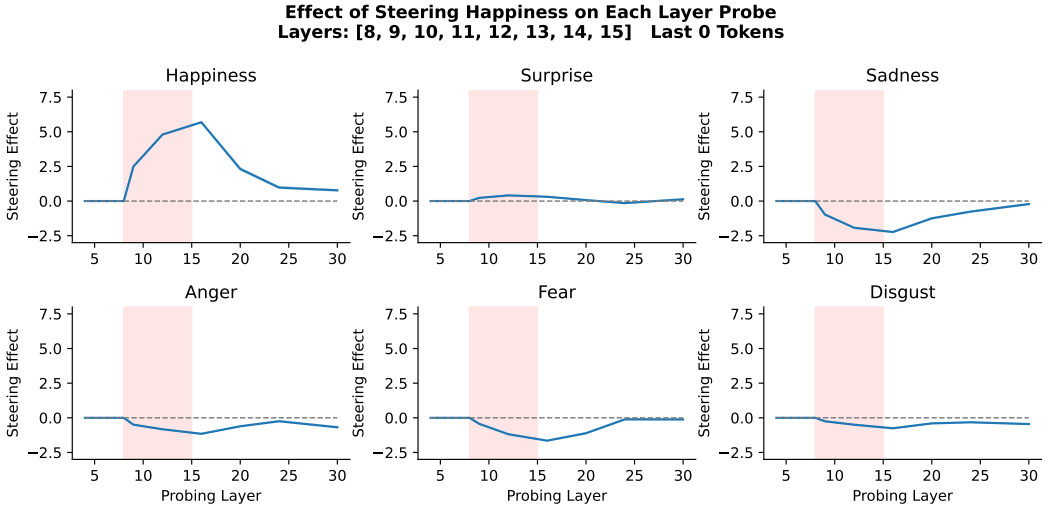


Figure 14: Using probe weights to steer at layers 7-14, we see that steering effect perseveres until the output layer. Note: layer indexes in title are off by 1.

F Probe Steering Effects by Layer

We empirically find that steering with linear probe weights at a single layer does not change model behavior. To better understand why this happened, we used our linear probes for a set of layers ℓ spaced throughout the model to measure whether steering effect actually occurs, but disappears at later layers due to model self-repair.

We find that indeed, with one layer, steering effect occurs at the targeted layer but disappears quickly thereafter (Fig. 13). If we instead steer at a range of layers, the effect is able to persevere all the way to the model output (Fig. 14).

G BEHAVIORAL DATA DISTRIBUTIONS

As a basic metric of how correlated our raw data $y_{t,c}$ is across concepts, in Figs. 15, 16, 17 we show correlations and histograms for each individual concept c and each pair of concepts across all domains.

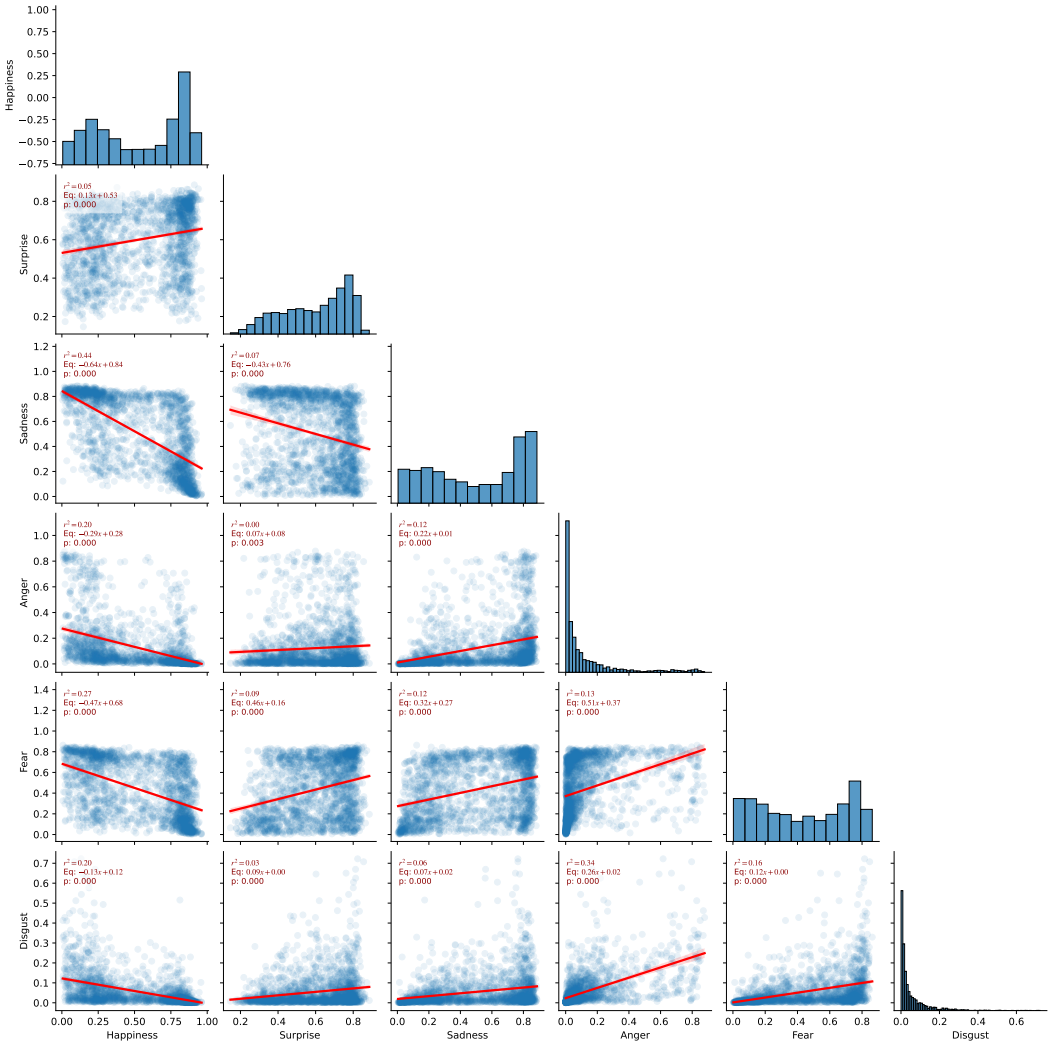


Figure 15: Correlations in $y_{t,c}$ for each pair of concepts in the *Emotions* domain. A subset of 2000 random data points are shown here.

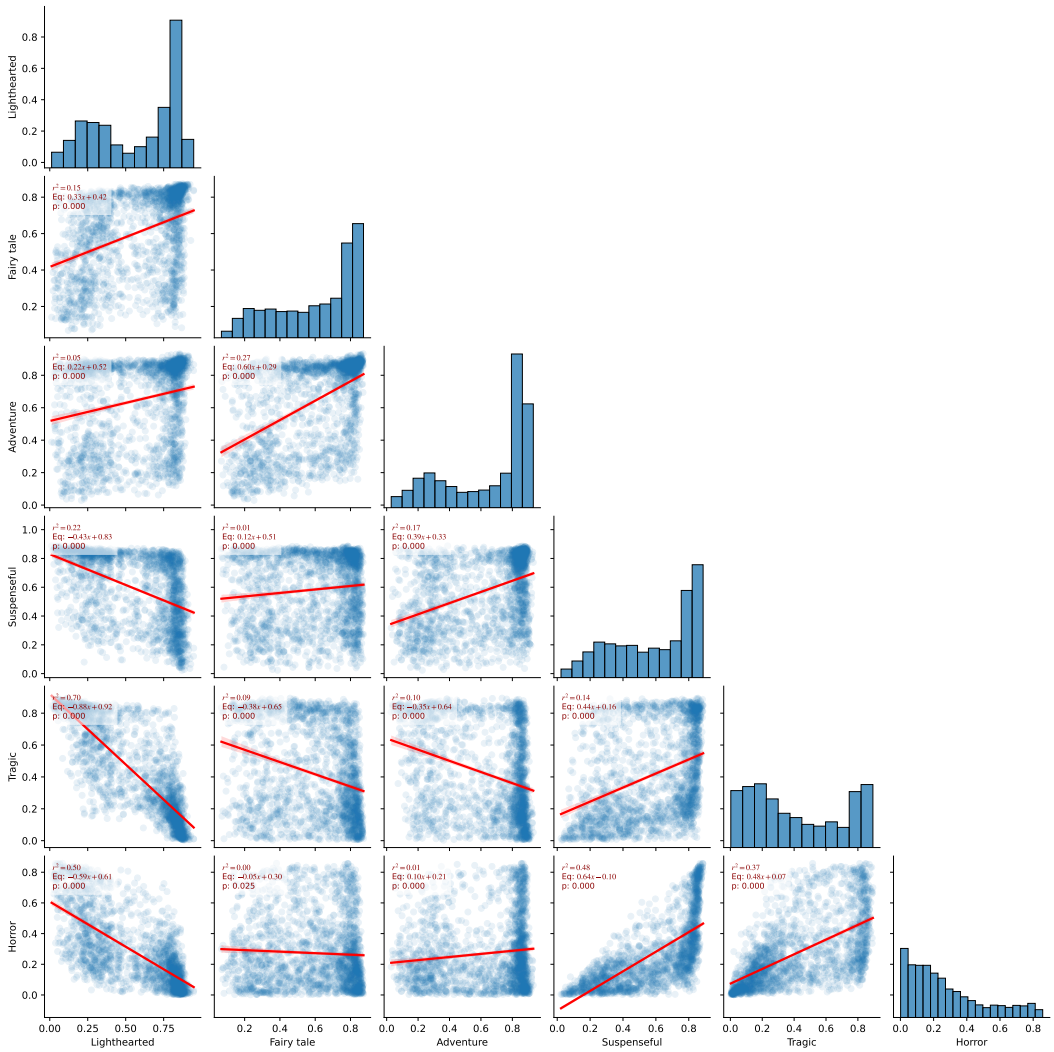


Figure 16: Correlations in $y_{t,c}$ for each pair of concepts in the *Genres* domain. A subset of 2000 random data points are shown here.

H CALIBRATING LINEAR PROBES

We empirically observed that when plotting probe predictions $\widehat{y}_{t,c}$ against ground truth $y_{t,c}$, predictions were systematically biased, indicative of miscalibration. To rectify this, we applied isotonic regression, which largely corrects this bias (Fig. 18).

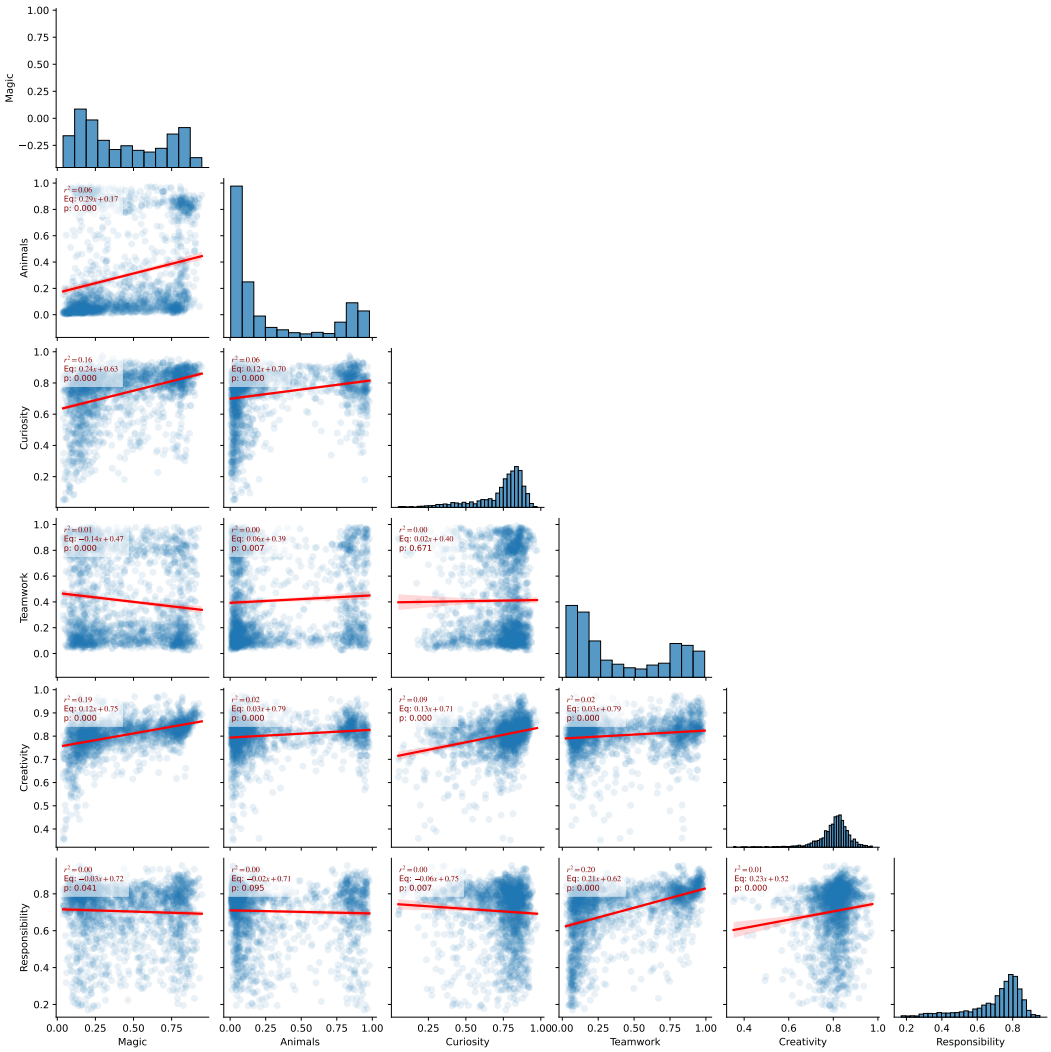


Figure 17: Correlations in $y_{t,c}$ for each pair of concepts in the *Arbitrary* domain. A subset of 2000 random data points are shown here.

I Further Analysis of Steering Effects

In Fig. 19 we show the effects of steering for two concepts *happiness* and *sadness* on belief dynamics in $y_{t,c}$, as well as the effect of steering at varying magnitude α. Results shown here are for the linear probe.

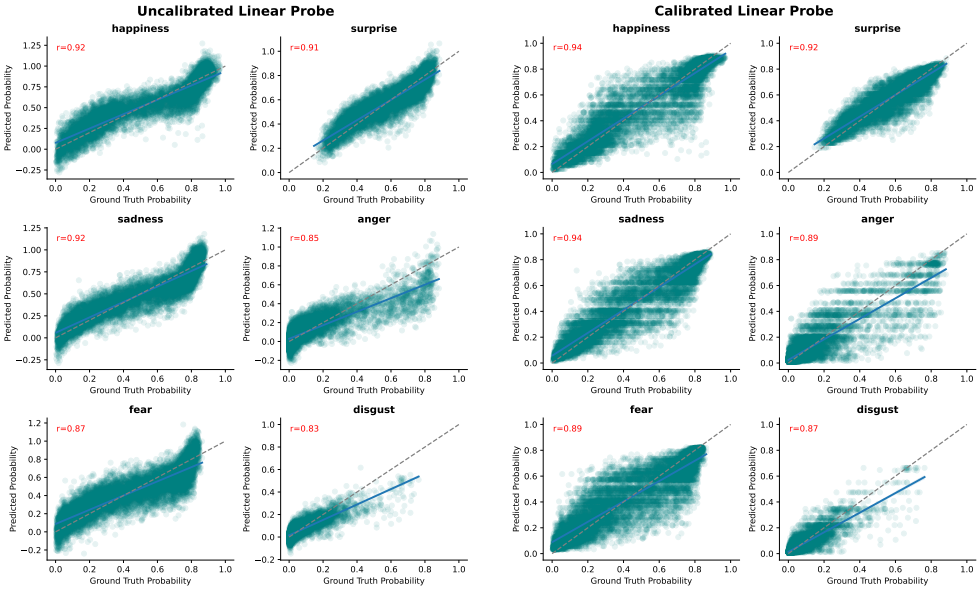


Figure 18: (Left) We find that probe predictions $\widehat{y}_{t,c}$ are initially miscalibrated with respect to ground truth $y_{t,c}$. (Right) After applying isotonic regression, predictions are more calibrated and less systematically biased.

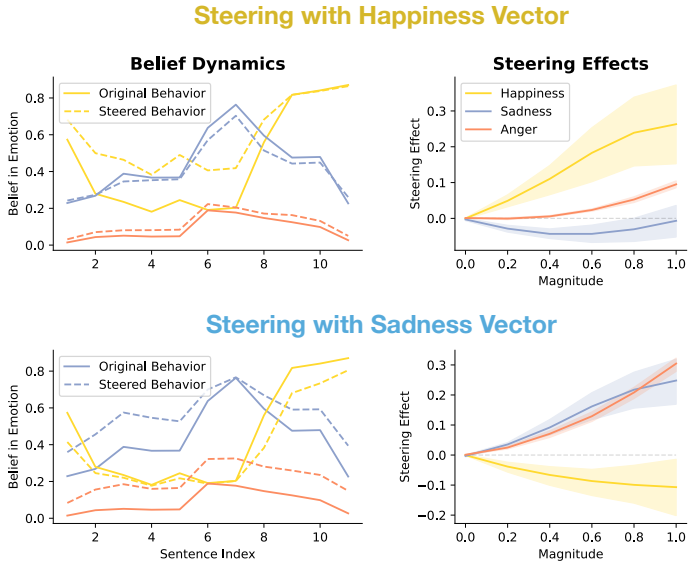


Figure 19: (Left) We observe that activation steering alters model belief dynamics, depending on which concept is targeted. (Right) Steering effect varies systematically as a function of steering vector magnitude. We also observe entanglement where steering effects unintended concepts, e.g. steering for *sadness* impacts belief in *anger*.

J Belief Dynamics for Additional Stories

The belief dynamics for each input text has its own nuanced story to tell. Here, in Fig. 20 we show two additional stories other than the one story considered in the main text. An interactive dashboard with belief dynamics and trajectories for all stories in our test set is available at: https://conceptual-beliefs.streamlit.app.

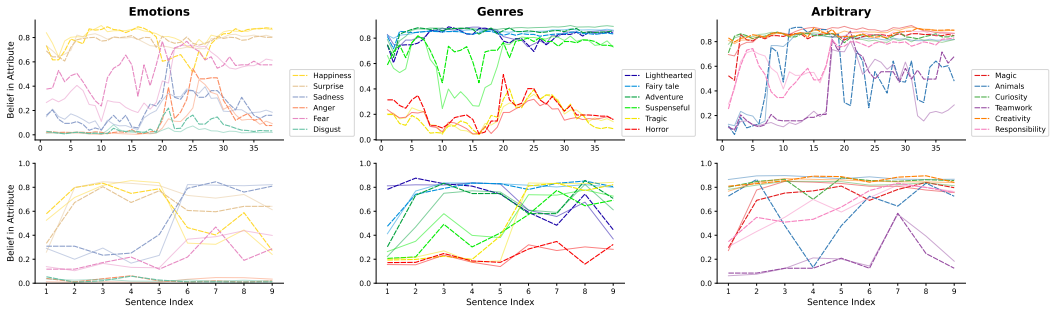


Figure 20: Belief Dynamics in $y_{t,c}$ for two additional stories.

K Sentence Index Encoded in Activation Manifold

We find that $\mathcal{M}_z$ not only encodes the structure of the domain, as in Fig. 5, but it also encodes sentence index t, i.e., an estimate of the current progress within a story (Fig 21).



Figure 21: PCA of activations for max activating examples, colored by the story progress, i.e., sentence index (Top) and by max-activating emotion label (Bottom)

L SENSITIVITY TO UMAP DIMENSIONALITY

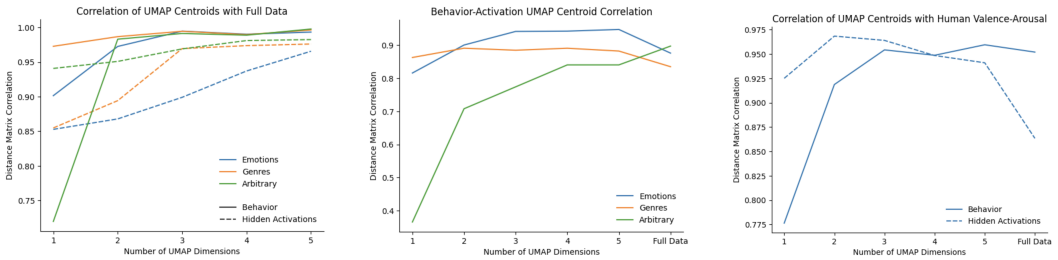


Figure 22: **Effect of UMAP dimensionality on concept-centroid geometry.** All panels report Pearson correlations between concept-centroid distance matrices. (Left) Correspondence between distances in each UMAP embedding and its original behavior or activation space. (Center) Correspondence between behavior and activation distances at matched embedding dimensionalities and in the original spaces. (Right) Correspondence between model distances for Emotions and distances in human valence–arousal space. Colors identify conceptual domains; solid and dashed lines in the left and right panels denote behavior and activations, respectively. "Full Data" denotes distances computed without UMAP. Low-dimensional embeddings capture substantial centroid structure, although additional dimensions improve preservation of original-space distances and do not uniformly improve correspondence with human data.

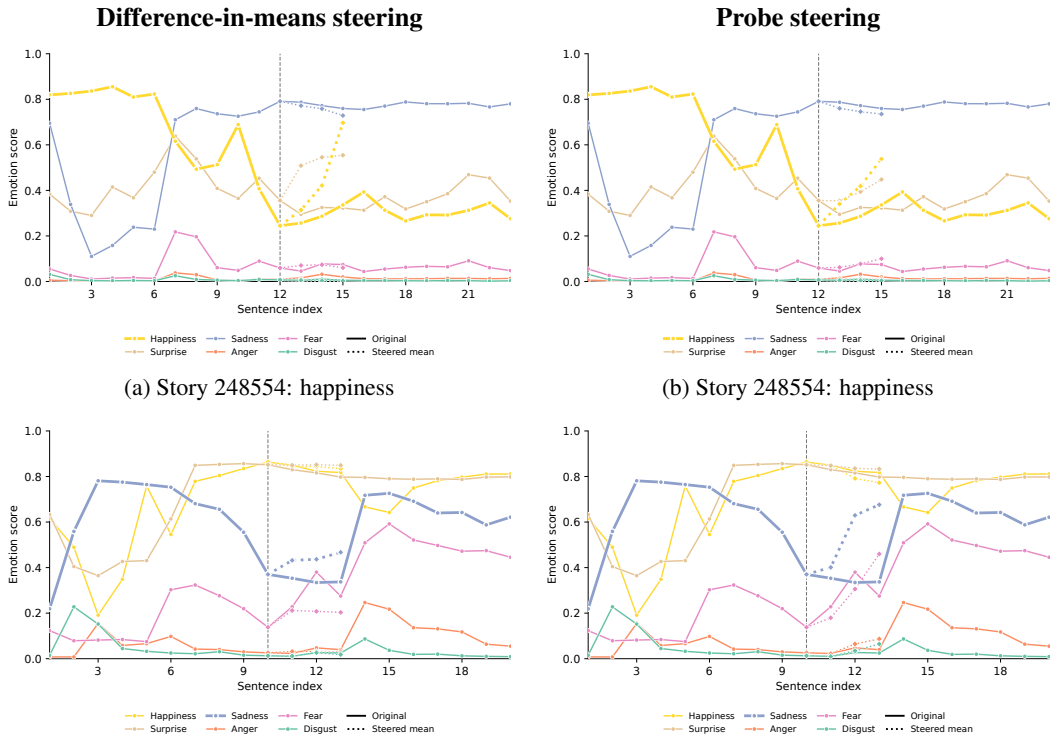
To assess how well low-dimensional embeddings preserve relationships among concepts in the original behavior and activation spaces, we repeated our distance-matrix analyses with UMAP embeddings ranging from one to five dimensions. We evaluated three properties: preservation of concept-centroid distances, correspondence between behavior and activations, and correspondence with human valence–arousal space.

First, Fig. 22 (Left) shows Pearson correlations between concept-centroid distance matrices computed in each original space and those computed in its UMAP embeddings. Two-dimensional embeddings closely preserve these relationships in behavior across all three domains. Correlations are also high for activations, although additional dimensions yield further improvements, particularly for Emotions and Genres. Thus, two dimensions capture much of the structure in concept-centroid distances, without preserving all of the structure captured by higher-dimensional embeddings.

Second, Fig. 22 (Center) compares concept-centroid distance matrices between behavior and activations at each embedding dimensionality, as well as in the original spaces ("Full Data"). For Emotions and Genres, correlations remain high across all tested dimensionalities. For Arbitrary, correlations are lower in low-dimensional embeddings but increase with dimensionality and are high in the original spaces. These results suggest that shared behavioral and representational structure is more readily captured in low dimensions for Emotions and Genres than for Arbitrary. This is consistent with our hypothesis that domains containing related concepts are particularly amenable to a low-dimensional conceptual-space account, although shared structure is also present for the Arbitrary domain.

Finally, Fig. 22 (Right) compares concept-centroid distance matrices for Emotions with the distance matrix derived from human valence–arousal coordinates. For behavior, the correlation is high with two dimensions and increases further with three, with relatively little change thereafter. For activations, the correlation is highest with two dimensions, decreases as embedding dimensionality increases, and is lower in the original space. This pattern is consistent with low-dimensional activation embeddings emphasizing relationships aligned with human valence–arousal space, while additional dimensions capture structure less closely aligned with that space.

Together, these analyses support using two-dimensional embeddings to visualize relationships among concept centroids, while showing that higher-dimensional embeddings preserve additional structure, particularly in activations. These analyses concern centroid geometry and do not establish that two dimensions preserve all information in the underlying data.



(a) Story 248554: happiness

(b) Story 248554: happiness

(c) Story 1670970: sadness

(d) Story 1670970: sadness

Figure 23: **Emotion evaluations of steered story continuations.** Llama generates four continuations per condition, with a requested length of three sentences each, using the steering vectors from the main text. An unsteered model evaluates the emotional content of the resulting text. Rows show happiness steering for story 248554 and sadness steering for story 1670970; columns compare difference-in-means steering (left) and probe steering (right).

M STEERING STORY GENERATION

The experiments in the main text examine how activation steering changes a model's judgments about a story. Here, we test whether the same steering vectors also affect the stories the model generates. Starting from a story prefix, we apply the difference-in-means and probe steering vectors used in the main text to Llama and sample four continuations per condition, with a requested length of three sentences each. We then use an unsteered model to evaluate the emotional content of the generated text. This separates steering during generation from the subsequent evaluation, so changes in the evaluator's judgments cannot arise from directly steering the evaluator.

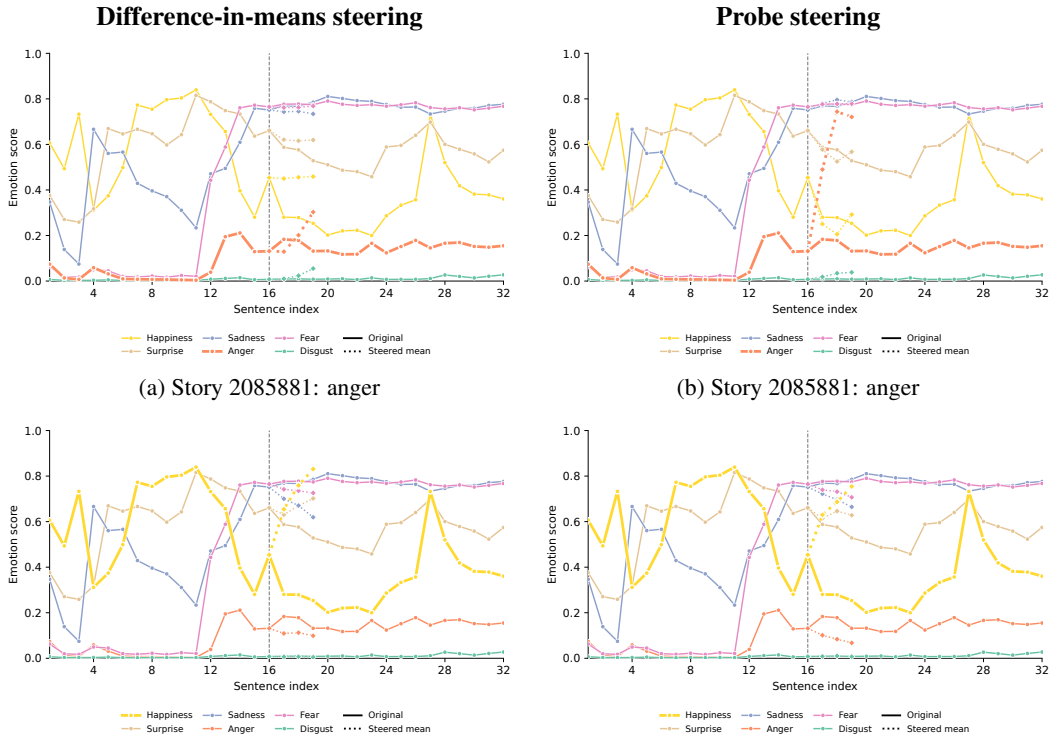
Figures 23 and 24 present the emotion evaluations for the supplied story–target pairs under both steering methods. Table 1 provides complementary examples for story 2085881, using the same prefix for anger and happiness steering. The anger-steered continuations emphasize hostility, conflict, and rejection of the protagonist's invention, often accompanied by sadness. The happiness-steered continuations instead emphasize optimism, curiosity, and new possibilities, although some uncertainty remains in individual samples.

These examples provide qualitative evidence that the steering vectors influence generated narrative content in addition to elicited judgments. Taken together, these results suggest that the conceptual representations we study in the main text are causally involved in both story evaluation and generation.

Table 1: **Example continuations under anger and happiness steering.** All continuations begin from the shared prefix below (story 2085881). Each row lists an independent story continuation.

Story prefix before steering.
Tired of the old ways, a man dreamed of making something new. He thought about how he could help people. Every day, he worked hard on his idea. But no one understood him. "What is light as a feather but cannot be held?" he would ask. His friends would laugh, not knowing he meant hope.
One day, he built a small machine that could help others see. It could show pictures of things far away. He was proud of his work. He showed it to a crowd. They stared in awe. But soon, they began to worry. "What if this machine takes away our jobs?" one said. Fear spread among them like fire.
Feeling sad, he tried to explain. "If you use this machine, you will find new ways to work.

Sample	Steered toward anger	Steered toward happiness
0	He pleaded for them to understand, but they wouldn't listen. As he watched his machine, a brilliant invention of his, being torn apart by angry fists, he felt his heart break. His dreams, once so full of promise, now seemed impossible.	Together, we can create a brighter future Your new jobs will be more exciting and fulfilling, he continued. The crowd's concerns began to fade away, replaced by a sense of optimism and curiosity. With a few more tweaks, the machine was ready for its maiden voyage.
1	His words fell on deaf ears, and the crowd's anger turned into a violent storm. The man's words of reason were drowned out by the cries of those who feared change. In the chaos, he saw a young girl holding a small child, her eyes fixed on the machine with a mix of wonder and hope.	New Horizons You will have more time to enjoy life and explore new opportunities," he said with a smile. But they weren't convinced yet. "Will this machine bring more than just pictures?" they asked, their curiosity piqued.
2	But the crowd was too angry, and it stormed away, leaving him heartbroken. He wondered if anyone would ever understand the message behind his creation, the message of hope.	His Artwork and Vision He shared pictures of the countryside and the mountains they had known. He spoke of the progress that was to come. The crowd's concerns began to fade, replaced by excitement and curiosity.
3	New Horizons He paused, his words heavy with the weight of hurt feelings, but the people wouldn't listen. The anger in their voices rose, and the man's heart sank, but he refused to give up. He stood firm, his eyes fixed on a young girl in the crowd, who watched with tears in her eyes, and a spark of understanding in her gaze.	Imagineers and Dreamers They will find new pictures to paint, new words to write, and new paths to walk. They will see new things, and their imaginations will soar. Together, we will create a brighter future.



(a) Story 2085881: anger

(b) Story 2085881: anger

(c) Story 2085881: happiness

(d) Story 2085881: happiness

Figure 24: **Steering toward different emotions from the same story prefix.** Emotion evaluations of continuations of story 2085881 under anger steering (top) and happiness steering (bottom), using difference-in-means vectors (left) or probe vectors (right). Each condition includes four generated continuations, with a requested length of three sentences each, evaluated by an unsteered model. Table 1 shows supplied continuation examples for this story.

N Related Work

In-context learning and belief-state geometry. In-context learning is often studied as few-shot task induction from input–output examples (Brown et al., 2020; Chan et al., 2022; Min et al., 2022), but recent work argues for a broader view in which context-dependent improvement over a sequence constitutes a form of in-context learning (Lampinen et al., 2024). Bayesian accounts similarly interpret ICL as inference over latent hypotheses, tasks, or data-generating processes (Xie et al., 2021; Zhang et al., 2023; Panwar et al., 2024). Closest to our belief-state framing, Shai et al. (2024) show that transformers can linearly represent the geometry of belief states in the residual stream for controlled sequence-prediction problems; our setting instead considers behaviorally elicited beliefs over semantic concepts in naturalistic stories.

Conceptual spaces and emotion structure. Our use of a low-dimensional conceptual space follows the conceptual-spaces tradition, which represents concepts geometrically and treats semantic similarity as distance in a metric space (Gärdenfors, 2000). The Emotions domain is also motivated by psychological accounts of affective structure, including valence–arousal and hierarchical models of emotion (Russell, 1980; Plutchik, 1980; Shaver et al., 1987). Recent work shows that LLM outputs exhibit hierarchical organization over emotion concepts (Zhao et al., 2025), providing a closely related example of cognitively grounded structure in model behavior.

Narrative dynamics and sentiment arcs. Computational narrative analysis has long studied how sentiment, topics, personas, and social relations vary over the course of a story. Reagan et al. (2016) find that many stories in Project Gutenberg can be summarized by a small set of emotional arcs, while Antoniak et al. (2019) analyze sentiment, topic, persona, and power dynamics across normalized positions in online birth stories. These works motivate the idea that narrative meaning is dynamic rather than fixed, and that temporal structure is central to understanding how stories are processed.

Representation-space trajectories and low-dimensional geometry. A growing line of work studies model computation as movement through representation space rather than as static representations at isolated tokens. For example, Zhou et al. (2025) model LLM reasoning as flows through activation space, relating reasoning steps to geometric quantities such as position, velocity, and curvature. More broadly, representational similarity analysis and dimensionality-reduction methods provide tools for comparing the geometry of neural or model representations (Kriegeskorte et al., 2008; McInnes et al., 2018). This literature is closely related to our use of low-dimensional manifolds and distance matrices to compare behavioral and activation-space structure.

Probing, steering, and causal representations. Linear probes are widely used to test whether linguistic, semantic, or task-relevant information is decodable from model activations (Alain & Bengio, 2017; Belinkov et al., 2017; Conneau et al., 2018; Tenney et al., 2019a; Hewitt & Manning, 2019; Belinkov, 2022). Activation steering and related causal-intervention methods go further by modifying hidden states and measuring resulting behavioral changes (Bau et al., 2019b; Subramani et al., 2022; Li et al., 2023; Turner et al., 2024; Panickssery et al., 2024). Todd et al. (2024) identify compact activation vectors that causally induce task behavior in ICL settings. Our methods draw on this probing-and-steering tradition while applying it to dynamically updated beliefs about narrative concepts.

O LLM Use Statement

LLMs were used for this work in discussing research ideas, in executing the experiments in App. M, in drafting the writing for Apps. M, L, in coding the interactive dashboard, and in proofreading the manuscript for typos.