Localizing and Evaluating Multinomial Concepts with Affine Subspaces

Divya Appapogu [1] **Freya Behrens** [1] **Yonatan Belinkov** [2 3] **Aaron Mueller** [1]

Abstract

The Linear Representation Hypothesis suggests that concepts in language models can be represented as linear directions in activation space. While empirically effective for binary concepts, more recent work suggests that representations of multinomial concepts may exhibit inherently multidimensional structure, such as circular or curved manifold geometries. Efficiently discovering multinomial concept representations and evaluating whether they precisely cover the full concept space remains a significant challenge. Building on prior success in identifying non-basis-aligned directions for binary concepts, we model multinomial concepts as affine subspaces, which can be viewed as multidimensional generalizations of directions (optionally with an offset). We extend existing methods of locating linear subspaces to affine concept subspaces in language model activation spaces, along with evaluation methods to characterize the precision and recall of recovered multinomial concept spaces. We demonstrate an application where we steer LLMs by sampling points from the recovered subspace; unlike one-dimensional steering, sampling enables us to steer a model's behavior toward diverse but concept-related behaviors.

1. Introduction

Language Models (LMs) have demonstrated remarkable capabilities (Anthropic, 2025; Singh et al., 2025; Comanici et al., 2025; Wei et al., 2022), and our knowledge of their internal workings is rapidly increasing. Many advances have been driven by mechanistic interpretability, which aims to understand trained models post hoc (Olah et al., 2020) via their internal computations. A central line of work in this area seeks to identify human-interpretable concept representations in neural network activations.

Much work on concept discovery defines mappings between concepts and one-dimensional linear directions in a model's representation space (Elhage et al., 2021; Park et al., 2023; 2025; Bricken et al., 2023; Nanda et al., 2023; Huben et al.,

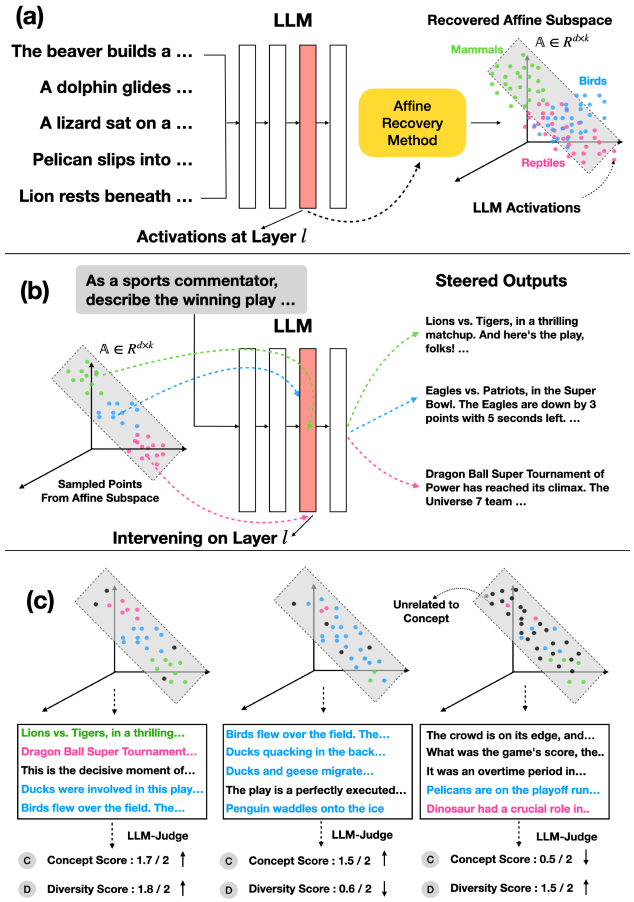

Figure 1. **(a)** We propose methods for recovering affine representations of multinomial concepts in LM activations (§2). **(b)** We sample points from the recovered affine subspace, steer the LM toward the sampled point, and generate (§4.2). **(c)** To evaluate the quality of the affine subspace, we evaluate the steered outputs for concept relevance via a concept score, and concept coverage via a diversity score (§4.3).

[1]Department of Computer Science, Boston University, Boston, USA [2]Department of Computer Science, Technion – Israel Institute of Technology, Israel [3]Department of Computer Science, Kempner Institute, Harvard University, Boston, USA. Correspondence to: Divya Appapogu <divsp@bu.edu>, Aaron Mueller <amueller@bu.edu>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

2024; Mikolov et al., 2013). However, recent evidence suggests that concepts may exhibit more complex geometric structure: concepts may be polytopes (Park et al., 2025), or may lie on multidimensional (and potentially non-linear) manifolds embedded in the representation space, rather than along one-dimensional linear directions (Engels et al., 2025; Gurnee et al., 2025; Shafran et al., 2026; Tiblias et al., 2026; Karkada et al., 2026; Bhalla et al., 2026). This perspective is especially compelling when considering *multinomial* concepts (Engels et al., 2025; Gurnee et al., 2025; Shafran et al., 2026), whose values cannot be cleanly conceptualized as varying degrees of simply on or off (Sharma et al., 2024; Marks & Tegmark, 2024; Bolukbasi et al., 2016).

In this work, we propose a geometric representation that builds on the success of non-basis-aligned directions for binary concepts. We extend this framework to multinomial concepts by modeling them as *affine subspaces* (Rajendran et al., 2024). Affine subspaces can be viewed both as multidimensional generalizations of directions and as local linear approximations to nonlinear manifolds, and could thus provide a balance between computational tractability and geometric fidelity.

We also propose evaluation metrics and settings for measuring the quality of a multinomial concept representation. Here, quality encompasses both whether the subspace captures only the target concept (analogous to precision) and the coverage of concept it represents (analogous to recall). Given a ground-truth representation (§3.1), we propose to evaluate both criteria using the Grassmannian Distance. In the absence of a reference (§4)—as in most realistic settings—we instead propose metrics based on steering via *sampling* from the space.

In this work, we propose methods for locating affine concept representations, two applications of these representations, and evaluation methods to characterize their quality. While we focus on affine subspaces of LLM activations as a proof of concept, our evaluation framework is geometry-agnostic and can be applied to any proposed concept geometry, including non-linear manifolds. Our contributions are:

- We introduce a toy evaluation framework with known ground-truth concept geometry. Using this setting, we propose and benchmark various methods to construct affine subspaces.
- We develop reference-free evaluation methods for settings where ground-truth geometry is unavailable, using precision and recall to measure how well a recovered affine subspaces in language models.
- We develop a method for sampling diverse concept-related representations within a learned subspace, enabling diverse interventions that allow targeted control over subsets of a concept in LLMs.

2. Preliminaries and Methods

Our goal is to discover a concept in a collection of observed points $\{\boldsymbol{x}^{(i)}\}_{i=1}^{m}$ where all $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$, e.g. activation vectors of a LLM. In this context, a concept is a low-dimensional affine subspace $\mathbb{A}$ that best explains the data[1]. ‘ Formally, we define an affine subspace $\mathbb{A} \subset \mathbb{R}^d$ of the form $\mathbb{A} = \{\boldsymbol{L}\boldsymbol{c} + \boldsymbol{b} \mid \boldsymbol{c} \in \mathbb{R}^k\}$ where $\boldsymbol{L} \in \mathbb{R}^{d\times k}$ is a matrix with orthonormal columns defining a k-dimensional linear subspace, $\boldsymbol{b} \in \mathbb{R}^d$ is an offset vector, and vector $\boldsymbol{c} \in \mathbb{R}^k$ denotes the coefficients in the subspace basis.

We assume that observed points $\boldsymbol{x} \in \mathbb{R}^d$ are generated from this affine subspace with additive Gaussian noise ϵ and fulfill $\boldsymbol{x} = \boldsymbol{L}\boldsymbol{c} + \boldsymbol{b} + \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Our motivation for finding low-dimensional affine subspaces is grounded in the assumption that LLM activations corresponding to a single concept lie on a manifold of intrinsic dimension $k \ll d$, where d is the ambient activation dimension. This assumption is supported by empirical evidence from prior work showing that semantic concepts occupy low-dimensional manifolds embedded in high-dimensional LLM activation spaces (Gurnee et al., 2025; Karkada et al., 2026; Engels et al., 2025).

2.1. Locating Affine Subspaces

To obtain suitable affine concepts for a collection of points, we generalize existing binary or rank-one methods (i.e., methods that identify one-dimensional subspaces) to multinomial or higher-dimensional settings. The first three methods are straightforward extensions of existing one-dimensional concept discovery approaches; we then propose a multidimensional extension of Difference in Means and introduce a method using Schatten Probes. In the following, we briefly summarize these approaches using a unified notation.

Notation. Let $\boldsymbol{X} \in \mathbb{R}^{m\times d}$ denote the matrix with activation vectors $\boldsymbol{x}^{(i)}$ in its rows and $i \in \{1, \dots, m\}$. Formally, $\boldsymbol{X}_{i,:} = (\boldsymbol{x}^{(i)})^\top$ for $i = 1, \dots, m$.

SVD. For any matrix $\boldsymbol{M} \in \mathbb{R}^{a\times b}$, we denote its singular value decomposition (SVD) by

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top,$$

where $\boldsymbol{U} \in \mathbb{R}^{a\times a}$ and $\boldsymbol{V} \in \mathbb{R}^{b\times b}$ are orthogonal matrices containing the left and right singular vectors respectively, and $\boldsymbol{\Sigma} \in \mathbb{R}^{a\times b}$ is a diagonal matrix containing singular values in decreasing order.

[1] Evidence is emerging that concepts are not necessarily linearly encoded (Bhalla et al., 2026; Karkada et al., 2026). We view affine subspaces as linear approximations of concept structure; like other linear approximations in interpretability such as attributions, they sacrifice some accuracy for efficiency at scale.

Principal Component Analysis (PCA; Pearson, 1901). We compute the mean activation vector: $\boldsymbol{\mu}_C = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}^{(i)}$. Next, we center the data by constructing the centered matrix $\tilde{\boldsymbol{X}} \in \mathbb{R}^{m\times d}$ where $\tilde{\boldsymbol{X}}_{i,:} = (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_C)^\top, \quad i = 1, \ldots, m.$

We then compute the SVD of the centered matrix $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. The affine subspace $\mathbb{A}$ is defined by: $\mathbb{A} = \{\boldsymbol{L}\boldsymbol{c} + \boldsymbol{\mu}_C \mid \boldsymbol{c} \in \mathbb{R}^k\}$, where $\boldsymbol{L} \in \mathbb{R}^{d\times k}$ consists of the first k columns of $\boldsymbol{V}$, i.e., $\boldsymbol{L} = [\boldsymbol{v}_1 \mid \boldsymbol{v}_2 \mid \cdots \mid \boldsymbol{v}_k]$ where $\boldsymbol{v}_i$ is the i-th column of $\boldsymbol{V}$ (the i-th right singular vector).

Factor Analysis (FA; Spearman, 1904). Recent work suggests that leveraging local geometry may be more effective for concept localization than assuming a global affine concept representation (Shafran et al., 2026). In this method, we model the generative process as $\boldsymbol{x}^{(i)} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z}^{(i)} + \epsilon^{(i)}$, where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the global mean, $\boldsymbol{L} \in \mathbb{R}^{d\times k}$ is the factor loading matrix, $\boldsymbol{z}^{(i)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ are latent factors, and $\epsilon^{(i)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$ is noise with diagonal covariance matrix $\boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_d)$. We estimate parameters $\{\boldsymbol{L}, \boldsymbol{\Psi}, \boldsymbol{\mu}\}$ using the Expectation-Maximization (EM) algorithm. The recovered affine subspace is $\mathbb{A} = \{\boldsymbol{L}\boldsymbol{c} + \boldsymbol{\mu} \mid \boldsymbol{c} \in \mathbb{R}^k\}$ with basis $\boldsymbol{L}$ and offset $\boldsymbol{\mu}$.

Linear Probe (Alain & Bengio, 2017). We train a linear classifier with parameters $\boldsymbol{W} \in \mathbb{R}^{d\times|C|}$ and $\boldsymbol{b} \in \mathbb{R}^{|C|}$ to predict concept class labels $y^{(i)} \in \{1, \ldots, |C|\}$ from activations $\boldsymbol{x}^{(i)}$ via cross-entropy loss. Performing SVD $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, we extract the top k left singular vectors to form $\boldsymbol{L} = [\boldsymbol{u}_1 \mid \cdots \mid \boldsymbol{u}_k] \in \mathbb{R}^{d\times k}$. We set the offset to the global mean $\boldsymbol{\mu}_C = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}^{(i)}$ (rather than the learned bias $\boldsymbol{b}$) The affine subspace is $\mathbb{A} = \{\boldsymbol{L}\boldsymbol{c} + \boldsymbol{\mu}_C \mid \boldsymbol{c} \in \mathbb{R}^k\}$.

Difference in Means (DiffMean; Marks & Tegmark, 2024). The directions of greatest variance may not necessarily capture the concept's semantic structure (Mueller et al., 2025). We hypothesize that directions separating distinct concept values more accurately characterize the concept's affine geometry. We therefore propose a method based on difference-in-means (Marks & Tegmark, 2024) that captures pairwise differences between concept values.

Let $C = \{c_1, c_2, \ldots, c_{|C|}\}$ denote the set of distinct classes for concept C. For each class $c_j \in C$, let $\mathbb{X}_{c_j} = \{\boldsymbol{x}^{(i)} : \boldsymbol{x}^{(i)} \text{ corresponds to class } c_j\}$ with $|\mathbb{X}_{c_j}| = m_j$ samples. We compute the mean activation vector for each class:

$$\boldsymbol{\mu}_{c_j} = \frac{1}{m_j} \sum_{\boldsymbol{x}\in\mathbb{X}_{c_j}} \boldsymbol{x}.$$

For each pair of classes (c_i, c_j) where $i < j$, we compute the difference vector:

$$\boldsymbol{d}_{i,j} = \boldsymbol{\mu}_{c_i} - \boldsymbol{\mu}_{c_j} \in \mathbb{R}^d.$$

We stack these as columns to form $\boldsymbol{D} = [\boldsymbol{d}_{1,2} \mid \boldsymbol{d}_{1,3} \mid \cdots \mid \boldsymbol{d}_{|C|-1,|C|}] \in \mathbb{R}^{d\times p}$, where $p = \binom{|C|}{2}$ is the number of unique pairs.

We would like to be able to control the rank k of the resulting affine subspace. To do so, we perform SVD on the difference matrix: $\boldsymbol{D} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, and extract the top k right singular vectors to form the basis matrix $\boldsymbol{L} = [\boldsymbol{v}_1 \mid \cdots \mid \boldsymbol{v}_k] \in \mathbb{R}^{d\times k}$. The offset is set to the global mean: $\boldsymbol{\mu}_C = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}^{(i)}$, where $m = \sum_{j=1}^{|C|} m_j$ is the total number of samples. The recovered affine subspace is $\mathbb{A} = \{\boldsymbol{L}\boldsymbol{c} + \boldsymbol{\mu}_C \mid \boldsymbol{c} \in \mathbb{R}^k\}$.

Schatten Probe. (Braun et al., 2025; Giampouras et al., 2020) We also propose an extension to linear probe to encourage low-rank weights through approximate Schatten-p norm regularization. We factorize the weight matrix as $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{B}^\top$, where $\boldsymbol{A} \in \mathbb{R}^{d\times k}$ and $\boldsymbol{B} \in \mathbb{R}^{|C|\times k}$ for chosen rank k. Let $\boldsymbol{a}_i \in \mathbb{R}^d$ and $\boldsymbol{b}_i \in \mathbb{R}^{|C|}$ denote the i-th columns of $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively. We minimize the following objective:

$$\mathcal{L}(\boldsymbol{A}, \boldsymbol{B}) = \text{CrossEntropy}(\boldsymbol{A}\boldsymbol{B}^\top, \text{labels}) + \lambda \cdot \text{reg}(\boldsymbol{A}, \boldsymbol{B}),$$

where the regularizer is defined as:

$$\text{reg}(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{2^p} \sum_{i=1}^{k} \left(\|\boldsymbol{a}_i\|_2^2 + \|\boldsymbol{b}_i\|_2^2\right)^p,$$

where $\lambda > 0$ is a regularization coefficient, and $p < 1$ penalizes non-zero singular values.[2] This regularizer encourages low-rank structure by penalizing paired column norms. After training, we reconstruct the full weight matrix $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{B}^\top$ and perform SVD: $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$.

We select the top k_{eff} left singular vectors based on the effective rank of $\boldsymbol{W}$, where $k_{\text{eff}} \leq k$ is determined by the singular value spectrum. These form the basis matrix $\boldsymbol{L} = [\boldsymbol{u}_1 \mid \cdots \mid \boldsymbol{u}_{k_{\text{eff}}}] \in \mathbb{R}^{d\times k_{\text{eff}}}$. The offset is $\boldsymbol{\mu}_C = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}^{(i)}$, yielding affine subspace $\mathbb{A} = \{\boldsymbol{L}\boldsymbol{c} + \boldsymbol{\mu}_C \mid \boldsymbol{c} \in \mathbb{R}^{k_{\text{eff}}}\}$.

3. Toy Task: Locating Known Affine Subspaces

We first compare which methods best localize known affine subspaces. To test this, we devise a toy setting where the basis vectors and offset of an affine subspace are randomly generated, and then sample vectors from the subspace to form a dataset from which a method must recover the true basis vectors and offset.

3.1. Gaussian Affine Subspaces

We generate a ground-truth affine subspace by randomly sampling a set of orthonormal basis vectors $\boldsymbol{L}_{\text{true}} = [\boldsymbol{v}_1 \mid \cdots \mid \boldsymbol{v}_k] \in \mathbb{R}^{d\times k}$ and a random offset vector

[2] Empirically, we find $p = 0.9$ performs best; we use this value unless otherwise stated.

$\boldsymbol{b}_{\text{true}} \in \mathbb{R}^d$. This defines a ground-truth affine subspace $\mathbb{A}_{\text{true}} = \{\boldsymbol{L}_{\text{true}}\boldsymbol{c} + \boldsymbol{b}_{\text{true}} \mid \boldsymbol{c} \in \mathbb{R}^k\}$.

We generate m synthetic data points from the ground-truth affine subspace with added noise. We sample latent coefficients $\boldsymbol{c}^{(i)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ for $i = 1, \ldots, m$, where $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ denotes a standard isotropic k-dimensional Gaussian with zero mean and unit variance in each dimension. We generate points as $\boldsymbol{x}^{(i)} = \boldsymbol{b}_{\text{true}} + \sigma \boldsymbol{L}_{\text{true}}\boldsymbol{c}^{(i)} + \epsilon^{(i)}$, where $\epsilon^{(i)} \sim \mathcal{N}(\boldsymbol{0}, \eta^2 \boldsymbol{I}_d)$ is isotropic d-dimensional Gaussian noise with variance η^2 per dimension, controlling the off-subspace perturbation. The parameter σ scales the spread of points along the subspace directions, while η controls how far points deviate from the subspace. Without noise ($\eta = 0$), all points lie exactly on the affine subspace and recovery is trivial via least-squares (see Appendix A.1); adding noise makes the problem non-trivial.

To simulate a multinomial concept with $|C|$ distinct values, we sample points in $|C|$ clusters corresponding to different concept classes. For each class $c_j \in C$ where $j \in \{1, \ldots, |C|\}$, we first sample a cluster center by drawing coefficients $\boldsymbol{c}_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ from the standard Gaussian and setting $\boldsymbol{\mu}_j = \boldsymbol{b}_{\text{true}} + \sigma \boldsymbol{L}_{\text{true}}\boldsymbol{c}_j$. This places cluster centers at random locations along the ground-truth subspace. We then generate m_j points around each cluster center (where $\sum_{j=1}^{|C|} m_j = m$) as $\boldsymbol{x}^{(j,i)} = \boldsymbol{\mu}_j + \rho \boldsymbol{L}_{\text{true}}\boldsymbol{c}^{(j,i)} + \epsilon^{(j,i)}$, where $\boldsymbol{c}^{(j,i)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ and $\epsilon^{(j,i)} \sim \mathcal{N}(\boldsymbol{0}, \eta^2 \boldsymbol{I}_d)$ for $i = 1, \ldots, m_j$. Here $\rho \ll \sigma$ controls the within-cluster spread along the subspace points within a cluster vary with small scale ρ around their cluster center, while cluster centers themselves are spread with larger scale σ. The off-subspace noise variance η^2 applies uniformly to all points.

We apply each of the discovery methods described in §2 to the synthetic dataset $\{\boldsymbol{x}^{(i)}\}_{i=1}^m$ with class labels $\{y^{(i)}\}_{i=1}^m$, where $y^{(i)} \in \{1, \ldots, |C|\}$ denotes the cluster/class that point i was generated from. Each method recovers an affine subspace $\mathbb{A}_{\text{recovered}} = \{\boldsymbol{L}_{\text{recovered}}\boldsymbol{c} + \boldsymbol{b}_{\text{recovered}} \mid \boldsymbol{c} \in \mathbb{R}^{k'}\}$, where k' is the estimated subspace dimension (which may differ from the ground-truth dimension k). The rank selection criterion varies by method: for PCA, we select k' as the smallest dimension capturing 95% of the variance in the centered data matrix $\tilde{\boldsymbol{X}}$; for difference-in-means, we select k' capturing 95% of the variance in the difference matrix $\boldsymbol{D}$; for linear probe, we select k' capturing 95% of the variance in the weight matrix $\boldsymbol{W}$; for factor analysis we use the effective rank of the loadings matrix $\boldsymbol{L}$; and for the Schatten probe, we use the effective rank of the learned factorization $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{B}^\top$.

Grassmann Distance. To evaluate the quality of recovery in this controlled setting, we use the affine Grassmann distance to measure the distance between the recovered affine subspace $\mathbb{A}_{\text{recovered}}$ and the ground-truth affine subspace $\mathbb{A}_{\text{true}}$. The Grassmann distance is a natural extension of the Grassmannian manifold from linear subspaces to affine subspaces, and accounts for both differences in orientation and differences in offset. Given two affine subspaces $\mathbb{A} + b$ and $\mathbb{B} + c$ of dimensions k and k', respectively, the affine Grassmann distance is defined as:

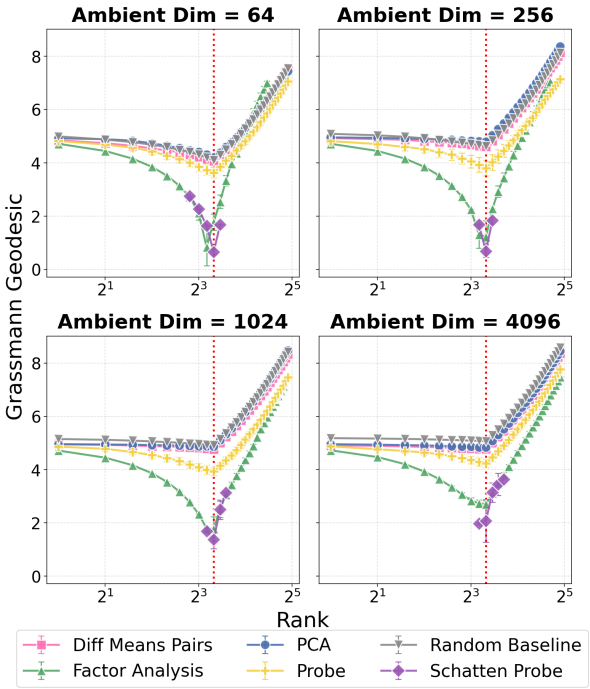

Figure 2. **Rank controlled toy setup.** We vary the rank from 1 to 30 for each recovery method and report the Grassmann distance to the ground-truth 10-dimensional affine subspace (indicated by the dashed vertical red line). Results are averaged over 5 random seeds with standard deviations on the order of 10^{-1} or lower. Here, the ground-truth parameters are $\sigma = 0.5$ (spread of points along the affine subspace) and $\eta = 0.5$ (deviation from the affine subspace). See Fig. 5 (App. B) for complete results.

$$d_{\text{Graff}(\infty,\infty)}(\mathbb{A} + b,\, \mathbb{B} + c) = \left(\frac{\pi^2}{4}|k - k'| + \sum_{i=1}^{\min(k+1,\, k'+1)} \theta_i^2 \right)^{1/2} \quad (1)$$

where $\theta_i \in [0, \frac{\pi}{2}]$ are the affine principal angles between $\mathbb{A} + b$ and $\mathbb{B} + c$, computed from the singular values of $Y_{\mathbb{A}+b}^T Y_{\mathbb{B}+c}$ where $Y_{\mathbb{A}+b}$ and $Y_{\mathbb{B}+c}$ are the Stiefel coordinate representations.

The Stiefel coordinates for an affine subspace $\mathbb{A} + b \in \text{Graff}(k, n)$ are constructed from orthogonal affine coordinates $[\boldsymbol{A}, b_0]$, where $\boldsymbol{A} \in \mathbb{R}^{n \times k}$ is an orthonormal basis matrix for the linear subspace $\mathbb{A}$ (i.e., $\mathbb{A} = \text{span}(\boldsymbol{A})$ and $\boldsymbol{A}^\top \boldsymbol{A} = I_k$), and $b_0 \in \mathbb{R}^n$ is the displacement vector orthogonal to $\mathbb{A}$ (i.e., $\boldsymbol{A}^\top b_0 = 0$). The Stiefel coordinates

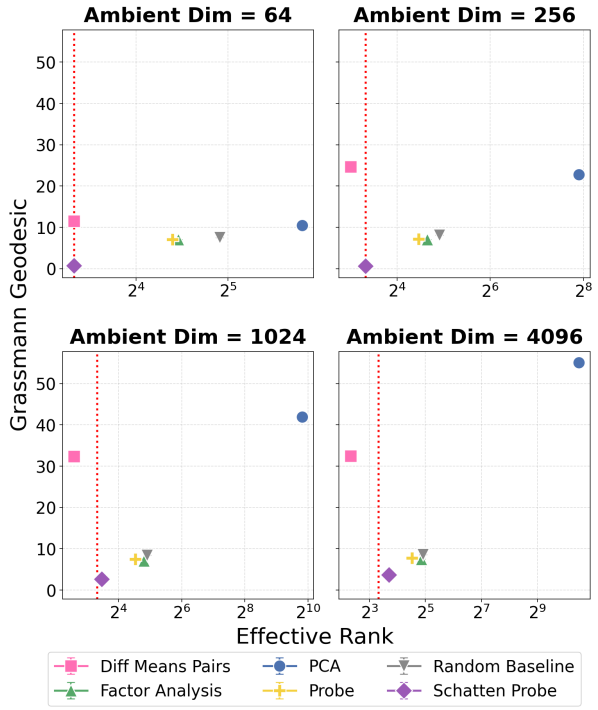

Figure 3. **Rank estimation for toy setup.** We evaluate each method's ability to estimate the rank of the ground-truth 10-dimensional subspace (dashed vertical red line). For PCA, linear probes, and DiffMean, rank is estimated using 95% variance explained on $\tilde{\boldsymbol{X}}$, $\boldsymbol{W}$, and $\boldsymbol{D}$ respectively. For Schatten probe and factor analysis, we use the effective rank of the weight matrix $\boldsymbol{W}$ and loading matrix $\boldsymbol{L}$, respectively. Results are averaged over 5 random seeds; standard deviations are on the order of 10^{-1} or lower. The ground truth parameters are $\sigma = 0.5$ (spread of points along the affine subspace) and $\eta = 0.5$ (deviation from the affine subspace). See Fig 6 (App. B) for full results.

are:

$$Y_{\mathbb{A}+b} = \begin{bmatrix} \boldsymbol{A} & b_0/\sqrt{1+\|b_0\|^2} \\ 0 & 1/\sqrt{1+\|b_0\|^2} \end{bmatrix} \in \mathbb{R}^{(n+1)\times(k+1)} \quad (2)$$

Lower values of $d_{\mathrm{Graff}(\infty,\infty)}$ reflect a closer match between $\mathbb{A}_{\text{recovered}}$ and $\mathbb{A}_{\text{true}}$. We choose this metric because it can handle dimensionality mismatches where $k' \neq k$. To minimize the distance, a method must accurately recover the true subspace's intrinsic dimension k, the correct basis directions ($\boldsymbol{L}_{\text{recovered}} \approx \boldsymbol{L}_{\text{true}}$), and the correct offset ($\boldsymbol{b}_{\text{recovered}} \approx \boldsymbol{b}_{\text{true}}$). See App. B.1 for a more detailed intuitive description of the metric and related metrics.

Results. We first plot rank-controlled results for our methods and a random affine subspace as a baseline. We observe (Fig. 2) that, at many ranks, methods like PCA and DiffMean are no better than the random baseline; meanwhile, Schatten probes and factor analysis perform significantly better. How well does each method select the correct rank? When selecting a single rank per method,[3] we observe (Fig. 3) that Schatten probes are best at recovering the true affine subspace; they also often recover subspaces of a dimensionality closest to the true dimensionality. Linear probes tend to recover subspaces with dimensionalities larger than the true dimensionality. PCA generally recovers dimensionality close to the ambient dimensionality. DiffMean often under-predicts the rank. Given these results, we hypothesize that Schatten probes will perform best when scaling these methods to large language model activations. If the ranking of methods significantly differs between this setting and LLM activations, this could be indirect evidence that affine subspaces are not faithful approximations of concept representations.

4. Natural Language Experiments

In the LLM setting, a point $\boldsymbol{x}$ is a hidden activation for a given token. We would like to discover an affine subspace of the activations that both (i) captures all values a particular concept can take, and (ii) only includes concept-related points. We would also like to evaluate the quality of the different methods.

However, unlike in the toy setting where ground-truth geometry is known, we do not have access to the true concept structure in language models trained on natural data; we therefore cannot use the Grassmann distance. Thus, we propose a sampling-based reference-free evaluation metric to assess the quality of discovered affine subspaces. Our goal is to measure two aspects of subspace quality. First, we measure the **concept score** (Wu et al., 2025; analogous to *precision*); this measures whether points sampled from the recovered subspace correspond to the target concept. Second, we propose a **diversity score** (analogous to *recall*); this measures whether the subspace captures the full range of values a concept can take.

4.1. Measuring Concept and Diversity Scores via Sampling and Steering

Given a recovered affine subspace $\mathbb{A}_{\text{recovered}} = \{\boldsymbol{L}\boldsymbol{c} + \boldsymbol{b} \mid \boldsymbol{c} \in \mathbb{R}^k\}$, we sample n points $\{\boldsymbol{p}^{(i)}\}_{i=1}^n$ from the subspace by drawing coefficients $\boldsymbol{z}^{(i)} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_k)$ and computing $\boldsymbol{p}^{(i)} = \boldsymbol{L}\boldsymbol{z}^{(i)} + \boldsymbol{b}$. Since affine subspaces are unbounded, we select σ empirically to ensure sampled points remain within the typical range of the model's activation space.

To test whether the sampled directions correspond to a target concept (e.g., animals), we intervene on the hidden representations of a language model and evaluate whether the generations under this intervention remain within the concept space. More details about interventions are presented in §4.2.

[3] We describe how the ranks are selected in §2.

To evaluate them, for a set of intervention-generated samples for a given concept, we compute (i) a *concept score* (AXBENCH, Wu et al. (2025)), and (ii) a *diversity score*. We introduce the latter to quantify the range of distinct concept realizations. Each score is an integer in $[0, 2]$. High concept scores indicate that the sampled directions induce coherent generations related to the target concept, which itself suggests that the recovered affine subspace aligns well with the underlying concept representation in the language model (i.e., has high precision). High diversity scores indicate that the samples correspond to semantically diverse values, which indicates that the space more fully captures the concept (i.e., has high recall). To compute these scores, we use an LLM judge; see Sec 4.3 and an overview of sampling can be seen in Fig 4.

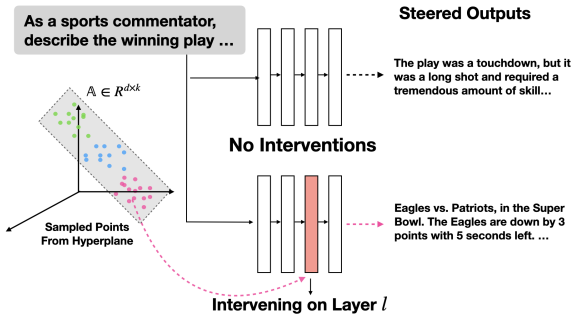

Figure 4. **Sampling Evaluation.** LLM outputs before and after intervention with samples from the affine subspace. Interventions steer the model to generate concept-related content. Additional qualitative results are in App. F (Tables 7–10).

4.2. Experimental Setting

Datasets. We construct datasets for four categorical concepts: ANIMALS, PLANTS, VEHICLES, COUNTRIES. Each categorical dataset is organized into subcategories. For example, ANIMALS are partitioned into six subcategories: `mammals`, `reptiles`, `birds`, `fish`, `amphibians`, and `insects`, with 5 representative instances per subcategory, yielding 30 concept values in total.

For each instance, we generate a sentence describing its attributes, and then prompt the model to predict the target instance. To generate descriptive attributes for each concept value, we use Claude Opus 4.6 (Anthropic, 2025), prompting it to produce 100 factual descriptions per instance. For example, here is one factual description for `rabbit` (in the `mammals` subcategory of the ANIMALS dataset):

Animals Activation Extraction Prompt

```
The Rabbit is a small mammal
belonging to the family Leporidae
and the order Lagomorpha, which also
includes hares and pikas.
Animal:
```

Further details on the datasets are in App. C. Further examples of prompts are in App. G.7.

Extracting activations. All experiments are conducted on the instruction-tuned variants of Llama-3.1-8B (Grattafiori et al., 2024) and Gemma-2-2B (Team et al., 2024). We extract the activation vector from the final token position (immediately before the model predicts the concept value). Intuitively, we expect the representation at this position to contain the most relevant contextual information needed to predict the concept token. These activation vectors serve as inputs to all methods considered in this work. The prediction target is the specific concept label (e.g., *beaver*, *lion* in the ANIMALS dataset). Our goal is to construct low-dimensional affine subspaces that capture the concept's structure.

Steering toward sampled points. To evaluate how well the affine subspace represents the concept, we prompt a model with concept-unrelated prompts from AlpacaEval (Dubois et al., 2024), and intervene on the activation vector. We randomly select 5 prompts and use this fixed set across all experiments; see Tab. 4 (App. C.1). To fully take advantage of the diversity of the affine subspace, we sample points from the recovered affine subspace. For each generation, we sample a point $\boldsymbol{s} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_k)$ from the affine subspace via $\boldsymbol{s} = \boldsymbol{b} + \boldsymbol{L}\boldsymbol{z}$ where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_k)$. We sample a number of points proportional to the number of concept classes in the dataset, generating one steered completion per sample. At each generation step, we modify the hidden representation by adding the sampled steering vector: $\boldsymbol{h}' = \boldsymbol{h} + \alpha \boldsymbol{s}$, where $\alpha \in \mathbb{R}^+$ controls the steering strength. To avoid overly aggressive perturbations that lead to degenerate outputs, we decay α exponentially over the course of generation (details in App. C.2). The sampled point $\boldsymbol{s}$ determines the steering direction, while α controls the intervention magnitude, ensuring the model maintains fluency. Here we do not consider α to be a specific direction as in one dimensional steering, rather we use this as a hyper parameter to control the steering strength.

4.3. Evaluating Generated Samples via LLM-Judges

Following prior work (Lee et al., 2025), we use Mistral-Large as an LLM-based judge to evaluate outputs generated by our sampling procedure.

Table 1. Concept scores, diversity scores, and their harmonic means for each multinomial concept. All sampling methods outperform prompting with respect to both the concept score and diversity score. The best sampling method depends on the concept and model. For Llama, DiffMean and factor analysis perform best. For Gemma, DiffMean and both probes perform well.

		ANIMALS			VEHICLES			PLANTS			COUNTRIES		
Model	Method	Concept	Diversity	Harmonic	Concept	Diversity	Harmonic	Concept	Diversity	Harmonic	Concept	Diversity	Harmonic
Llama-3.1-8B	Prompting	0.400	1.479	0.630	0.226	0.948	0.366	0.270	1.628	0.463	0.776	1.251	0.958
	DiffMean	0.620	1.731	0.913	0.867	**1.722**	1.153	0.520	1.641	0.790	1.240	1.696	1.432
	FA	0.713	1.807	1.023	**1.013**	1.663	**1.259**	**0.580**	1.605	**0.852**	**1.296**	**1.837**	**1.520**
	PCA	0.627	**1.852**	0.936	0.760	1.687	1.048	0.540	1.607	0.808	1.136	1.757	1.380
	Probe	0.720	1.795	1.028	0.680	1.603	0.955	0.460	**1.714**	0.725	1.136	1.724	1.369
	Schatten Probe	**0.760**	1.840	**1.076**	0.840	1.652	1.114	0.500	1.653	0.768	1.200	1.752	1.424
Gemma-2-2B	Prompting	0.333	0.858	0.480	0.040	0.200	0.006	0.130	0.371	0.192	0.533	0.784	0.642
	DiffMean	**0.647**	1.748	**0.944**	**0.973**	**1.840**	**1.273**	**0.690**	1.716	**0.984**	0.792	1.715	1.084
	FA	0.307	**1.861**	0.527	0.467	1.535	0.716	0.400	**1.771**	0.653	0.376	1.476	0.599
	PCA	0.507	1.777	0.789	0.800	1.737	1.095	0.640	1.846	0.950	0.608	1.558	0.875
	Probe	0.527	1.684	0.802	0.800	1.763	1.101	0.510	1.757	0.791	**0.800**	**1.800**	**1.108**
	Schatten Probe	0.493	1.814	0.776	**0.973**	1.791	1.261	0.480	1.765	0.755	0.760	1.700	1.050

Concept Score. We adapt the **concept score** of AXBENCH (Wu et al., 2025) for our multinomial concepts. The concept score measures how well a generated output reflects the intended concept or steering direction; see App. G for LLM-judge prompts.

Diversity Score. We introduce a *diversity score* to quantify how well the recovered affine subspace covers the intended concept. The score is computed in two stages: (i) We use an LLM to extract relevant entities from each generated sentence (e.g., `cat`, `sparrow`), along with their associated category labels (e.g., `mammal`, `bird` for the ANIMALS dataset). (ii) We then aggregate the extracted entities and categories across samples and compute the entropy (Shannon, 1948) at two levels: 1. Entity-level: captures the number and distribution of distinct instances (e.g., specific animals), measuring coverage within a concept. 2. Category-level: captures how broadly the samples span higher-level groupings, measuring whether the subspace covers diverse subcategories of the concept. For concepts with well-defined categorical structure, we provide the category schema explicitly to the LLM judge to ensure consistency; see App. G for examples. We compute the Diversity Score at the instance level, and then separately at category level. The diversity score is the sum of these two components; it lies in $[0, 2]$. We also prompt the LLM judge to record all entities that were not present in the training distribution for each concept (the Unseen score); see App. G.5 for the prompt, and Tab. 5 and 6 for results.

The diversity score captures the extent to which the recovered affine subspace covers the intended concept. A higher entity-level score indicates broad coverage across individual instances; a higher category-level score indicates that the subspace spans diverse subcategories rather than collapsing onto a narrow region of the concept.

4.4. Results

We report the concept score and diversity score in Tab. 1. We include a prompting-based baseline where we prepend an instruction to mention the concept in the output before the AlpacaEval instruction. We ensure that the numbers of generated samples across all methods and concepts match.

For Llama-3.1-8B-Instruct, Factor Analysis achieves the best overall performance, having the highest harmonic mean between concept and diversity scores for the VEHICLES, PLANTS, and COUNTRIES concepts. For ANIMALS, the Schatten probe performs best. DiffMean is never best, but is generally competitive with Schatten probes.

For Gemma-2-2B-IT, trends differ. DiffMean performs best for ANIMALS, VEHICLES, and PLANTS, indicating that simple class-separating directions are often sufficient in these concepts. However, for COUNTRIES, the linear probe achieves the best harmonic mean.

Across all datasets and models, steering methods almost always outperform simple prompting on *both* the concept and diversity scores. Wu et al. (2025) found that prompting was often highly performant for binary concepts.

Qualitative meta-evaluation. We manually examined a subset of LLM judge outputs to validate the reliability of our evaluation pipeline. We found that evaluation quality improved significantly when processing outputs individually rather than in batches. Specifically, rather than asking the LLM to rate concept scores for multiple sentences simultaneously or extract entities from multiple outputs at once, we query the judge once per sentence for concept scoring and once per sampled output for entity extraction. This approach yields more consistent and accurate judgments. We then compute scores per method and per rank across different *Alpaca Eval* prompts. Entity-level and category-level extractions for the Diversity Score are performed at

the sentence level, but the final entropy is computed over all prompts for a given method and rank combination. We report the harmonic mean of the overall Diversity Score and Concept Score as our primary evaluation metric.

5. Related Work

Concept geometry. Early concept geometry studies operated under the Linear Representation Hypothesis (Park et al., 2023), which holds that concepts are linear (and typically one-dimensional) subspaces of LM activations. More recent work has proposed that concepts may be multidimensional (Gurnee et al., 2025; Engels et al., 2025; Park et al., 2025; 2026). This structure can be non-linear, and includes geometries such as convex hulls (Fel et al., 2026), polytopes (Park et al., 2025), vector magnitudes (rather than directions; Csordás et al., 2024), or curved manifolds (Bhalla et al., 2026). Affine subspaces lie at an intermediate point of complexity: they generalize beyond single dimensions, but retain the simplicity and tractability of linear methods.

Evaluating concept representations. Given a concept representation, how can one verify its quality? Prior work has proposed correlational and causal metrics for evaluating their quality. Some evaluate geometric goodness-of-fit (Tiblias et al., 2026), whereas others evaluate counterfactual interventions on downstream behavior (Engels et al., 2025; Mueller et al., 2025; Shafran et al., 2026; Wu et al., 2025; Lee et al., 2026). Others still use toy settings with known ground truths (Gupta et al., 2024). These works evaluate whether a concept was found, but not whether the full concept space was covered; our diversity score fills this gap.

Multi-dimensional concept steering. Most steering approaches assume that concepts correspond to single directions in activation space, and thus intervene by adding a scaled steering vector to the hidden state (Li et al., 2023; Zou et al., 2025; Turner et al., 2025; Subramani et al., 2022; Marks & Tegmark, 2024). Engels et al. (2025) instead steer by replacing the angular coordinate corresponding to one value (e.g., Monday) with another in a circular latent space, and then reconstruct the activation; this alters the model's predictions on periodic variables like days of the week. Shafran et al. (2026) propose steering based on a Mixture of Factor Analyzers (MFA) decomposition of activation space, which models activations as belonging to local Gaussian regions with associated low-dimensional subspaces. This representation enables two forms of intervention: steering activations toward a component centroid to move the model toward a concept region, and modifying the coordinates within the component's local subspace to control finer semantic variations. Our work builds on growing recognition that concepts in language models often exhibit intrinsically multi-dimensional structure. We propose a new evaluation framework to evaluate steering on multinomial concepts, which has only preliminarily been explored in prior work (though see concurrent work by Wurgaft et al., 2026).

5.1. Discussion and Conclusions

We have proposed methods for recovering affine approximations of multinomial concepts from language model activations, as well as metrics and methods for evaluating the quality of multinomial concept representations. In a toy task, Schatten probes perform well, but in realistic settings, DiffMean, Factor Analysis, and the Schatten Probe perform well—and significantly outperform a prompting-based approach.

A benefit of multidimensional concept representations is that they enable more diverse steering than one-dimensional representations; one can sample diverse points from the concept space if one does not mind which value of the concept is chosen. Meaningful higher-rank spaces could enable more fine-grained control in monitoring and steering methods by giving one more control over the level of granularity used in defining a concept.

The relative ranking of methods differed significantly between the toy setting (§3.1) and natural language setting (§4). This could be indirect evidence that affine subspaces are not the most faithful abstraction over concepts. Given recent findings (Bhalla et al., 2026), this is not necessarily surpring. Nonetheless, high scores in the sampling evaluation reveal that these affine approximation may often be sufficient for practical purposes. As our evaluation framework is applicable to any steering setup, future work should directly compare affine subspaces to other geometries.

Limitations

We focus primarily on affine methods. Recent studies have suggested that concepts may be better characterized by more general manifold geometries (Karkada et al., 2026; Bhalla et al., 2026). In theory, our evaluations are compatible with any concept localization method, so future work should compare the performance of affine subspaces to non-linear representations, such as piecewise linear manifolds or continuous manifolds. Another limitation is that we focus on categorical concepts with no internal structure; our understanding could be strengthened by additionally investigating ordinal, periodic, or continuous concepts.

Acknowledgments

This material is supported by the National Science Foundation under Grant No. 2530728 and the U.S.-Israel Binational Science Foundation (BSF) under Grant No. 2025670. This

research was partly supported by the Israel Science Foundation (grant No. 2942/25) and the European Union (ERC, Control-LM, 101165402). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the National Science Foundation, U.S.-Israel Binational Science Foundation, European Union, or the European Research Council Executive Agency. Neither the European Union nor the granting authorities can be held responsible for them. FB was supported by the Swiss National Science Foundation (SNSF) under Grant 239054.

References

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

Anthropic. Claude (claude-sonnet-4-6). https://www.anthropic.com, 2025. Large language model.

Bhalla, U., Fel, T., Rager, C., Feucht, S., Haklay, T., Wurgaft, D., Boppana, S., Kowal, M., Shyam, V., Merullo, J., Geiger, A., and Lubana, E. S. Do sparse autoencoders capture concept manifolds?, 2026. URL https://arxiv.org/abs/2604.28119.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Braun, D., Bushnaq, L., Heimersheim, S., Mendel, J., and Sharkey, L. Interpretability in parameter space: Minimizing mechanistic description length with attribution-based parameter decomposition, 2025. URL https://arxiv.org/abs/2501.14926.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., and et al, A. A. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Csordás, R., Potts, C., Manning, C. D., and Geiger, A. Recurrent neural networks learn to store and generate sequences using non-linear representations. In *The 7th BlackboxNLP Workshop*, 2024. URL https://openreview.net/forum?id=NUQeYgg8x4.

Dubois, Y., Liang, P., and Hashimoto, T. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Engels, J., Michaud, E. J., Liao, I., Gurnee, W., and Tegmark, M. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Fel, T., Wang, B., Lepori, M. A., Kowal, M., Lee, A., Balestriero, R., Joseph, S., Lubana, E. S., Konkle, T., Ba, D. E., and Wattenberg, M. Into the rabbit hull: From task-relevant concepts in DINO to minkowski geometry. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=pbbRmnFkjG.

Giampouras, P., Vidal, R., Rontogiannis, A., and Haeffele, B. A novel variational form of the schatten-p quasi-norm. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21453–21463. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f53eb4122d5e2ce81a12093f8f9ce922-Paper.pdf.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A.,

Hinsvark, A., and et al, A. R. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Gupta, R., Arcuschin, I., Kwa, T., and Garriga-Alonso, A. Interpbench: Semi-synthetic transformers for evaluating mechanistic interpretability techniques. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=R9gR9MPuD5.

Gurnee, W., Ameisen, E., Kauvar, I., Tarng, J., Pearce, A., Olah, C., and Batson, J. When models manipulate manifolds: The geometry of a counting task. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/linebreaks/index.html.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Karkada, D., Korchinski, D. J., Nava, A., Wyart, M., and Bahri, Y. Symmetries in language statistics shape the geometry of model representations. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=XQeMPEkfdd.

Lee, A., Belinkov, Y., Viégas, F., and Wattenberg, M. Decomposing query-key feature interactions using contrastive covariances. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=t5cfrJfuzD.

Lee, Y., Kim, J., Kim, J., Cho, H., Kang, J., Kang, P., and Kim, N. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15771–15798, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.796. URL https://aclanthology.org/2025.emnlp-main.796/.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

Lim, L.-H., Wong, K. S.-W., and Ye, K. The grassmannian of affine subspaces. *Foundations of Computational Mathematics*, 21(2):537–574, Apr 2021. ISSN 1615-3383. doi: 10.1007/s10208-020-09459-8. URL https://doi.org/10.1007/s10208-020-09459-8.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

Mueller, A., Geiger, A., Wiegreffe, S., Arad, D., Arcuschin, I., Belfki, A., Chan, Y. S., Fiotto-Kaufman, J. F., Haklay, T., Hanna, M., Huang, J., Gupta, R., Nikankin, Y., Orgad, H., Prakash, N., Reusch, A., Sankaranarayanan, A., Shao, S., Stolfo, A., Tutek, M., Zur, A., Bau, D., and Belinkov, Y. MIB: A mechanistic interpretability benchmark. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=sSrOwve6vb.

Nanda, N., Lee, A., and Wattenberg, M. Emergent linear representations in world models of self-supervised sequence models. In Belinkov, Y., Hao, S., Jumelet, J., Kim, N., McCarthy, A., and Mohebbi, H. (eds.), *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.2. URL https://aclanthology.org/2023.blackboxnlp-1.2/.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=T0PoOJg8cK.

Park, K., Choe, Y. J., Jiang, Y., and Veitch, V. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bVTM2QKYuA.

Park, K., Nief, T., Choe, Y. J., and Veitch, V. The information geometry of softmax: Probing and steering. In *Forty-third International Conference on Machine Learning*,

2026. URL https://openreview.net/forum?id=BrIGLICBWy.

Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11): 559–572, 1901.

Rajendran, G., Buchholz, S., Aragam, B., Schölkopf, B., and Ravikumar, P. K. From causal to concept-based representation learning. In *Causality and Large Models @NeurIPS 2024*, 2024. URL https://openreview.net/forum?id=FcVnIBYbkW.

Shafran, O. D., Ronen, S., Fahn, O., Ravfogel, S., Geiger, A., and Geva, M. From directions to regions: Decomposing activations in language models via local geometry. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Tz7n7pX9cO.

Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., DURMUS, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S. M., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., Nathan, A., Luo, A., Helyar, A., and et al, A. M. Openai gpt-5 system card, 2025. URL https://arxiv.org/abs/2601.03267.

Spearman, C. " general intelligence" objectively determined and measured. 1904.

Subramani, N., Suresh, N., and Peters, M. Extracting latent steering vectors from pretrained language models. In Muresan, S., Nakov, P., and Villavicencio, A. (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/.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., and et al, A. T. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Tiblias, F., Bigoulaeva, I., Niu, J., Balloccu, S., and Gurevych, I. Hypothesis-driven feature manifold analysis in LLMs via supervised multi-dimensional scaling. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=vCKZ40YYPr.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering, 2025. URL https://openreview.net/forum?id=2XBPdPIcFK.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Wu, Z., Arora, A., Geiger, A., Wang, Z., Huang, J., Jurafsky, D., Manning, C. D., and Potts, C. Axbench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=K2CckZjNy0.

Wurgaft, D., Rager, C., Kowal, M., Shyam, V., Feucht, S., Bhalla, U., Haklay, T., Bigelow, E., Sarfati, R., McGrath, T., Lewis, O., Merullo, J., Goodman, N., Fel, T., Geiger, A., and Lubana, E. S. Manifold steering reveals the shared geometry of neural network representation and behavior. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=QklBm5nnR9.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A. Recovering an Affine Subspace

A.1. Trivial Solution

If we have $k+1$ points $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\} \subset \mathbb{R}^n$ in general position, we recover a k-dimensional affine subspace $\mathbb{A}$ as follows. We designate $\boldsymbol{x}_0$ as the offset, setting $\boldsymbol{b} = \boldsymbol{x}_0$, and form k direction vectors by subtracting the offset from the remaining points,$\boldsymbol{d}_i = \boldsymbol{x}_i - \boldsymbol{x}_0, \quad i = 1, \ldots, k$. The vectors $\{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_k\}$ are linearly independent. We then orthonormalize them via the Gram-Schmidt procedure to obtain an orthonormal basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$, and collect them as columns of $\boldsymbol{V} = [\boldsymbol{v}_1 \mid \cdots \mid \boldsymbol{v}_k] \in \mathbb{R}^{n \times k}$. The recovered k-dimensional affine subspace is then $\mathbb{A} = \left\{\boldsymbol{b} + \boldsymbol{V}\boldsymbol{c} \mid \boldsymbol{c} \in \mathbb{R}^k\right\}$,

where any point $\boldsymbol{x} \in \mathbb{A}$ can be written as $\boldsymbol{x} = \boldsymbol{x}_0 + \sum_{i=1}^{k} c_i \boldsymbol{v}_i$ for coefficients $\boldsymbol{c} = (c_1, \ldots, c_k)^\top \in \mathbb{R}^k$. In real-world settings, we do not have access to which points belong to the underlying subspace, so the trivial solution is not applicable. For example, when retrieving activations corresponding to a specific concept from a language model, these activations do not lie cleanly on an affine subspace. If they did, recovering the subspace would be trivial. Instead, we assume that the activations are corrupted by noise, so the points only approximately lie on an affine subspace. Consequently, recovering the underlying affine subspace becomes a non-trivial problem.

To mimic this setting, we introduce a toy setup where we systematically vary the noise conditions. As described in Section 3.1, we control both the noise scale and noise level. Specifically, we evaluate noise scales and levels in $0.1, 0.3, 0.5, 0.7, 0.9$ and compare the methods introduced in Section 2. The results are shown in Fig. 7. In this setup, the true dimension of the affine subspace is 10, while the ambient dimension is 4906, mimicking typical LLM representations.

We observe that the Schatten probe consistently recovers a dimension close to the ground-truth value. Correspondingly, it achieves a low Grassmann distance. In contrast, DiffMeans systematically underestimates the rank of the affine subspace, while other methods overestimate the rank. Additional details on the Grassmann metric are provided in App. B.1.

B. Further Toy Task Results

Here, we present more detailed results on the toy task. Figure 6 shows results when selecting a single rank, following the methods in §2. Figure 5 shows results for each method when varying the rank. Figure 7 shows results when varying the degree of noise while generating points from the ground-truth affine subspace.

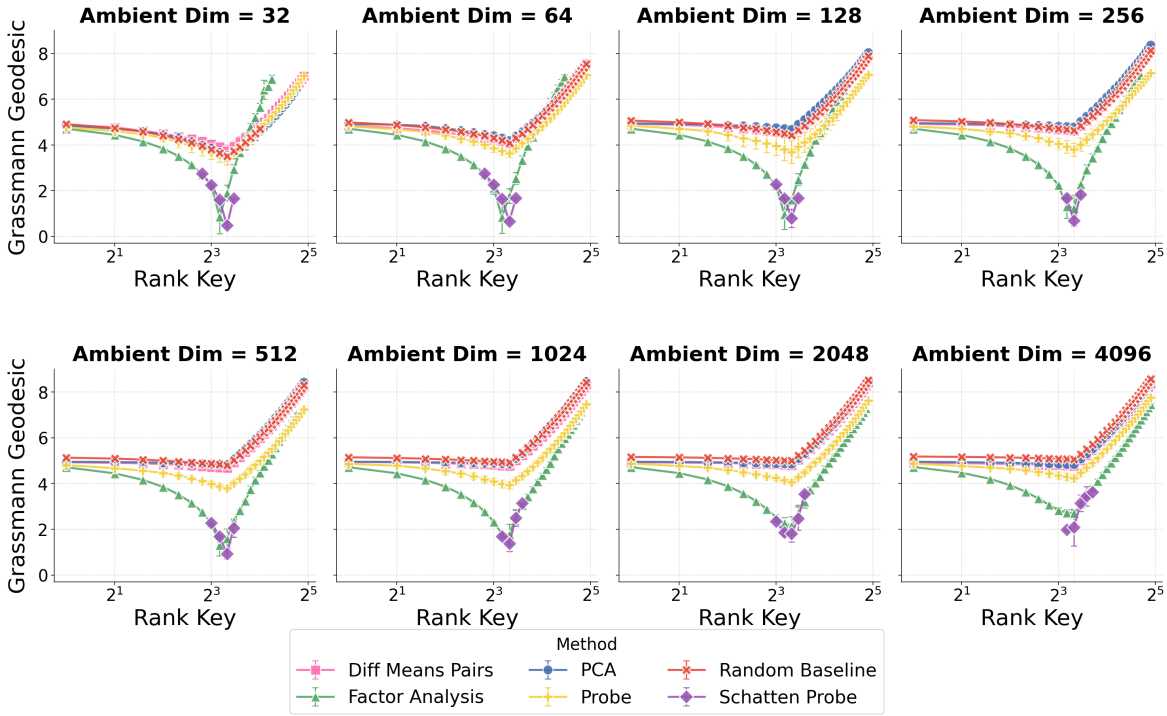

Figure 5. Toy task methods with varying ranks across different ambient dimensions. We vary the rank from 1 to 30 for each recovery method and report the Grassmann distance to the ground-truth 10-dimensional affine subspace (indicated by the dashed vertical red line). Results are averaged over 5 random seeds with standard deviations on the order of 10^{-1} or lower.

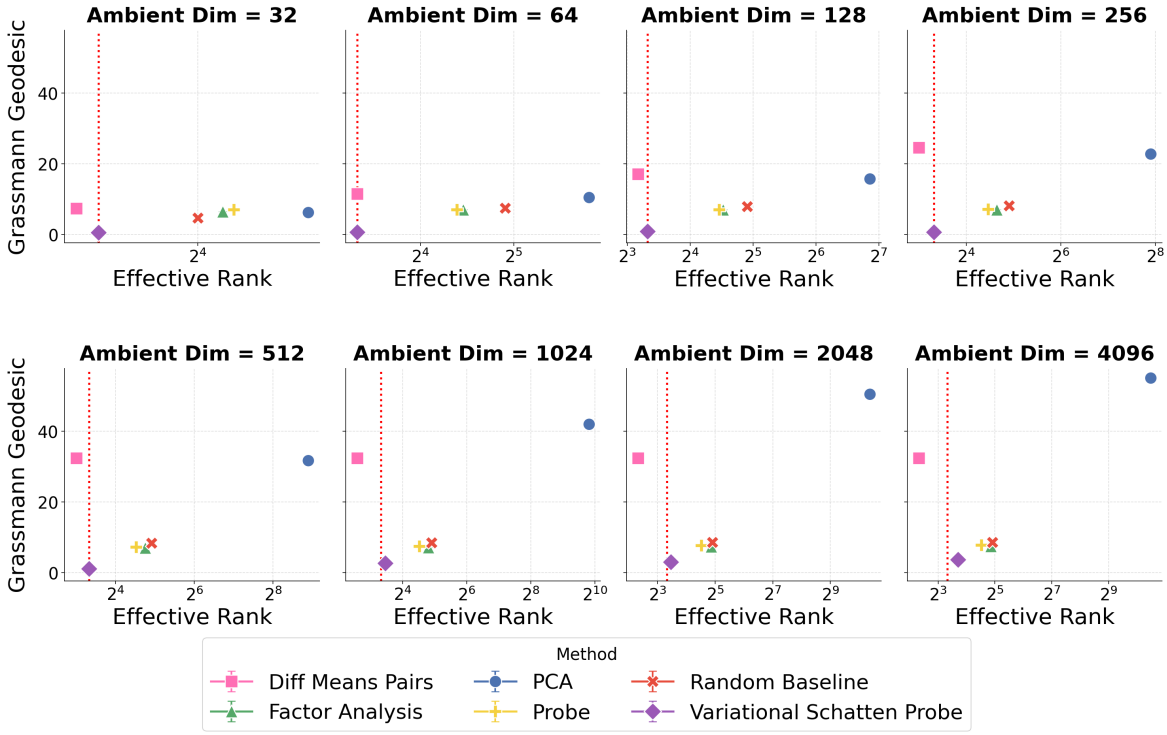

Figure 6. Results for toy setup. The true dimensionality of the ground-truth plane is indicated by a dashed vertical red line. We tune the rank of each method and report the best point (with the lowest Grassmann distance) only. We run 5 different seeds; the standard deviation is in the order of 10^{-1} or lower.

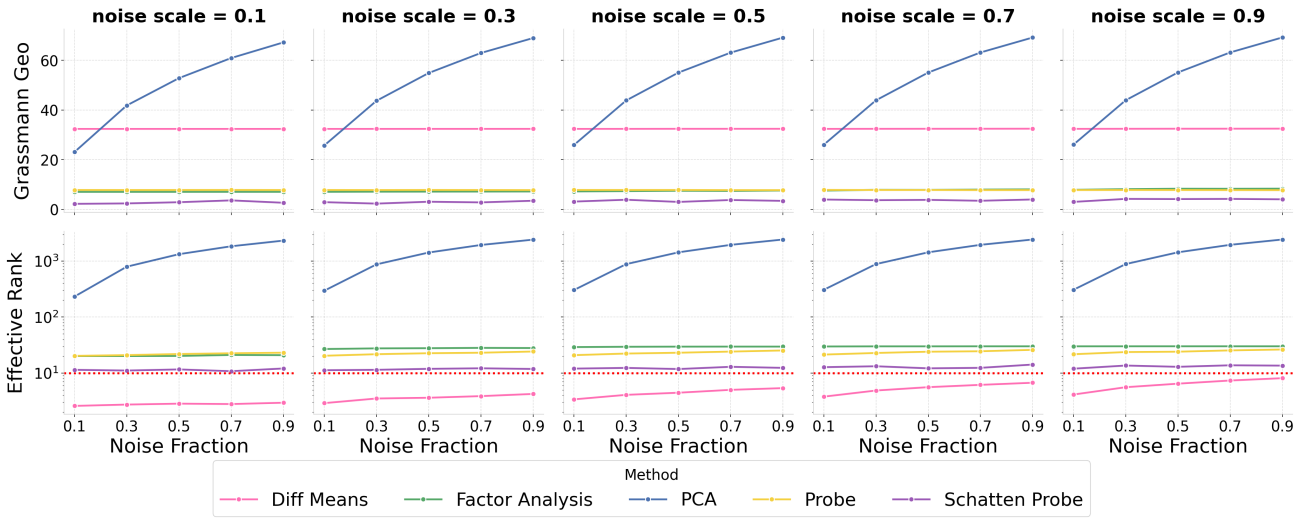

Figure 7. Results for toy setup with different levels of noise scale and noise fraction. The true dimensionality of the ground-truth plane is indicated by a dashed vertical red line. Here Noise scale refers to how far away from the affine subspace the points are and noise fraction refers to how many points are away from the subspace. Here the true dimension is 10 and the ambient dimension is 4096.

B.1. Grassmannian Metrics for Affine Subspaces

To evaluate our toy setup, we use the Grassmannian metric (Lim et al., 2021). The Grassmannian of affine subspaces $\mathrm{Graff}(k, n)$ is a natural generalization of the classical Grassmannian, with linear subspaces arising as the special case of zero offset. The metrics on $\mathrm{Graff}(\infty, \infty)$ are defined in Tab. 2, and we provide an intuitive visualization in Fig. 9. For all experiments we use the Grassmannian geodesic distance d_{geo} as our primary metric. Empirically, we find that all three metrics are similar across different evaluations, so we default to d_{geo} for consistency.

Intuitively, d_{geo} jointly accounts for three sources of differences between affine subspaces: the *principal angle*, the *offset* between subspaces, and *dimension mismatch*. Given two affine subspaces $\mathbb{A} + \boldsymbol{b}$ and $\mathbb{B} + \boldsymbol{c}$ of dimensions k and k' respectively, the geodesic distance is computed as:

$$d_{\mathrm{geo}}(\mathbb{A} + \boldsymbol{b}, \mathbb{B} + \boldsymbol{c}) = \left(\frac{\pi^2}{4} |k - k'| + \sum_{i=1}^{\min(k+1,\, k'+1)} \theta_i^2 \right)^{1/2} \tag{3}$$

where $\theta_i = \cos^{-1}(\sigma_i) \in [0, \frac{\pi}{2}]$ are the affine principal angles, with σ_i being the i-th singular value of $\boldsymbol{Y}_{\mathbb{A}+\boldsymbol{b}}^\top \boldsymbol{Y}_{\mathbb{B}+\boldsymbol{c}}$.

The Stiefel coordinates $\boldsymbol{Y}_{\mathbb{A}+\boldsymbol{b}} \in \mathbb{R}^{(n+1)\times(k+1)}$ embed the affine subspace into a higher-dimensional representation suitable for computing distances via standard linear algebra. For an affine subspace $\mathbb{A} + \boldsymbol{b}$ with orthogonal affine coordinates $[\boldsymbol{A}, \boldsymbol{b}_0]$, where $\boldsymbol{A} \in \mathbb{R}^{n\times k}$ is an orthonormal basis for $\mathbb{A}$ (with $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_k$) and $\boldsymbol{b}_0 \in \mathbb{R}^n$ satisfies $\boldsymbol{A}^\top \boldsymbol{b}_0 = \boldsymbol{0}$, the Stiefel coordinates are:

$$\boldsymbol{Y}_{\mathbb{A}+\boldsymbol{b}} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b}_0 / \sqrt{1 + \|\boldsymbol{b}_0\|^2} \\ \boldsymbol{0}^\top & 1 / \sqrt{1 + \|\boldsymbol{b}_0\|^2} \end{bmatrix} \in \mathbb{R}^{(n+1)\times(k+1)} \tag{4}$$

where $\boldsymbol{0} \in \mathbb{R}^k$ is the zero vector. The normalization factor $\sqrt{1 + \|\boldsymbol{b}_0\|^2}$ ensures that $\boldsymbol{Y}_{\mathbb{A}+\boldsymbol{b}}$ has orthonormal columns, making it an element of the Stiefel manifold $V(k+1, n+1)$.

The dependence of d_{geo} on each of these three factors is illustrated in Fig. 8 as well as in Fig. 9.

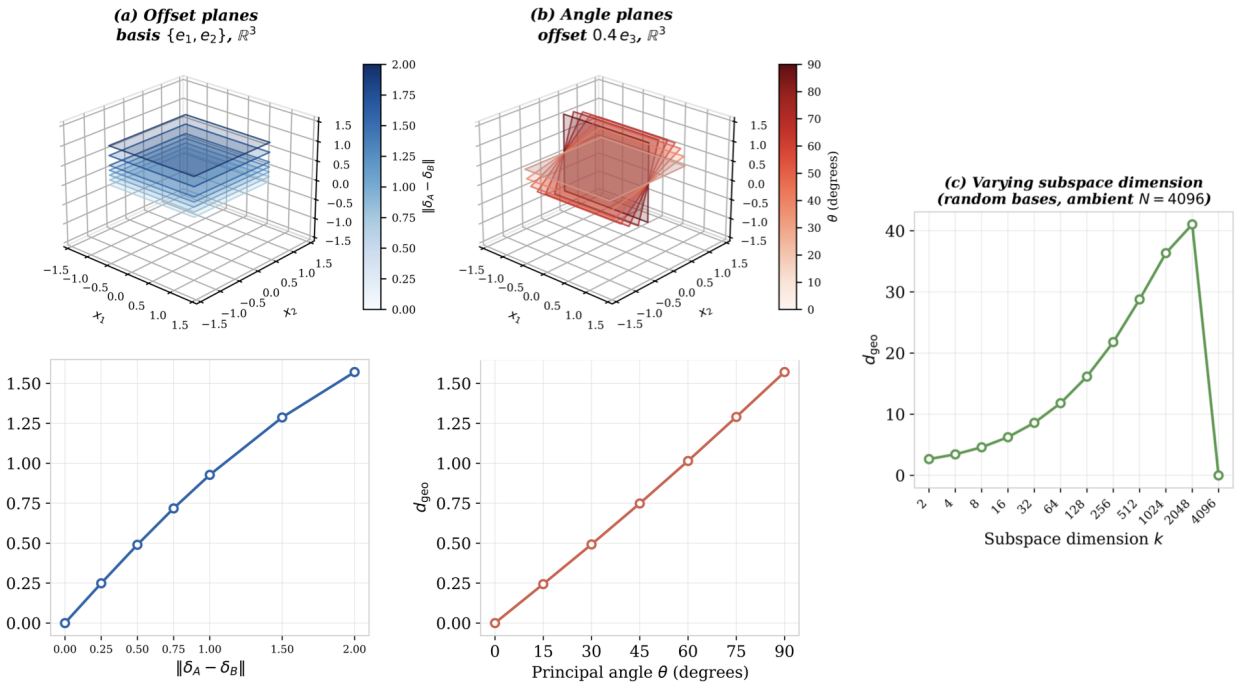

Figure 8. Intuition for the Grassmannian geodesic metric d_{geo} across three geometric settings, all using affine subspaces of dimension $k = 2$. **(a)** Varying offset: both subspaces share the same basis $\{e_1, e_2\}$ in $\mathbb{R}^3$ but are displaced along e_3 by offsets $\delta \in \{0, 0.25, 0.5, 0.75, 1.0, 1.5, 2.0\}$ (shown light to dark); d_{geo} grows monotonically with the separation $\|\delta_A - \delta_B\|$. **(b)** Varying angle: both subspaces share offset $0.4\,e_3$ in $\mathbb{R}^3$ but the recovered subspace (red, light to dark) is tilted away from the true subspace (blue) by principal angle $\theta \in \{0, 15, 30, 45, 60, 75, 90\}$; d_{geo} increases monotonically with θ. **(c)** Varying subspace dimension: two random affine subspaces of matching dimension k are drawn in ambient $\mathbb{R}^{4096}$; d_{geo} grows with k as higher-dimensional random subspaces are increasingly likely to be nearly orthogonal.

Table 2. Metrics on $\text{Graff}(\infty, \infty)$, where $k = \dim \mathbb{A}$, $l = \dim \mathbb{B}$, and θ_i are the affine principal angles between $\mathbb{A} + b$ and $\mathbb{B} + c$.

Distance Metric	Definition
Grassmann	$d_{\text{Graff}(\infty,\infty)}(\mathbb{A}+b,\, \mathbb{B}+c) = \left(\frac{\pi^2}{4}\lvert k-l\rvert + \sum_{i=1}^{\min(k+1,\, l+1)} \theta_i^2 \right)^{1/2}$
Chordal	$d^{\kappa}_{\text{Graff}(\infty,\infty)}(\mathbb{A}+b,\, \mathbb{B}+c) = \left(\lvert k-l\rvert + \sum_{i=1}^{\min(k+1,\, l+1)} \sin^2\theta_i \right)^{1/2}$
Procrustes	$d^{\rho}_{\text{Graff}(\infty,\infty)}(\mathbb{A}+b,\, \mathbb{B}+c) = \left(\lvert k-l\rvert + 2\sum_{i=1}^{\min(k+1,\, l+1)} \sin^2(\theta_i/2) \right)^{1/2}$

C. Concept Dataset

In Table 3, we present an overview of the concepts, subcategories, and instances we use in our multinomial concept datasets.

C.1. Prompts

Similar to AxBench eval, we randomly select 5 prompts from *Alpaca Eval* dataset and use this fixed set across all experiments; see Tab. 4.

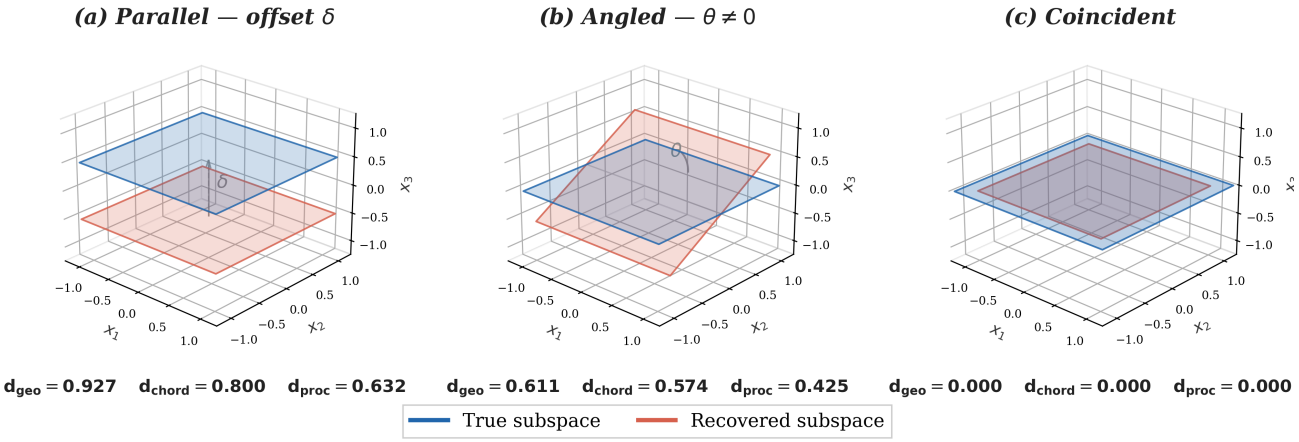

Figure 9. Affine subspace recovery in $\mathbb{R}^3$ under three geometric configurations. The blue plane denotes the true subspace and the red plane the recovered subspace. **(a)** Parallel planes separated by offset δ, with no angular difference between the subspaces. **(b)** Planes intersecting at a principal angle $\theta \neq 0$, sharing a common origin. **(c)** Coincident planes, where the recovered subspace exactly matches the true subspace ($d_{\text{geo}} = d_{\text{chord}} = d_{\text{proc}} = 0$). Distances are computed using the geodesic (d_{geo}), chordal (d_{chord}), and Procrustes (d_{proc}) metrics. The affine subspaces are defined by the following basis vectors and offsets: **(a)** true subspace spanned by $\{e_1, e_2\}$ with offset $+0.5\, e_3$, recovered subspace spanned by $\{e_1, e_2\}$ with offset $-0.5\, e_3$; **(b)** true subspace spanned by $\{e_1, e_2\}$ with offset $+0.4\, e_3$, recovered subspace spanned by $\{e_1,\ \cos\theta\, e_2 + \sin\theta\, e_3\}$ with offset $-0.4\, e_3$, where $\theta = 35°$; **(c)** both subspaces coincide, spanned by $\{e_1, e_2\}$ with shared offset $+0.3\, e_3$, giving zero distance under all three metrics.

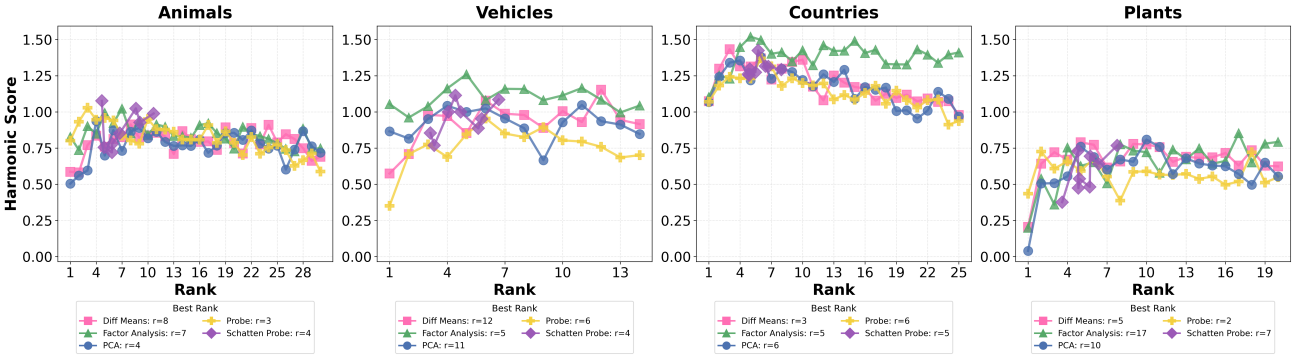

Figure 10. Sampling results across different methods for Llama-3.1-8B-Instruct Layer 25. Harmonic scores and best rank per method.

C.2. Hyperparameters

For sampling evaluation, we introduce α as a hyperparameter that controls the strength of intervention when steering the model's activations. We perform a grid search over $\alpha \in [0.1, 0.9]$ with a step size of 0.1 to identify the optimal intervention strength for each concept and model combination.

Through this search, we found that different models and concepts converge to different optimal α values. For Gemma, concepts: animals, countries, plants perform best with $\alpha = 0.78$, and vehicles require stronger intervention with $\alpha = 0.9$. For Llama we use $\alpha = 0.15$ for animals, vehicles and 0.35 for plants and countries. To ensure the model maintains fluent generation throughout the sequence, we exponentially decay α from its initial value to 0.1 over the course of generation.

Schatten Probe Hyperparameters. We train variational Schatten-p probes with $p \in \{0.1, 0.2, \ldots, 0.9\}$. Unless otherwise specified, we use the Adam optimizer with a learning rate of 10^{-2}, a regularization coefficient of $\lambda = 10^{-2}$, and train for 2,000 optimization steps. The factorization rank is initialized to 30 (corresponding to the number of entities per concept), factor columns are pruned every 50 optimization steps, and columns with energy below 10^{-5} are removed.

Table 3. Dataset overview. Blue indicates In Distribution (ID) and Orange represents Out of Distribution (OOD)

Domain	Category	Instances
Animals	Mammals	Beaver, Lion, Rabbit, Giraffe, Dolphin, **Elephant**
	Reptiles	Lizard, Turtle, Dinosaur, Crocodile, Gecko, **Chameleon**
	Birds	Parrot, Duck, Owl, Eagle, Penguin, **Crow**
	Amphibians	Caecilian, Salamander, Axolotl, Frog, Diplocaulus, **Toad**
	Fish	Carp, Seahorse, Shark, Anchovy, Salmon, **Tuna**
	Insects	Dragonfly, Mantis, Beetle, Fly, Ant, **Grasshopper**
	Molluscs	**Octopus, Squid, Slug, Clam, Oyster**
Plants	Trees	Oak, Pine, Maple, Mango, Coconut, **Banyan**
	Shrubs	Rose, Lavender, Hibiscus, Azalea, Boxwood, **Hydrangea**
	Herbs	Basil, Mint, Spinach, Wheat, Lettuce, **Cilantro**
	Climbers	Grapevine, Pea, Bean, Ivy, Pepper, **Passionflower**
	Creepers	**Pumpkin, Watermelon, Strawberry, Cucumber, Squash, Gourds**
Countries	Africa	Egypt, Kenya, Ethiopia, Ghana, Morocco, **Sudan**
	Asia	China, India, Japan, Thailand, Indonesia, **Vietnam**
	Europe	Germany, France, Italy, Spain, Portugal, **Poland**
	North America	Canada, Mexico, Cuba, Panama, Jamaica, **Belize**
	South America	Brazil, Argentina, Chile, Colombia, Peru, **Ecuador**
	Oceania	**Australia, Fiji, Samoa, Tonga, Kiribati, Vanuatu**
Vehicles	Land	Car, Bus, Truck, Bike, Scooter, **Van**
	Rail	Train, Tram, Metro, Subway, Monorail, **Funicular**
	Air	Plane, Jet, Glider, Drone, Copter, **Blimp**
	Water	**Boat, Ship, Yacht, Canoe, Raft, Skiff**

D. LLM-as-a-Judge

We use Mistral-Large to compute all evaluation scores through the prompts detailed in App. G. Our diversity score consists of two components: a category score that measures whether the generated text mentions the target concept category, and an instance score that evaluates whether specific instances within that category are present. Following the evaluation framework from AxBench, we compute a concept score which measures whether the concept is present in the LLM output. We then calculate the final metric as the harmonic mean between the diversity score (recall) and the concept score (precision).

Rank-wise results comparing different steering methods are presented in Fig. 11 for Gemma-2-2B-IT and Fig. 10 for Llama-3.1-1B-Instruct. More detailed breakdowns of performance across various sampling strategies and layer selections are shown in Fig. 12 and Fig. 13.

E. Additional Metrics

Projection. We define a projection-based metric to measure how close a point lies to the recovered affine subspace. Given the recovered affine subspace

$$\mathcal{A} = \{L\mathbf{t} + \mathbf{b} \mid t \in \mathbb{R}^k\},$$

where $L \in \mathbb{R}^{d\times k}$ defines the linear subspace and $\mathbf{b} \in \mathbb{R}^d$ is the offset. Let $U \in \mathbb{R}^{d\times k}$ denote a basis for the linear subspace L.

The projection of a point $\mathbf{x} \in \mathbb{R}^d$ onto the affine subspace $\mathcal{A}$ is given by

$$\text{proj}_{\mathcal{A}}(x) = \mathbf{b} + \text{proj}_L(\mathbf{x} - \mathbf{b}),$$

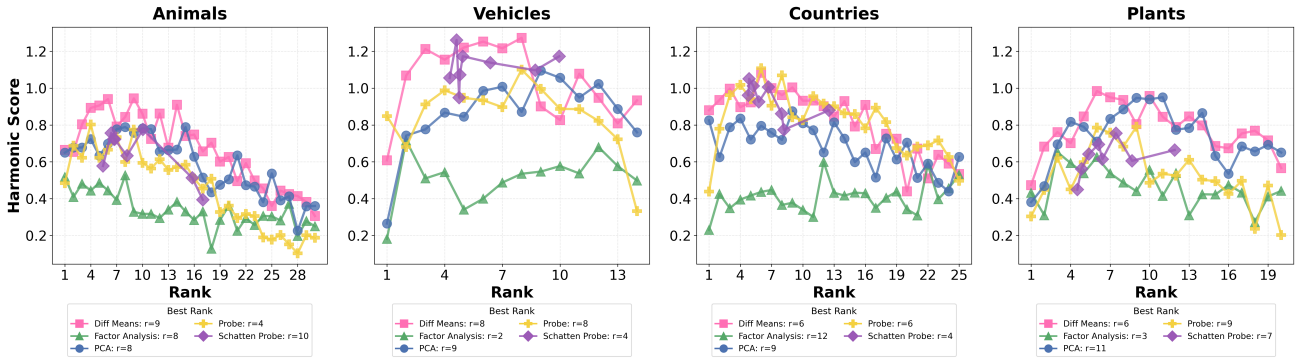

Figure 11. Sampling results across different methods for Gemma-2-2B-it Layer 16. Harmonic scores and best rank per method.

Table 4. Randomly sampled prompts from the Alpaca-Eval dataset.

#	Alpaca Eval Prompts
1	I like to host guests at my home from time to time, and I am gathering recipes of different dishes and drinks to keep things interesting. I am interested in trying some Argentinian dishes. Can you give me a recipe for Milanesa de peceto?
2	Can you write a short story where Hildibrand Manderville somehow gets transformed into Gandalf the Grey and he must work with The Warrior of Light and Nashu to restore his regular, most gentlemanly form.
3	Write a code block in Markdown containing an example of a code block in Markdown. Don't forget those quadruple backticks.
4	Create a daily itinerary based on the given information. Our family is looking for a 9-day Morocco trip that has light to moderate activity levels.
5	As a sports commentator, describe the winning play in the final seconds of a championship game.

where the projection onto the linear subspace L is defined as

$$\text{proj}_L(x) = U(U^\top U)^{-1}U^\top \mathbf{x}.$$

We then measure the distance between the activation x and the recovered affine subspace as

$$d(\mathbf{x}, \mathcal{A}) = \|\mathbf{x} - \text{proj}_{\mathcal{A}}(\mathbf{x})\|_2.$$

Intuitively, if the recovered affine subspace correctly captures the geometry of a target concept, then activations corresponding to *in-distribution* (ID) concept examples $\mathbf{x}_C$ should lie close to the subspace. Similarly, activations from related but held-out out-of-distribution (OOD) concept examples $\mathbf{x}_{C'}$ may also lie near the subspace, while activations corresponding to unrelated concepts $\mathbf{x}_{\tilde{C}}$ should lie farther away. We say that $\mathcal{A}$ is *selective* for the concept if $d(\mathbf{x}_{\tilde{C}}, \mathcal{A}) - d(\mathbf{x}_C, \mathcal{A})$ is generally large.

Because projection distance naturally decreases as the subspace rank increases, this metric is inherently rank-dependent. To ensure fair comparison across methods, we evaluate all methods under a rank-controlled setting by computing projection distances across multiple ranks.

Table 5. Performance metrics across all concepts for layer 25 Llama-3.1-8B-Instruct

	Animals							**Vehicles**						
	Concept	Diversity	Div Breakdown		Harmonic			Concept	Diversity	Div Breakdown		Harmonic		
Method	Score	Score	Inst	Cat	Score	Unseen	Rank	Score	Score	Inst	Cat	Score	Unseen	Rank
Prompting	0.400	1.479	0.842	0.637	0.630	-	-	0.226	0.948	0.581	0.637	0.366	-	-
Diff Means	0.620	1.731	**0.962**	0.769	0.913	43	8.00	0.867	1.722	0.862	**0.860**	1.153	34	12.00
FA	0.713	1.807	0.949	0.858	1.023	**60**	7.00	**1.013**	1.663	0.817	0.846	**1.259**	**39**	5.00
PCA	0.627	**1.852**	0.953	0.899	0.936	42	4.00	0.760	**1.687**	**0.904**	0.783	1.048	38	11.00
Probe	0.720	1.795	0.920	0.875	1.028	32	3.00	0.680	1.603	0.851	0.752	0.955	17	6.00
Sparse Probe	**0.760**	1.840	0.929	**0.911**	1.076	26	4.67	0.840	1.652	0.851	0.800	1.114	25	4.42
	Plants							**Countries**						
	Concept	Diversity	Div Breakdown		Harmonic			Concept	Diversity	Div Breakdown		Harmonic		
Method	Score	Score	Inst	Cat	Score	Unseen	Rank	Score	Score	Inst	Cat	Score	Unseen	Rank
Prompting	0.270	1.628	0.942	0.686	0.463	-	-	0.776	1.251	0.624	0.626	0.958	-	-
Diff Means	0.520	1.641	**0.913**	0.727	0.790	**30**	5.00	1.240	1.696	0.901	0.795	1.432	17	3.00
FA	**0.580**	1.605	0.879	0.726	**0.852**	18	17.00	**1.296**	**1.837**	**0.910**	**0.928**	**1.520**	**31**	5.00
PCA	0.540	1.607	0.838	0.770	0.808	9	10.00	1.136	1.757	0.881	0.875	1.380	19	6.00
Probe	0.460	**1.714**	0.871	**0.843**	0.725	6	2.00	1.136	1.724	0.879	0.845	1.369	16	6.00
Sparse Probe	0.500	1.653	0.840	0.813	0.768	8	7.77	1.200	1.752	0.892	0.860	1.424	16	5.75

Table 6. Performance metrics across all concepts for Gemma-2-2B-it layer 16

	Animals							**Vehicles**						
	Concept	Diversity	Breakdown		Harmonic			Concept	Diversity	Breakdown		Harmonic		
Method	Score	Score	Inst	Cat	Score	Unseen	Rank	Score	Score	Inst	Cat	Score	Unseen	Rank
Prompting	0.333	0.858	0.610	0.248	0.480	-	-	0.040	0.200	0.200	0.00	0.006	-	-
Diff Means	**0.647**	1.748	0.902	0.846	**0.944**	**35**	9.00	**0.973**	**1.840**	0.941	**0.899**	**1.273**	32	8.00
FA	0.307	**1.861**	**0.954**	**0.907**	0.527	28	8.00	0.467	1.535	0.941	0.595	0.716	10	2.00
PCA	0.507	1.777	0.902	0.875	0.789	12	8.00	0.800	1.737	**0.977**	0.759	1.095	10	9.00
Probe	0.527	1.684	0.864	0.820	0.802	16	4.00	0.800	1.763	0.889	0.874	1.101	28	8.00
Sparse Probe	0.493	1.814	0.933	0.880	0.776	22	10.05	**0.973**	1.791	0.901	0.890	1.261	**39**	4.62
	Plants							**Countries**						
	Concept	Diversity	Breakdown		Harmonic			Concept	Diversity	Breakdown		Harmonic		
Method	Score	Score	Inst	Cat	Score	Unseen	Rank	Score	Score	Inst	Cat	Score	Unseen	Rank
Prompting	0.130	0.371	0.188	0.182	0.192	-	-	0.544	0.784	0.392	0.391	0.642	-	-
Diff Means	**0.690**	1.716	0.970	0.746	**0.984**	13	6.00	0.792	1.715	0.906	0.808	1.084	10	6.00
FA	0.400	1.771	**0.982**	0.789	0.653	7	3.00	0.376	1.476	0.845	0.631	0.599	11	12.00
PCA	0.640	**1.846**	0.965	**0.881**	0.950	**14**	11.00	0.608	1.558	0.800	0.758	0.875	8	9.00
Probe	0.510	1.757	0.972	0.785	0.791	9	9.00	**0.800**	**1.800**	**0.914**	**0.886**	**1.108**	**14**	6.00
Sparse Probe	0.480	1.765	0.967	0.799	0.755	10	7.44	0.760	1.700	0.850	0.849	1.050	8	4.85

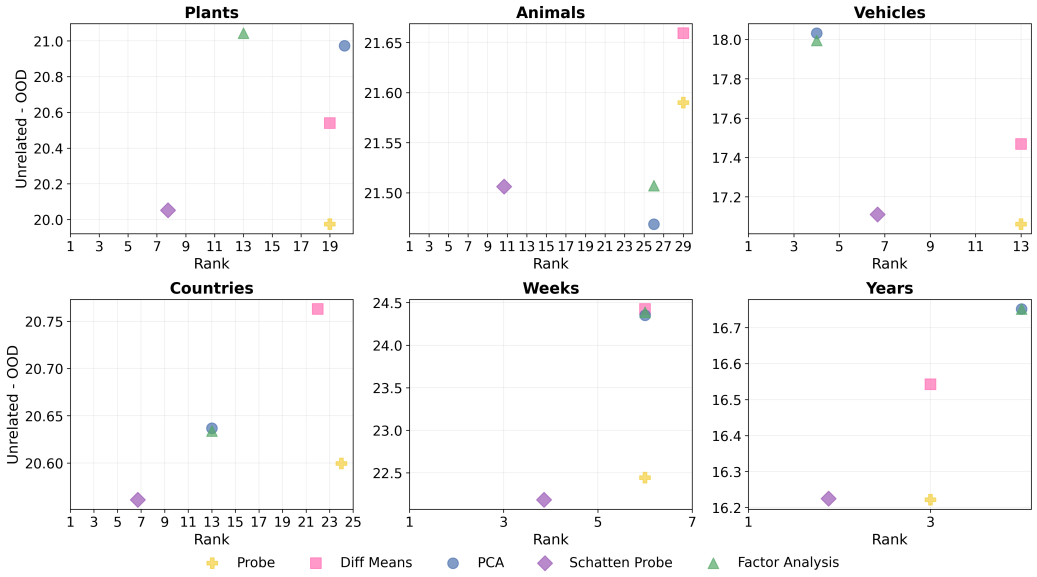

Figure 14. Projection Distance results for layer 25 Llama-3.1-8B-Instruct varied across lighest unrealted -ood distance rank shown per method.

F. Qualitative Samples

Here, we provide examples of model outputs for each sampling method. We provide examples for ANIMALS (Tab. 7), VEHICLES (Tab. 8), PLANTS (Tab. 9), and COUNTRIES (Tab. 10).

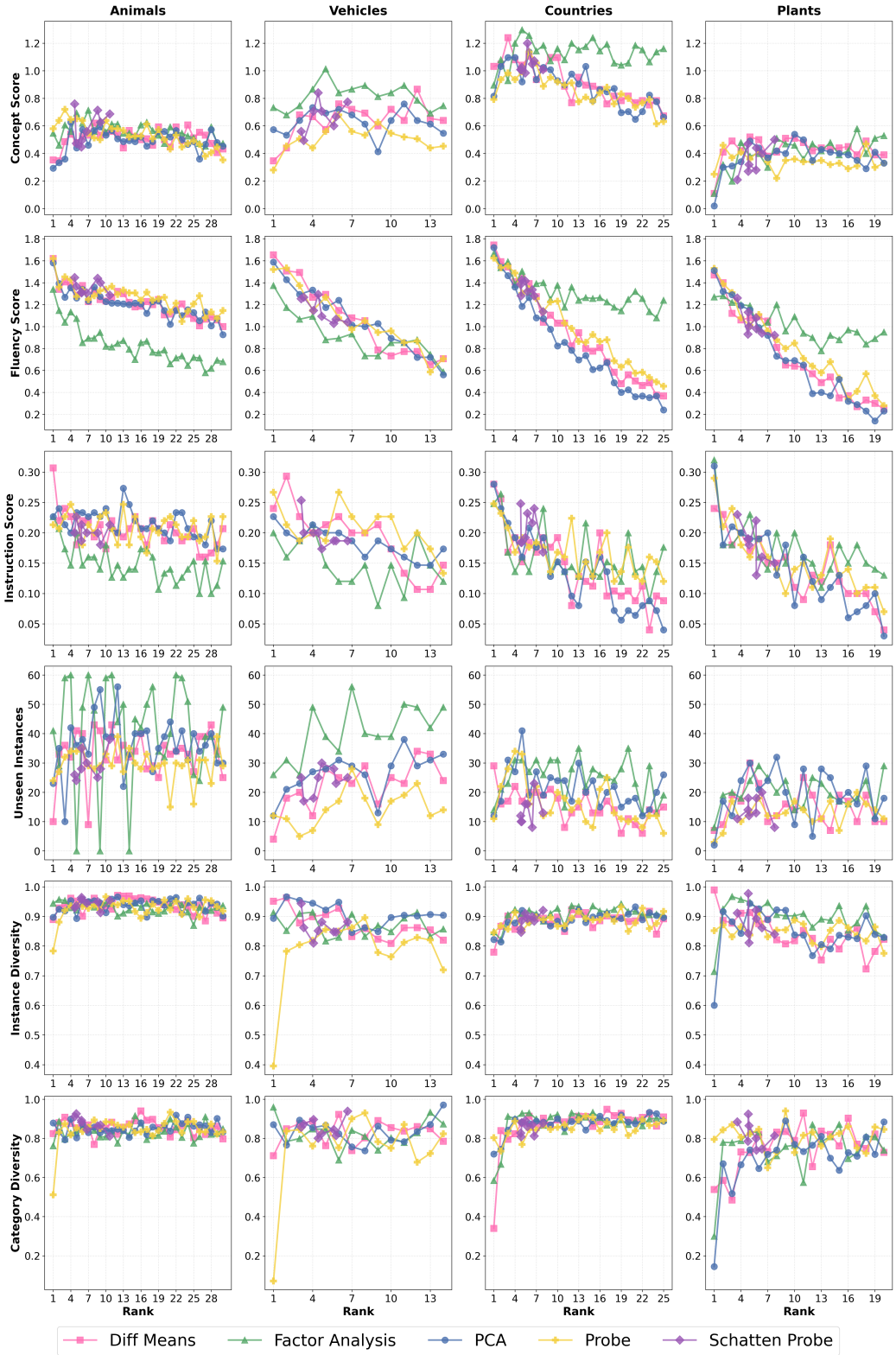

Figure 12. Sampling results across different methods for Llama-3.1-8B-Instruct Layer 25. Each row corresponds to a different concept (plants, animals, vehicles), and each column shows a different metric. Instance Diversity and Category Diversity measure the variety of generated samples, while Concept Score, Fluency Score, and Instruct Score evaluate the quality of generated text.

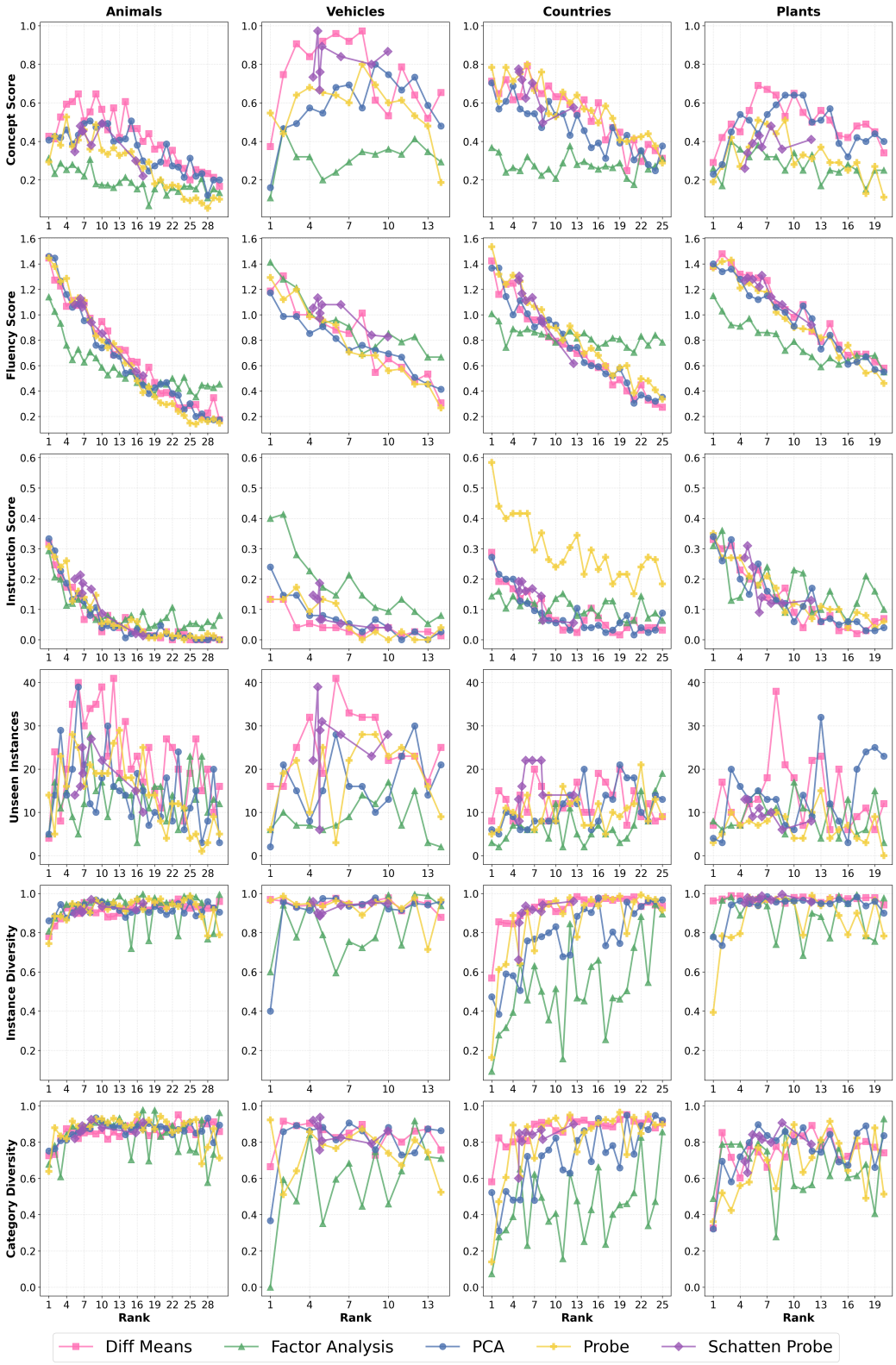

Figure 13. Sampling results across different methods for Gemma-2-2B-it Layer 16. Each row corresponds to a different concept (plants, animals, vehicles), and each column shows a different metric. Instance Diversity and Category Diversity measure the variety of generated samples, while Concept Score, Fluency Score, and Instruct Score evaluate the quality of generated text.

Table 7. Qualitative samples for ANIMALS concept.

Concept	**Main Prompt: Create a daily itinerary based on the given information. Our family is looking for a 9-day Morocco trip that has light to moderate activity levels.**
Prompting	1. We will be traveling with our two young children aged 6 and 8. We also have a dog, but we won't be bringing her along.**Day 1: Arrival in Marrake 2. We are a group of four, including two children aged 9 and 11. The children love animals, and my wife is interested in exploring local. 3. Our interests include history, culture, and wildlife. Assuming that we will be in Morocco from April 15th to April 23rd, we have the following preferences:
Diff Means	1. Giraffe, monkeys, and monkeys (specifically lemurs) are the animals we want to see. Morocco is known for its vibrant culture and beautiful landscapes. 2. Seals, dolphins, and Dune bashing are not on the trip agenda. 3. Bird watching, camel ride, and cooking class are must-do activities. We're interested in exploring the desert, imperial cities
Factor Analysis	1. Ants in the the Ant Moth Grasshopper Beetle Fly and various butterflies and Dragonflies are all common in our area. We will be traveling with a 9-year-old. 2. Dinosaur fossil remains have been found in some parts of the country and we're interested in visiting sites with dinosaur fossils. This is not a birdwatching trip. 3. Seafood lovers and interested in history, culture, and architecture. Here are some of our preferences: - We prefer to stay in central
PCA	1. Dragon Boat Racing is scheduled for day 8. Day 1 starts in Marrakech, day 3 is in Essaouira, day 6 is in Chefchaouen, and 2. Reptiles and other small animals are a valid interest for our family. Thus, we would like to visit the desert, a park with reptiles and a garden. 3. Insect repellent and sunscreen are must-haves. We want to explore the beautiful cities of Morocco and learn about
Probe	1. giraffe-shaped camels with a red and white striped blanket and a cheerful crew, and the trip to be a blend of culture, history, 2. Gecko's Frontera is a hotel in Marrakech, which is a 20-minute walk from Jardin Majorelle. We have a 1-day Sahara Desert Tour booked on Day 3. Whaleshark, a popular snorkeling destination, and Sahara desert. We want to explore cities, culture, and architecture
Schatten Probe	1. Salmon is one of the famous dishes in Morocco. If we plan to try salmon for dinner at a local restaurant on the third day, 2. Lizards, camels, and the Sahara Desert are the highlights of our trip. We will be traveling in February. 3. Llama trek, desert night in a tent, and a camel trek are the top three experiences we want to have during this trip.

Table 8. Qualitative samples for VEHICLES concept.

Concept	**Main Prompt: Create a daily itinerary based on the given information. Our family is looking for a 9-day Morocco trip that has light to moderate activity levels.**
Prompting	1. We are interested in visiting the famous Chefchaouen, Tangier, and Marrakech. We prefer to drive ourselves rather than rely on public transportation. 2. Our family consists of two adults and two children aged 6 and 8. We have booked a 7-night stay at a riad in Marrakech and have a 2-day car rental 3. We have already booked our flights from Europe to Marrakech, which will arrive in the morning. We are staying in a riad in the o
Diff Means	1. scooters and Bikes are not a good option for our 9-year-old and 3-year-old. Therefore, we have selected the following itinerary: 2. Tramways, monorochrom, and monochrometers. The 9-day Morocco trip includes visits to Casablanca, Marrakech, Fez, Essaouira, 3. Subway (metro) and bus train subway metro and other means of transportation will be used for all destinations, no driving.
Factor Analysis	1. Plane and car rental are already booked. We have the following interests: - Explore the local culture and history 2. Trains in the Marrakech metro is a convenient and efficient way to get around. The family consists of two adults and three children 3. Sleeper train, 4x4 vehicle, and flights are not available. We want to explore the vibrant city of Marrakech and experience the stunning beauty of the Atlas Mountains.
PCA	1. Electric scooters will be used for transportation in Fez and Marrakech, as well as public transportation and walking in other areas. 2. Underground Trubia metro line in Trubia, Spain is a fascinating place. I love tapas, and my family would love a visit to the famous Spanish restaurant, 3. Electric mobility scooters are not available for rent, but you can use a city map and/or a guide to navigate. In Morocco, the people are very friendly.
Probe	1. Plane travel takes up two days (day 1-2), and you want to spend 3 days in the Atlas Mountains and 2 days in Marrakech. 2. Metro and walking are our preferred modes of transportation. We have a few special requests: Explore the beautiful and historic city of Marrakech 3. Jet ski for a time, rent a scooter or rent a car to drive the Atlantic coast. Jet Skiing: Go Jet Skiing at the beautiful lagoon
Schatten Probe	1. Trams and public transportation in Casablanca is very limited, so we will not use it. Instead, we will use taxis. 2. Bus-train combo ticket, including transportation, accommodation and food, will be purchased. 3. Truck tours, boat trips, and ouch camel rides. We're interested in seeing the Sahara Desert, visiting cities,

Table 9. Qualitative samples for PLANTS concept.

Concept	**Main Prompt: Create a daily itinerary based on the given information. Our family is looking for a 9-day Morocco trip that has light to moderate activity levels.**
Prompting	1. We have kids aged 3 and 6, so we want to make sure there are plenty of kid-friendly activities and attractions. Here's a breakdown of our preferences: We love nature and 2. We would like to explore some of the vibrant cities and historical sites, but also experience the natural beauty of the country. 3. We want to explore the cities, towns, and natural wonders of this fascinating country. We're interested in learning about the local flora and fauna and would like to include a visit to a garden or botanical
Diff Means	1. Mint tea is our favorite, so we want to visit a mint tea farm. We're also interested in seeing the Sahara Desert 2. Pine trees are a must, and we want to stay in accommodations with some extra space. We have a family of five, including two kids. 3. Ivy, the mom, loves trying new foods, meeting new people, and learning about history and culture. Her husband, Tom, is interested
Factor Analysis	1. Forests, mountains, and desert regions are all part of the Morocco experience. Our family of four consists of two adults, 2. Wheat bread, lentil, and vegetable dishes are staple foods for us. We are interested in experiencing local culture and learning 3. Spin and Spin Pepper (P) is one of the desert peppers. The other pepper family member, which is more well-known, is Bell (B).
PCA	1. Spinach and Fava bean soup, Moroccan bread, and dates are some of the food items we're interested in trying. 2. Maple syrup maple pine needle pine oak and birch are the tree species we are interested in learning about. 3. Spin and harvest beans for the desert peas and then spin and harvest wheat to store for the winter. Our trip will begin in Marrakech,
Probe	1. Coconut, our 10-year-old, is interested in exploring the desert, camels, and markets. Our children, 9-year-old Coco and 7-year-old Lila, 2. Basil is a history buff and wants to explore the ancient medinas and historical sites. Rachel is an art lover and wants to see 3. Pine trees, and the palm trees they've been dreaming of, but we also want to explore the ancient cities, medinas, and local culture.
Schatten Probe	1. Mint-flavored tea and fresh orange juice are favorite drinks of family members. On the day of arrival, we check into a riad in 2. Grapevine, Oued Lettouf, and the nearby beaches of Bouregreg will be the primary focus of our trip. 3. Grape and lettuce let, wild asparagus, and pease p Our family is looking for a 9-day Morocco trip that

Table 10. Qualitative samples for COUNTRIES concept.

Concept	**Main Prompt: Create a daily itinerary based on the given information. Our family is looking for a 9-day Morocco trip that has light to moderate activity levels.**
Prompting	1. We're planning to explore Morocco's imperial cities, Sahara Desert, and the Mediterranean coast. We want to balance culture, history, 2. We are interested in exploring Morocco's diverse cities, history, and culture. We'd like to stay in riads, which are traditional 3. We are planning to visit Marrakech, Essaouira, and Chefchaouen. We will travel by car and would like to visit some historical sites,
Diff Means	1. Canada is our home country, and we have a one-day layover in London. The trip must include a 4-day desert safari adventure, 2. Caribbean Asian food is not a priority, but rather a variety of authentic local dishes. The family has a few specific requests: 3. Bolivia, Argentina, and Spain were all considered, but Morocco was the winner! We have two children, aged 7 and 10.
Factor Analysis	1. France, Italy, and Spain are not considered, but we are interested in visiting Morocco's cities, natural landscapes, 2. Great Wall of China and German autobahung do not make sense in this context. Since the't are traveling with kids, they will want to experience local culture a 3. Peru Trip is a 14 day trip to Peru that offers a range of activities, from easy day hikes to more strenuous mountaineering. The Morocco trip costs $3,500
PCA	1. France and Spain are too long of a flight for my husband, so we're looking at destinations within Europe or North Africa. Morocco has been a long-standing favorite of ours. 2. Italy's cuisine is world-famous, but Morocco has its own unique blend of African, Arab, and Mediterranean flavors, 3. Germany is not a requirement, but we have a time-sensitive flight out of Munich. Thus, we want to avoid traveling to the north of Morocco as that would add extra travel time.
Probe	1. China? No. We're going to Morocco. And we'th going to explore the vibrant cities, stunning beaches, and historic sites. 2. France was our previous destination and we are now looking for something different from our previous experience. Morocco will be our third international trip. 3. Thailand may be our next destination, but we're unsure, so let's explore Morocco first. We'll be traveling with children aged 6 and 9 years old.
Schatten Probe	1. India and China are also being considered, but for this trip, we're focusing on Morocco. Given the time of year and weather, 2. Cubans and 4 kids are coming along, with the first child aged 6. The family wants to explore culture, try local cuisine, 3. Canada is our country of origin, and we'll be visiting during the jamboree (winter) season. In our party, there are 3 adults

G. Evaluation Prompts

G.1. Animals Evaluation Prompt

Animals Evaluation Prompt

You are evaluating a sentence that was generated in response to an instruction about animals.
Original instruction: `{instruction}`
Sentence: `{sentence}`

Concept score — does this sentence mention or clearly relate to animals?

- **2** = explicit animal mention: names a specific animal or species (e.g. "lion", "eagle", "salmon", "labrador"), a body part unique to animals, or animal behaviour
- **1** = indirect or metaphorical animal reference (e.g. "the python library", "Mustang car", "bear market") — animal word present but used in a non-animal sense
- **0** = no animal word, reference, or implication whatsoever

Return ONLY this JSON:

```
{
    "concept_score": 0|1|2,
    "reason": "<one sentence explaining all three scores>"
}
```

G.2. Plants Evaluation Prompt

Plants Evaluation Prompt

You are evaluating a sentence that was generated in response to an instruction about plants.
Original instruction: `{instruction}`
Sentence: `{sentence}`

Concept score — does this sentence mention or clearly relate to plants?

- **2** = explicit plant mention: names a specific plant, tree, flower, or crop (e.g. "oak", "rose", "fern", "wheat"), plant anatomy (leaf, root, petal), or plant behaviour (photosynthesis, germination)
- **1** = indirect or metaphorical plant reference (e.g. "Ivy League", "Rose" as a person's name, "budding career") — plant word present but not used in a botanical sense
- **0** = no plant word, reference, or implication whatsoever

Return ONLY this JSON:

```
{
    "concept_score": 0|1|2,
    "reason": "<one sentence explaining all three scores>"
}
```

G.3. Countries Evaluation Prompt

Countries Evaluation Prompt

You are evaluating a sentence that was generated in response to an instruction about countries or nations.
Original instruction: `{instruction}`
Sentence: `{sentence}`

Concept score — does this sentence mention or clearly relate to a country or nation?

- **2 =** explicit country mention: names a specific country, nation, or sovereign state (e.g. "France", "Japan", "Brazil"), its demonym (e.g. "French", "Japanese"), its capital, or a nationally iconic landmark
- **1 =** indirect or ambiguous geographical reference — mentions a region, continent, or city without naming a country, or uses a country name in a non-geographical sense (e.g. "Turkey" as food)
- **0 =** no country, nation, or geographical reference whatsoever

Return ONLY this JSON:

```
{
    "concept_score": 0|1|2,
    "reason": "<one sentence explaining all three scores>"
}
```

G.4. Vehicles Evaluation Prompt

Vehicles Evaluation Prompt

You are evaluating a sentence that was generated in response to an instruction about vehicles.
Original instruction: `{instruction}`
Sentence: `{sentence}`

Concept score — does this sentence mention or clearly relate to a vehicle?

- **2 =** explicit vehicle mention: names a specific vehicle, model, or type (e.g. "Boeing 747", "bicycle", "submarine", "freight train"), vehicle components (engine, cockpit, hull), or vehicle actions (accelerating, docking, taking off)
- **1 =** indirect or ambiguous vehicle reference — brand name used ambiguously (e.g. "Mercury" as a planet, "Mustang" as a horse), or a general transport-related word without naming a vehicle
- **0 =** no vehicle word, reference, or implication whatsoever

Return ONLY this JSON:

```
{
    "concept_score": 0|1|2,
    "sub_concept": "<specific vehicle or category mentioned, or 'none'>",
    "reason": "<one sentence explaining all three scores>"
}
```

G.5. Extraction Prompts

G.5.1. Animals Extraction Prompt

Animals Extraction Prompt

Read the following sentence and extract all animal mentions.
Sentence: `{sentence}`

Rules:

- Extract every animal word, even if used in a non-animal context (e.g. "Python" in programming → snake, "Jaguar" as a car → big cat)
- For each animal found, infer its taxonomic group (mammals, reptiles, birds, fish, insects, amphibians, arachnids, etc.)
- Breed-level mentions are distinct (e.g. "labrador" is not collapsed to "dog")
- Generic terms count too ("creature", "beast", "mammal", "bird")
- If no animals are present, return empty lists

Return ONLY this JSON:

```
{
    "items": [
        {"name": "<animal name>", "taxonomy": "<taxonomic group>"},
        ...
    ]
}
```

G.5.2. Plants Extraction Prompt

Plants Extraction Prompt

Read the following sentence and extract all plant mentions.
Sentence: `{sentence}`

Rules:

- Extract every plant word, even if used in a non-plant context (e.g. "Ivy League" → ivy → climber, "Rose" as a name → rose → shrub)
- For each plant found, infer its plant type from exactly: trees, shrubs, herbs, climbers, creepers
- Species-level mentions are distinct (e.g. "oak" is not collapsed to "tree")
- Generic terms count too ("plant", "flower", "tree", "shrub")
- If no plants are present, return empty lists

Return ONLY this JSON:

```
{
    "items": [
        {"name": "<plant name>", "taxonomy": "<plant type>"},
        ...
    ]
}
```

G.5.3. Countries Extraction Prompt

Countries Extraction Prompt

Read the following sentence and extract all country or nation mentions.
Sentence: `{sentence}`

Rules:

- Extract every country or nation mention, even if indirect (e.g. "Amazon" as a company → Brazil, "Turkey" as a food → Turkey the country)
- For each country found, infer its continent from exactly: Africa, Asia, Europe, North America, South America, Oceania, Antarctica
- Cities, regions, or sub-national areas are NOT countries — skip them
- If no countries are present, return empty lists

Return ONLY this JSON:

```
{
    "items": [
        {"name": "<country name>", "taxonomy": "<continent>"},
        ...
    ]
}
```

G.5.4. Vehicles Extraction Prompt

Vehicles Extraction Prompt

Read the following sentence and extract all vehicle mentions.
Sentence: `{sentence}`

Rules:

- Extract every vehicle mention, even if used in a non-vehicle context (e.g. "Mustang" as a horse → skip, "Mercury" as a car brand → road)
- For each vehicle found, infer its category from exactly: rail, road, air, water
- Model-level variants are distinct (e.g. "Boeing 747" ≠ "Boeing 737")
- Generic terms count too ("vehicle", "aircraft", "ship", "train")
- If the word is clearly not being used as a vehicle, skip it
- If no vehicles are present, return empty lists

Return ONLY this JSON:

```
{
    "items": [
        {"name": "<vehicle name>", "taxonomy": "<category>"},
        ...
    ]
}
```

G.6. Out-of-Distribution Detection Prompts

G.6.1. Animals OOD Detection Prompt

Animals OOD Detection Prompt

You are analyzing a list of animal mentions to determine which are Out-of-Distribution (OOD).
Training set (IN-distribution): `[{training_str}]`
Extracted instances: `[{instances_str}]`

Rules:

- An instance is OOD if it is NOT in the training set (case-insensitive comparison)
- EXCLUDE generic terms: “animal”, “animals”, “creature”, “beast”, “mammal”, “bird”, “fish”, “reptile”, “amphibian”, “insect”, “arachnid”, etc.
- ONLY count specific animal names (e.g., “tiger”, “elephant”, “sparrow”)
- Match instances to training set flexibly (e.g., “tigers” matches “tiger”, “labrador retriever” matches “labrador” if in training)

Return ONLY this JSON:

```
{
    "ood_instances": ["<specific_animal_1>", "<specific_animal_2>", ...],
    "ood_count": <number>
}
```

G.6.2. Plants OOD Detection Prompt

Plants OOD Detection Prompt

You are analyzing a list of plant mentions to determine which are Out-of-Distribution (OOD).
Training set (IN-distribution): `[{training_str}]`
Extracted instances: `[{instances_str}]`

Rules:

- An instance is OOD if it is NOT in the training set (case-insensitive comparison)
- EXCLUDE generic terms: “plant”, “plants”, “tree”, “trees”, “flower”, “flowers”, “shrub”, “herb”, “vegetation”, etc.
- ONLY count specific plant names (e.g., “sunflower”, “cactus”, “fern”)
- Match instances to training set flexibly (e.g., “roses” matches “rose”, “oak tree” matches “oak” if in training)

Return ONLY this JSON:

```
{
    "ood_instances": ["<specific_plant_1>", "<specific_plant_2>", ...],
    "ood_count": <number>
}
```

G.6.3. Countries OOD Detection Prompt

Countries OOD Detection Prompt

You are analyzing a list of geographical mentions to determine which are Out-of-Distribution (OOD).
Training set (IN-distribution): `[{training_str}]`
Extracted instances: `[{instances_str}]`

Rules:

- An instance is OOD if it is NOT in the training set (case-insensitive comparison)
- INCLUDE specific countries, cities, regions, landmarks (e.g., "Russia", "Tokyo", "Sahara Desert", "Eiffel Tower")
- EXCLUDE generic terms: "country", "countries", "nation", "city", "place", "region", "continent", etc.
- Match instances to training set flexibly (e.g., "United States" could match "usa" if in training)

Return ONLY this JSON:

```
{
    "ood_instances": ["<specific_place_1>", "<specific_place_2>", ...],
    "ood_count": <number>
}
```

G.6.4. Vehicles OOD Detection Prompt

Vehicles OOD Detection Prompt

You are analyzing a list of vehicle mentions to determine which are Out-of-Distribution (OOD).
Training set (IN-distribution): `[{training_str}]`
Extracted instances: `[{instances_str}]`

Rules:

- An instance is OOD if it is NOT in the training set (case-insensitive comparison)
- EXCLUDE generic terms: "vehicle", "vehicles", "transport", "transportation", "aircraft", "ship", "boat", etc.
- ONLY count specific vehicle names (e.g., "tesla model 3", "boeing 747", "bicycle", "kayak")
- Match instances to training set flexibly (e.g., "cars" matches "car", "passenger train" matches "train" if in training)

Return ONLY this JSON:

```
{
    "ood_instances": ["<specific_vehicle_1>", "<specific_vehicle_2>", ...],
    "ood_count": <number>
}
```

G.7. Activation Extraction Prompts

Each prompt contains a one-shot example followed by the target completion. The example demonstrates the format, then the same sentence is repeated for the model to complete. We observed that while Llama-3.1-8B-Instruct is flexible with different prompts (requiring only that the concept word appears somewhere in the prompt), Gemma-2-2B-it requires precise formatting instructions and tends to deviate without explicit structure.

G.7.1. ANIMALS ACTIVATION PROMPT

Animals Activation Extraction Prompt

One-shot example:
```
The Rabbit is a small mammal belonging to the family Leporidae and the order
Lagomorpha, which also includes hares and pikas.
Animal:  Rabbit
```
Model completes:
```
The Rabbit is a small mammal belonging to the family Leporidae and the order
Lagomorpha, which also includes hares and pikas.
Animal:
```

G.7.2. COUNTRIES ACTIVATION PROMPT

Countries Activation Extraction Prompt

One-shot example:
```
Egypt is officially called the Arab Republic of Egypt.
Country:  egypt
```
Model completes:
```
Egypt is officially called the Arab Republic of Egypt.
Country:
```

G.7.3. PLANTS ACTIVATION PROMPT

Plants Activation Extraction Prompt

One-shot example:
```
Pine trees belong to the genus Pinus in the family Pinaceae, making Pine a major
group of gymnosperm conifers.
Plant:  Pine
```
Model completes:
```
Pine trees belong to the genus Pinus in the family Pinaceae, making Pine a major
group of gymnosperm conifers.
Plant:
```

G.7.4. VEHICLES ACTIVATION PROMPT

Vehicles Activation Extraction Prompt

One-shot example:
```
A Truck is also called a lorry in British English and other regions.
Vehicle:  Truck
```
Model completes:
```
A Truck is also called a lorry in British English and other regions.
Vehicle:
```