On the Identifiability of Concepts from Large Language Model Activations

Anonymous authors
Paper under double-blind review

Abstract

Unsupervised approaches to large language model (LLM) interpretability, such as sparse autoencoders (SAEs), offer a way to decode LLM activations into interpretable and, ideally, controllable concepts. On the one hand, these approaches alleviate the need for supervision from concept labels, paired prompts, or explicit causal models. On the other hand, without additional assumptions, SAEs are not guaranteed to be identifiable. In practice, they may learn latent dimensions that entangle multiple underlying concepts. If we use these dimensions to extract vectors for steering specific LLM behaviours, this non-identifiability might result in interventions that inadvertently affect unrelated properties. In this paper, we bring the question of identifiability to the forefront of LLM interpretability research. Specifically, we introduce Sparse Shift Autoencoders (`SSAE`s) which learn sparse representations of *differences* between embeddings rather than the embeddings themselves. Crucially, we show that `SSAE`s are identifiable from paired observations which differ in multiple unknown concepts, but not all. With this key identifiability result, we show that we can steer single concepts with only this weak form of supervision. Finally, we empirically demonstrate identifiable concept recovery across multiple real-world language datasets by disentangling activations from different LLMs.

1 Introduction

As increasingly powerful large language models (LLMs) are deployed and widely used, the need to interpret and steer their behavior grows. For both interpretability and steering, we require techniques to disentangle LLM activations into semantically meaningful, and ideally, manipulable concepts. A large class of LLM interpretability methods rely on supervision from ground truth concepts (Koh et al., 2020), paired prompts (Turner et al., 2024), target LLM completions (Subramani et al., 2022) and abstract causal models of behavior (Geiger et al., 2024) to map activations to concepts. For example, using contrastive pairs of prompts that differ by a single concept, recent papers have found vectors in activation space that encode sycophancy (Rimsky et al., 2024), truthfulness (Park et al., 2025), and refusal (Arditi et al., 2024). However, acquiring such supervision is costly, motivating unsupervised methods for concept learning.

Sparse autoencoders (SAEs) have emerged as a popular approach to unsupervised LLM interpretability (Cunningham et al., 2023). Taking inspiration from sparse dictionary learning, SAEs encode LLM activations in a sparse and overcomplete representation. While we might hope that there is a one-to-one correspondence between the learned dimensions and interpretable concepts, (Wu et al., 2025; Menon et al., 2025) show empirical evidence that SAEs significantly underperform supervised methods, suggesting that they may not be *identifiable*: that is, they could learn latent dimensions that entangle interpretable concepts. Consequently, if we use SAEs to extract activation directions to steer LLM behavior, non-identifiability could result in changes to unrelated properties.

In this paper, we propose Sparse Shift Autoencoders (`SSAE`s), models for provably recovering steering vectors without the need for concept labels, contrastive pairs and other supervision signals about concepts. Crucially, `SSAE`s learn from sparse multi-concept shifts: paired samples in which multiple unknown concepts differ, but not all of them. Such samples are cheap to obtain, for example, by pairing sentences from Wikipedia articles, or by using LLMs to synthetically generate contrastive texts. Briefly, an `SSAE` maps embedding differences between samples in a pair to a latent space that reflects the concept changes and uses a linear decoding function to reconstruct the difference vector. This architecture reflects the *linear representation hypothesis* (Mikolov et al., 2013; Jiang et al., 2024) in assuming that concepts are linearly encoded by LLMs. Crucially, we regularize the latent representation to be sparse, meaning that each shift is modelled using as few concept changes as necessary. We then leverage the results developed by Lachapelle et al. (2023) and Xu et al. (2024) to prove that the proposed `SSAE` approach identifies *some* concepts, under suitable assumptions on distribution that generated the data. We also show how this allow to extract extract valid steering vectors, i.e. direction in the LLM representation that changes a single concept. We study the `SSAE` empirically on challenging language datasets and models, finding many settings where they outperform SAEs as well as other related steering methods that require supervision.

In sum, this work: 1) formalizes the problem of recovering interpretable concepts from sparse multi-concept shifts, from the lens of identifiability; 2) proposes the `SSAE` to model these sparse multi-concept shifts and establishes identifiability guarantees for these models based on sparsity regularization; 3) using multiple real-world language datasets and LLMs, empirically verifies the identifiability result and demonstrates the benefits of an identifiable model for accurately predicting target steered embeddings.

2 Problem formulation

We observe texts $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ that are generated from underlying *concept representations* $\mathbf{c} \in \mathcal{C} \subseteq \mathbb{R}^{d_c}$ through an unknown generative process $g : \mathcal{C} \to \mathcal{X}$ so that $\mathbf{x} = g(\mathbf{c})$. While we cannot observe the concept representation $\mathbf{c}$ of an observation $\mathbf{x}$, we have access to *learned representations* $\mathbf{z} = f(\mathbf{x})$, where the function $f : \mathcal{X} \to \mathcal{Z} \subseteq \mathbb{R}^{d_z}$ maps observations $\mathbf{x}$ to d_z-dimensional real vectors $\mathbf{z} \in \mathcal{Z}$, known as their *embeddings*. Throughout this paper, we consider the case where $f(\mathbf{x})$ comes from an autoregressive language model and is the embedding of the final token $\mathbf{x}_T$ in the residual stream after the final layer. We assume that the concepts $\mathbf{c}$ are *encoded* in the representations $\mathbf{z}$ through the unknown composite function $\mathbf{z} = f(g(\mathbf{c}))$. We consider concept perturbations,

$$\tilde{\mathbf{c}} := \mathbf{c} + \boldsymbol{\delta}^c; \qquad \boldsymbol{\delta}^c_k = \lambda \mathbf{e}_k, \tag{1}$$

where $\boldsymbol{\delta}^c$ is called the **concept shift** vector, λ is the magnitude of the perturbation, and $\boldsymbol{\delta}^c_k \neq 0$ for all perturbed concepts k.

Main goal. We want to map unlabelled concept shifts $\boldsymbol{\delta}^c$ to their corresponding vectors in the space of LLM activations. (Refer to Apx. A.2 for a formal treatment of steering.).

The key challenge is that we only observe texts ($\mathbf{x}$) and their embeddings ($\mathbf{z}$), and thus, we cannot directly learn a mapping from concepts shifts $\boldsymbol{\delta}^c$ to LLM activation shifts. A naive unsupervised approach is to fit an autoencoder to LLM embeddings $\mathbf{z}$ so that for any input, we can encode it in a latent space, implement the desired concept shift $\boldsymbol{\delta}^c$ in that space, and decode it to obtain a perturbed embedding $\tilde{\mathbf{z}}$. However, unless the autoencoder is guaranteed to encode embeddings $\mathbf{z}$ in a latent space that captures concepts, this naive approach will result in perturbations $\tilde{\mathbf{z}}$ that *do not reflect the desired concept shifts*. Unfortunately, unconstrained autoencoding objectives are non-identifiable (Hyvärinen & Pajunen, 1999), and sparse autoencoding objectives (Cunningham et al., 2023) may not be able to invert embeddings to potentially billions of concepts. As such, there is no guarantee that such approaches recover latent concepts from observed embeddings $\mathbf{z}$, posing a risk for steering.

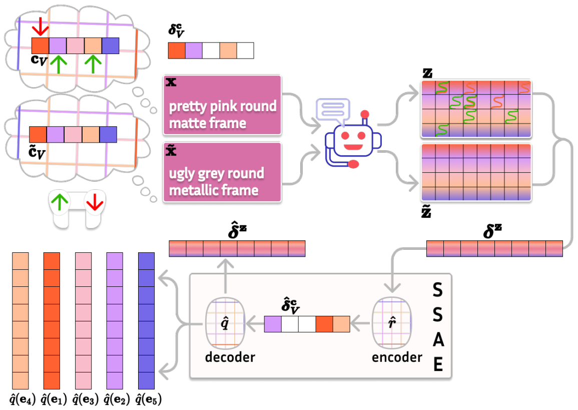
$\mathbf{c}_V$
$\boldsymbol{\delta}^c_V$
$\tilde{\mathbf{c}}_V$
$\mathbf{x}$
pretty pink round matte frame
$\tilde{\mathbf{x}}$
ugly grey round metallic frame
$\mathbf{z}$
$\tilde{\mathbf{z}}$
$\hat{\boldsymbol{\delta}}^z$
$\boldsymbol{\delta}^z$
$\hat{q}$
$\hat{\boldsymbol{\delta}}^c_V$
$\hat{r}$
decoder
encoder
S S A E
$\hat{q}(e_4)$ $\hat{q}(e_1)$ $\hat{q}(e_3)$ $\hat{q}(e_2)$ $\hat{q}(e_5)$

Figure 1: `SSAE`s map multi-concept shifts in embedding space to concept shifts, leveraging the latter's sparsity, thereby recovering steering vectors. The learnt steering vectors are identified up to permutation and scaling.

Key idea. We develop an identifiable autoencoding method called sparse shift encoders (`SSAEs`). The key idea behind `SSAEs` are multi-concept shift data, illustrated in Figure 1. As an example, consider two text snippets $\mathbf{x}$: `pretty pink round matte frame` and $\tilde{\mathbf{x}}$: `ugly grey round metallic frame`. Both $\mathbf{x}$ and $\tilde{\mathbf{x}}$ seem to be encoded by 5 concepts each– a descriptive adjective (pretty/ugly), colour (pink/grey), shape (round), texture (matte/metallic), and object (frame). However, when we consider what has changed from $\mathbf{x}$ to $\tilde{\mathbf{x}}$, it's a smaller set of (3) concepts, and it is also possible to imagine pairs which vary by just a single concept. An `SSAE` provably recovers these inter-sample concept shifts by regularizing the inferred concept shifts $\hat{\boldsymbol{\delta}}^c$ to be sparse.

3 Sparse Shift Autoencoders (SSAEs)

We start by describing the data-generating process and the set of concepts learnable via inter-sample differences, before proposing a method for learning steering vectors for these concepts. Following Locatello et al. (2020b), we consider paired observations $(\mathbf{x}, \tilde{\mathbf{x}})$ assumed to be sampled from the following generative process:

$$S \sim p(S), \quad (\mathbf{c}, \tilde{\mathbf{c}}) \sim p(\mathbf{c}, \tilde{\mathbf{c}} \mid S), \tag{2}$$

$$\mathbf{x} := g(\mathbf{c}), \quad \tilde{\mathbf{x}} := g(\tilde{\mathbf{c}})\,, \tag{3}$$

where $S \subseteq \{1, \ldots, d_c\}$ denotes the subset of concepts that vary between $\mathbf{x}$ and $\tilde{\mathbf{x}}$. More precisely, $p(\mathbf{c}, \tilde{\mathbf{c}} \mid S)$ is such that, with probability one, $\mathbf{c}_k = \tilde{\mathbf{c}}_k$ for all $k \notin S$. Crucially, across each pair of observations, an unknown set of concepts changes. For what follows, it will be useful to define $V \subseteq \{1, \ldots, d_c\}$ to be the set of *varying concepts*:

$$V := \bigcup_{S:\, p(S)>0} S\,. \tag{4}$$

The set V thus contains the concepts that can change in a pair $(\mathbf{x}, \tilde{\mathbf{x}})$. Even though concepts outside V are assumed to remain fixed *within* a pair $(\mathbf{x}, \tilde{\mathbf{x}})$, they can still vary *across* pairs. Without loss of generality, assume that $V := \{1, \ldots, |V|\}$.

Next, we consider difference vectors $\boldsymbol{\delta}^z := f(\tilde{\mathbf{x}}) - f(\mathbf{x}) = \tilde{\mathbf{z}} - \mathbf{z}$. These vectors capture how underlying concept differences between a pair of inputs $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are represented in the space of LLM embeddings. An important assumption made by (Rajendran et al., 2024; Park et al., 2023) helps us relate these difference vectors to concept shifts:

Assumption 1 (Linear representation hypothesis (LRH))**.** *The generative process $g : \mathcal{C} \to \mathcal{X}$ and the learned encoding function $f : \mathcal{X} \to \mathcal{Z}$ are such that $f \circ g : \mathcal{C} \to \mathcal{Z}$ is linear, implying there exists a $d_z \times d_c$ real matrix $\mathbf{A}$ such that:*

$$\mathbf{z} = f(g(\mathbf{c})) = \mathbf{A}\mathbf{c}\,. \tag{5}$$

Put simply, the LRH says that the learned representation $\mathbf{z}$ *linearly encodes concepts*. Consequently, difference vectors $\boldsymbol{\delta}^z$ are also linearly related to concept shifts so that $\boldsymbol{\delta}^z = \mathbf{A}\boldsymbol{\delta}^c$. A long line of work provides evidence for this hypothesis (c.f. Rumelhart & Abrahamson (1973); Hinton et al. (1986); Mikolov et al. (2013); Ravfogel et al. (2020b)). More recently, theoretical work justifies why linear properties could arise in these models (c.f. Jiang et al. (2024); Roeder et al. (2021); Marconato et al. (2024)). Section 6 provides a full list of related work, while Apx. A.7 provides an explanation of the equivalence between LRH's different interpretations.

Sparse Shift Autoencoders (`SSAEs`) take as input the observed difference vectors $\boldsymbol{\delta}^z_V$ and model them with an affine encoder $r : \mathbb{R}^{d_z} \to \mathbb{R}^{|V|}$ and an affine decoder $q : \mathbb{R}^{|V|} \to \mathbb{R}^{d_z}$ such that,

$$\hat{\boldsymbol{\delta}}^c_V := r(\boldsymbol{\delta}^z) := \mathbf{W}_e(\boldsymbol{\delta}^z - \mathbf{b}_d) + \mathbf{b}_e\,; \tag{6}$$

$$\hat{\boldsymbol{\delta}}^z := q(\hat{\boldsymbol{\delta}}^c_V) := \mathbf{W}_d\hat{\boldsymbol{\delta}}^c_V + \mathbf{b}_d\,. \tag{7}$$

The representation $r(\boldsymbol{\delta}^z)$ predicts $\boldsymbol{\delta}^c_V$, i.e., the concept shifts corresponding to $\boldsymbol{\delta}^z$, with $\boldsymbol{\delta}^c_V = (\boldsymbol{\delta}^c_i)_{i\in V}$ the subvector of $\boldsymbol{\delta}^c$ corresponding to the index set V. That is, `SSAEs` map differences in embedding space to their constituent concept shifts, focusing *only on the varying concepts*.

We train `SSAEs` to solve the following constrained problem:

$$(\hat{r}, \hat{q}) \in \arg\min_{r,q} \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}\left[||\boldsymbol{\delta}^z - q(r(\boldsymbol{\delta}^z))||_2^2\right] \tag{8}$$

$$\text{s.t. } \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}||r(\boldsymbol{\delta}^z)||_0 \leq \beta\,, \tag{9}$$

where Eqn. (8) is the standard auto-encoding loss that encourages good reconstruction and Eqn. (9) is a regularizer that encourages the predicted concept shift vector $\hat{\boldsymbol{\delta}}^c_V := \hat{r}(\boldsymbol{\delta}^z)$ to be sparse. Since the ℓ_0-norm is non-differentiable, in practice we replace it by an ℓ_1-norm leading to the following relaxed sparsity constraint:

$$\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}||r(\boldsymbol{\delta}^z)||_1 \leq \beta \,. \tag{10}$$

We then approximately solve this constrained problem by finding a saddle point of its Lagrangian using the ExtraAdam algorithm (Gidel et al., 2020) as implemented by Gallego-Posada & Ramirez (2022). Apx. B.1.2 provides a detailed discussion of the benefits of constraints as opposed to penalty to regularize objectives. Appropriate normalization is crucial for enforcing sparsity using the ℓ_1-norm. Further details, including other implementation aspects, are discussed in Section 5 and Apx. B.

Identifiability of `SSAE`s. In Section 4 we will show that, under suitable assumptions on the data-generating process and a suitable choice of β, the ℓ_0-regularized problem of Eqns. (8) and (9) is guaranteed to learn a $(\hat{r}, \hat{q})$ such that $\hat{r}(\boldsymbol{\delta}^z) = \mathbf{P}\mathbf{D}\boldsymbol{\delta}^c_V$ where $\mathbf{D}$ is an invertible diagonal matrix, $\mathbf{P}$ is a permutation matrix. In other words, the learned representation $\hat{r}(\boldsymbol{\delta}^z)$ can be related to the ground-truth concept shift vector $\boldsymbol{\delta}^c_V$ (considering only the varying concepts V) via a permutation-scaling matrix. We will later see how sparsity regularization is crucial for this to happen. Although our theoretical analysis assumes the learned representation has size $|V|$, we find in Apx. B.4 that, in practice, our method maintains a reasonable degree of identifiability when the representation size is larger than $|V|$. Linking identifiability back to steering, we conclude by showing how the identifiability guarantee implies that $\hat{q}(\mathbf{e}_k) \in \mathbb{R}^{d_z}$ are valid steering vectors for concepts in V.

4 Identifiability analysis

This section explains why we expect the representation learned in Eqn. (8) to identify the ground-truth concept shift vector $\boldsymbol{\delta}^c_V$ up to permutation and rescaling. To do so, we first demonstrate that, under suitable assumptions, the learned representation $\hat{r}(\boldsymbol{\delta}^z)$ identifies the ground-truth concept shift $\boldsymbol{\delta}^c_V$ *up to an invertible linear transformation* when we do not use sparsity regularization. Second, we show that by adding sparsity regularization, the learned representation identifies $\boldsymbol{\delta}^c_V$ *up to permutation and element-wise rescaling*.

Recall that, since we expect $d_c \gg d_z$, we cannot assume $\mathbf{A}$ to be injective; the same issue that arises when trying to encode $\mathbf{c}$ from $\mathbf{z}$. Fortunately, we do not need to make this assumption, thanks to the following decomposition. Let $\bar{V} := [d_c] \setminus V$ be the complement of V. Then:

$$\begin{aligned}\boldsymbol{\delta}^z &= \mathbf{A}\boldsymbol{\delta}^c = \mathbf{A}_V\,\boldsymbol{\delta}^c_V + \mathbf{A}_{\overline{V}}\boldsymbol{\delta}^c_{\overline{V}}\\ &= \mathbf{A}_V\,\boldsymbol{\delta}^c_V\;,\end{aligned} \tag{11}$$

where we used the fact that $\boldsymbol{\delta}^c_{\overline{V}} = 0$, by definition of V. By considering difference vectors, we focus on disentangling *only* the varying concepts, the linear entanglement of which the submatrix $\mathbf{A}_V$ captures. Since $|V| \leq d_c$, we can make the assumption that mixing function $\mathbf{A}_V$ is injective.

Assumption 2. *The matrix $\mathbf{A}_V \in \mathbb{R}^{d_z \times |V|}$ is injective.*

Note that this implies that $d_z \geq |V|$, i.e., $\mathbf{z}$ has at least as many dimensions as there are varying concepts. This is feasible given that d_z is typically around 10^3 (e.g., in LLMs), supporting a large set of varying concepts V.

To prove linear identifiability, we will need one more assumption. Let Δ^c_V be the support of the random vector $\boldsymbol{\delta}^c_V$. We will require that this support is diverse enough so that its linear span is equal to the whole space $\mathbb{R}^{|V|}$.

Assumption 3. *span*$(\Delta^c_V) = \mathbb{R}^{|V|}$.

With these assumptions, we can show linear identifiability by reusing proof strategies that are now common in the literature on identifiable representation learning (Khemakhem et al., 2020a; Roeder et al., 2021; Ahuja et al., 2022; Xu et al., 2024).

Proposition 1 (Linear identifiability). *Suppose $(\hat{r}, \hat{q})$ is a solution to the unconstrained problem of Eqn. (8). Under Apx. A.7 and Asm. 2 and 3, there exists an invertible matrix $\mathbf{L} \in \mathbb{R}^{|V|\times|V|}$ such that $\hat{q} = \mathbf{A}_V\mathbf{L}$ and $\hat{r}(\mathbf{z}) = \mathbf{L}^{-1}\mathbf{A}_V^{+}\mathbf{z}$ for all $\mathbf{z} \in \mathrm{Im}(\mathbf{A}_V)$, where $\mathrm{Im}(\mathbf{A}_V)$ is the image of $\mathbf{A}_V$.*[1]

We prove Prop. 1 in Apx. A.4. The result follows naturally from the linear representation hypothesis in Apx. A.7, but requires Asm. 2 and 3 for a complete proof. Rajendran et al. (2024) prove a similar result, showing that linear subspaces of representations that represent concepts are linearly identified from concept-conditional observations.

Identifiability up to permutation and rescaling. To go from identifiability up to linear transformation to identifiability up to permutation and rescaling, we need to make further assumptions. Let $\mathcal{S}$ be the support of the distribution $p(S)$, i.e., $\mathcal{S} := \{S \subseteq [d_c] \mid p(S) > 0\}$. The following is based on Lachapelle et al. (2023) and Xu et al. (2024).

Assumption 4 (Sufficient diversity of multi-concept shifts)**.** *The following two conditions hold.*

1. *(Sufficient support variability): For every varying concept $k \in V$, we have*

$$\bigcup_{S\in\mathcal{S}|k\notin S} S = V \setminus \{k\} \quad \forall k \in V\,; \tag{12}$$

2. *(Distribution $\mathbb{P}_{\boldsymbol{\delta}^c_S|S}$ continuous): For all $S \in \mathcal{S}$, the conditional distribution $\mathbb{P}_{\boldsymbol{\delta}^c_S|S}$ can be described using a probability density with respect to the Lebesgue measure on $\mathbb{R}^{|S|}$.*

Without the first assumption, two concepts $k, j \in V$ might always change together, meaning there is no data pair in which only one of them varies independently. Intuitively, this would prevent the model from disentangling them effectively. Importantly, our assumption accommodates a broad range of scenarios. E.g., it is not necessarily violated even in an extreme case where $|V| - 1$ concepts change in each pair. Moreover, it allows for the presence of statistically dependent concepts. The second criterion ensures the distribution $\mathbb{P}_{\boldsymbol{\delta}^c_S|S=s}$ does not concentrate mass on a subset of $\mathbb{R}^{|S|}$ of Lebesgue measure zero. In Apx. A.6, we provide examples of distributions that meet or fail the assumption.[2]

We are now ready to state the main identifiability result of this section. We note that its proof relies to a large extent on an existing result by Lachapelle et al. (2023).

Proposition 2 (Identifiability up to permutation). *Suppose $(\hat{r}, \hat{q})$ is a solution to the constrained problem of Eqns. (8) and (9) with $\beta = \mathbb{E}||\boldsymbol{\delta}^c_V||_0$. Under Apx. A.7 and Asm. 2 to 4, there exists an invertible diagonal matrix and a permutation matrix $\mathbf{D}, \mathbf{P} \in \mathbb{R}^{|V|\times|V|}$ such that $\hat{q} = \mathbf{A}_V\mathbf{D}\mathbf{P}$ and $\hat{r}(\mathbf{z}) = \mathbf{P}^\top\mathbf{D}^{-1}\mathbf{A}_V^{+}\mathbf{z}$ for all $\mathbf{z} \in \mathrm{Im}(\mathbf{A}_V)$, where $\mathrm{Im}(\mathbf{A}_V)$ is the image of $\mathbf{A}_V$.*

Proof sketch. We outline the proof here and defer the full details to Apx. A.5. We first show that all optimal solutions of the constrained problem must reach a reconstruction loss of zero. This means that optimal solutions to the constrained problem are also optimal for the unconstrained one. Thus, these solutions must identify $\mathbf{A}_V$ up to linear transformation, by Prop. 1. We can then rewrite the constraint as $\mathbb{E}||\boldsymbol{L}^{-1}\boldsymbol{\delta}^c_V||_0 \leq \beta = \mathbb{E}||\boldsymbol{\delta}^c_V||_0$. Here, we can reuse an argument initially proposed by Lachapelle et al. (2023) to leverage this inequality to conclude that $\boldsymbol{L}^{-1}$ must be a permutation-scaling matrix. For completeness, we present this argument in Lemma 4. It shows that, applying the matrix $\boldsymbol{L}^{-1}$ to $\boldsymbol{\delta}^c_V$ always strictly increases its expected sparsity, *unless* $\boldsymbol{L}^{-1}$ is a permutation-scaling matrix. Thus, to satisfy the inequality, $\boldsymbol{L}$ must be a permutation-scaling matrix.

Extracting steering vectors. Under Apx. A.7 and Asm. 2 to 4, Prop. 2 shows that $\hat{q} = \mathbf{A}_V\mathbf{D}\mathbf{P}$. From this identifiability result, we can see that,

$$\mathbf{z} + \hat{q}(\mathbf{e}_k) = \mathbf{A}\mathbf{c} + \mathbf{D}_{\pi(k),\pi(k)}\mathbf{A}\mathbf{e}_{\pi(k)} = \mathbf{A}(\mathbf{c} + \lambda\mathbf{e}_{\pi(k)}) = f(g(\mathbf{c} + \lambda\mathbf{e}_{\pi(k)})) = f(g(\tilde{\mathbf{c}}_{\pi(k),\lambda}))\,,$$

where $\lambda := D_{\pi(k),\pi(k)}$. In other words, when we add the decoded basis vector $\mathbf{e}_k$ to any embedding $\mathbf{z}$, i.e., add the k-column of the linear decoding matrix, the resulting vector represents $f(g(\tilde{\mathbf{c}}_{\pi(k),\lambda}))$,

[1] We might not have $\hat{r}(\mathbf{z}) = \mathbf{L}\mathbf{A}_V^{+}\mathbf{z}$ for $\mathbf{z} \notin \mathrm{Im}(\mathbf{A}_V)$, since the behavior of $\hat{r}$ is unconstrained by the objective outside the support of $\boldsymbol{\delta}^z$, i.e., outside $\mathrm{Im}(\mathbf{A}_V)$.

[2] See Lachapelle et al. (2023) for a strictly weaker but more technical assumption that is also sufficient for Prop. 2.

the embedding representation of the $\pi(k)$-th concept steered. Thus, identifiability directly leads to accurate unsupervised steering. A practitioner can use this result to try each steering vector $q(\mathbf{e}_k)$ in turn, generate tokens with an LLM, and directly interpret the changes to interpret the concept that was steered. By contrast, in Apx. A.3, we show that linear identifiability is insufficient to recover steering vectors up to permutation without the need for further labelled examples.

5 Empirical Studies

Using multiple language datasets and LLMs, our empirical studies proceed in two stages. First, we validate the core theoretical claim: do SSAEs recover concepts up to simple permutation and scaling transformations, and do they their decoder columns align with intended steering directions? Second, we turn to challenging language datasets where the target concepts are complex functions of prompts. Here, we evaluate whether SSAEs can identify steerable concepts across a range of complex real-world datasets including Bias in Bios (De-Arteaga et al., 2019), as well as refusal and sycophancy (Panickssery et al., 2024). We also include a case study on mitigating gender bias in text generation using Bias in Bios. We find that compared to SAEs (He et al., 2024; Anil et al., 2024; Biderman et al., 2023), SSAEs result in more accurate steering predictions.

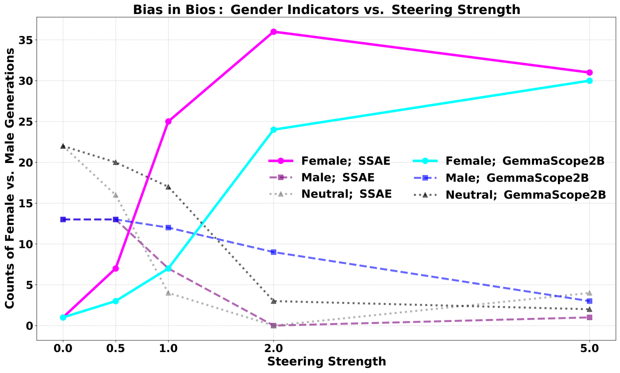

Figure 2: **SSAEs achieve earlier and stronger transitions to female indicators in generated text** (by strength 1.0), while GemmaScope2B requires stronger interventions.

Implementation details. We implement the autoencoder described in Eqn. (7) with an encoding dimension of $|V|$ when it is assumed that the number of concepts to be discovered in data is known, or $K \leq d_z$ in the more general case. As mentioned in Section 4, key to identifying steering vectors is the sparsity constraint from Eqn. (10). It is enforced using the cooper library (Gallego-Posada & Ramirez, 2022). For details on implementation, refer to Apx. B.

Experimental setup. We use text-based paired observations $(\mathbf{x}, \tilde{\mathbf{x}})$, to extract the final-layer token embedding (which is linearly identifiable following Roeder et al. (2021)) from one of Llama-3.1-8B (Llama Team et al., 2024), or Gemma2-2B (Anil et al., 2024), or Pythia-70M (Biderman et al., 2023) and use the embedding of the last token as the representation, following Ma et al. (2023), to obtain sentence representations $(f(\mathbf{x}), f(\tilde{\mathbf{x}}))$. We focus on evaluating LLM embeddings on language datasets here. In Apx. B.7, we validate the same conclusions with synthetic experiments.

Baselines. To validate our theory, we consider four different baselines: 1) SAEs trained on large-scale data (LlamaScope (He et al., 2024), GemmaScope (Lieberum et al., 2024), or PythiaSAE (EleutherAI, 2023) depending on the LLM the activations are obtained from), 2) an affine Autoencoder (aff) with an identical architecture as SSAEs but with no sparsity regularisation, 3) PCA on the same multi-concept difference vectors that SSAEs use, obtained from the last hidden layer (Liu et al., 2024), and 4) Mean Difference (MD) vectors that use paired observations differing in a single concept to compute $\frac{1}{n}\sum_{i=1}^{n}(\tilde{\mathbf{z}}_k^{(i)} - \mathbf{z}_k^{(i)}) = \frac{1}{n}\sum_{i=1}^{n}\lambda^{(i)}\mathbf{A}\mathbf{e}_k = \bar{\lambda}\mathbf{A}\mathbf{e}_k$ as the steering vector for concept k, which are used in contrastive activation addition methods (Panickssery et al., 2024) applied to different layers of an LLM. We denote the concept-steered embeddings produced by each method as $\tilde{\mathbf{z}}_{\texttt{SSAE}}$ (SSAE), $\tilde{\mathbf{z}}_{\texttt{SAE}}$ for the relevant SAE, $\tilde{\mathbf{z}}_{\texttt{PCA}}$ (PCA), $\tilde{\mathbf{z}}_{\texttt{aff}}$ (aff), and $\tilde{\mathbf{z}}_{\texttt{MD}}$ (MD). To compare discovery of meaningful concepts on complex real-world datasets, we compare against SAEs.

Evaluation criteria. We measure the *degree of identifiability* via the Mean Correlation Coefficient (MCC) (Hyvarinen & Morioka, 2016; Khemakhem et al., 2020b), which computes the highest average correlation between each learned latent dimension and the true latent dimension and equals 1.0 when they are aligned perfectly up to permutation and scaling. To evaluate the effect of applying a steering vector in embedding space, we consider held-out single concept shift data $(\mathbf{x}, \tilde{\mathbf{x}}_k)$ and evaluate how well steering vectors learnt using multi-concept shifts steer $f(\mathbf{x})$ towards $f(\tilde{\mathbf{x}}_k)$. Then we measure the accuracy of steering by comparing $\hat{\tilde{\mathbf{z}}}_k := f(\mathbf{x}) + \hat{q}(\mathbf{e}_{\pi(k)})$ and $f(\tilde{\mathbf{x}}_k)$ using cosine similarity as a measure of semantic similarity by searching over the columns of the decoding matrix $\mathbf{W}_d$. Refer to Apx. B.2 and Apx. B.3 for details. We also report reconstruction errors in Apx. B.1.4.

Validation of Theoretical Claims. We consider simple semi-synthetic datasets with single words where we can assume the number of underlying concept variations in pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ with a diverse range of concept variations. Datasets are named as: identifier of the dataset indicating why we consider it, followed by $|V|$ and $\max(|S|)$: IDENTIFIER($|V|$, $\max|S|$). Details on datasets can be found in Apx. B.1.3. Briefly, LANG$(1, 1)$ (e.g., *eng* → *french*) and GENDER$(1, 1)$ (e.g., *masculine* → *feminine* vary a single concept between $\mathbf{x}$ and $\tilde{\mathbf{x}}$. To stress-test the viability of our assumptions, we also consider the the multiple-choice track of TruthfulQA (Lin et al., 2022), creating $(\mathbf{x}, \tilde{\mathbf{x}})$ pairs by assigning $\mathbf{x}$ to be the question paired with a wrong answer that mimics human falsehoods, and $\tilde{\mathbf{x}}$ to be a question paired with the correct answer to capture the variation of the concept truthfulness from *false* → *true*. We include our findings for the activations stemming from Llama-3.1-8B here.

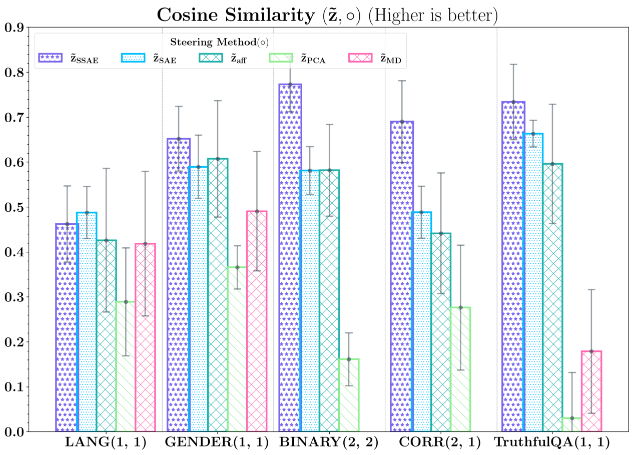

Figure 3: **A higher MCC value of the estimated decoder is associated with a greater cosine similarity.** Embeddings steered with vectors from a more disentangled decoder align more closely to target embeddings.

Table 1: **The mean MCC of the estimated decoder is close to** 1 across all datasets considering observations $(f(\mathbf{x}), f(\tilde{\mathbf{x}}))$, even for correlated concepts in CORR(2, 1).

	SSAE	aff
LANG$(1, 1)$	0.995 ± 0.001	0.985 ± 0.004
GENDER$(1, 1)$	0.993 ± 0.000	0.961 ± 0.000
BINARY$(2, 2)$	0.991 ± 0.001	0.936 ± 0.000
CORR$(2, 1)$	0.991 ± 0.001	0.928 ± 0.077
TruthfulQA	0.952 ± 0.006	0.885 ± 0.006

Table 2: **Sparsity regularisation is crucial** to identifying steering vectors, as demonstrated by using pairs of further entangled observations $(\mathbf{L}f(\mathbf{x}), \mathbf{L}f(\tilde{\mathbf{x}}))$.

	SSAE	aff
LANG$(1, 1)$	0.990 ± 0.000	0.876 ± 0.007
GENDER$(1, 1)$	0.991 ± 0.000	0.884 ± 0.005
BINARY$(2, 2)$	0.990 ± 0.001	0.796 ± 0.000
CORR$(2, 1)$	0.990 ± 0.001	0.630 ± 0.010
TruthfulQA	0.932 ± 0.008	0.751 ± 0.012

How well does `SSAE` identify steering vectors? For this first evaluation, we focus on the importance of sparsity regularisation for identifiability. We compare `SSAE` to the `aff` baseline that omits sparsity regularization, expecting that this baseline should result in lower MCC values. Since identifiability implies that learned decoders across such runs should be related by permutation-scaling transformations (Rolinek et al., 2019; Duan et al., 2019), we report the MCC between pairs of learned decoders as we train them using different random initializations, calling this variant of the metric MCC_D. We use 10 decoder pairs from 5 seeds for selected model hyperparameters. Table 1 shows that `SSAE` achieves consistently high MCC values, empirically corroborating Prop. 2, assuming a known $|V|$. As a sensitivity analysis, we further entangle the LLM embeddings by applying a dense linear invertible transformation $\mathbf{L}$ to the embeddings to generate $(\mathbf{L}f(\mathbf{x}), \mathbf{L}f(\tilde{\mathbf{x}}))$. As expected, Section 5 demonstrates that this widens the gap between `SSAE` and the affine baseline. The worsening performance of `aff` after the entanglement is applied suggests that LLM representations might already somewhat disentangle some concepts or encode them through sparse or simple transformations. Next, we evaluate whether the benefits of a higher MCC_D translate to performance improvements on steering embeddings to be more similar to those of the target concept.

Does identifiability translate to *better* steering? We hold out pairs $(\mathbf{x}, \tilde{\mathbf{x}}_k) \forall k \in V$, each varying by a single concept, and compare the cosine similarity between the steered embeddings and target embeddings. Figure 3 illustrates that SSAE's higher MCC_D performance generally translates to more accurate steering, with significant advantages over all related methods in the more challenging $\text{BINARY}(2, 2)$ and $\text{CORR}(2, 1)$ settings where multiple or correlated concepts change. Figure 3 also reveals that even slight differences in MCC_D values can translate into pronounced variations in steering accuracy. Next, we evaluate *out-of-distribution* (OOD) steering accuracy, based on the hypothesis that steering vectors that disentangle a single concept should transfer to different domains. For this evaluation, we learn a steering vector from *eng* → *french* using the $\text{BINARY}(2, 2)$ or $\text{CORR}(2, 1)$ dataset, where language changes are shown for occupation-related works, and use the steering vector on the $\text{LANG}(1, 1)$ dataset consisting of words related to household objects. Figure 4 shows that the steering vectors learned by SSAEs transfer effectively to OOD datasets while SAEs do not perform better than simple baselines, further substantiating the importance of identifiability for unsupervised steering.

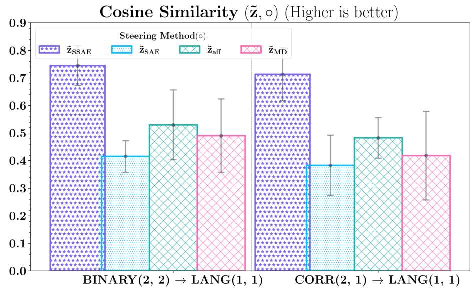

Figure 4: Embeddings steered using SSAE show **higher OOD generalisation performance**. SAEs (LlamaScope) generalise worse than other simpler baselines.

The identifiability theory requires assuming that the encoder dimension is known (equal to $|V|$), and relying on the last layer's embeddings for steering since the LRH – a key component of the identifiability results – is better theoretically motivated (c.f. (Marconato et al., 2024)) at the last layer. To test sensitivity to these assumptions, we conducted further studies training SSAEs on different layers of an LLM (Apx. B.5), and comparing values of MCC_D versus steering accuracy as we increase the encoding dimension size past $|V|$, finding that for encoding dimensions $> |V|$, there is an increase in steering accuracy even though MCC_D values drop substantially (see Apx. B.4 for details). Following these promising findings, in the next section, we conduct experiments with the maximum encoding dimension (equal to the embedding size) on challenging language datasets.

Real-World Steering. To demonstrate the utility of SSAEs in realistic settings, we consider three well-studied language datasets that contain more abstract concepts: (i) LLM sycophancy and (ii) LLM refusal benchmarks (Panickssery et al., 2024; Perez et al., 2022) consisting of multiple-choice questions with two answers demonstrating either the behaviour of interest or its opposite, and (iii) Bias in Bios dataset (De-Arteaga et al., 2019), consisting of biographies that differ by occupation and gender. We focus on Pythia-70M (Biderman et al., 2023) and Gemma2-2B (Anil et al., 2024) embeddings as inputs to train SSAEs. A key difference in these evaluations is that we train SSAEs with the maximum encoding size, giving the model more flexibility to recover concepts.

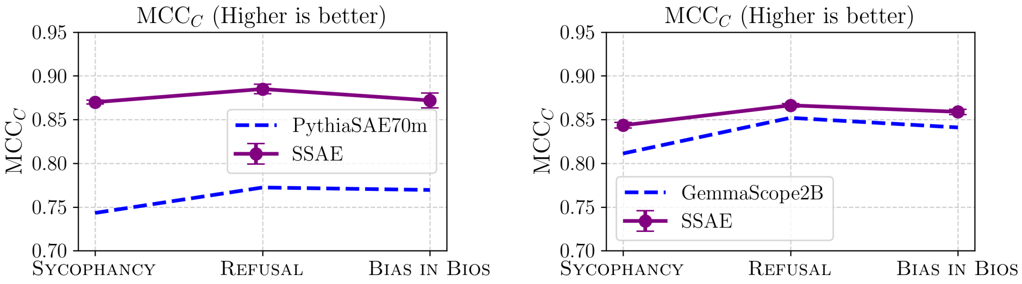

Figure 5: MCC_C values between learned activations and concept labels indicate that **SSAEs can outperform SAEs in identifying ground-truth concepts in the dataset**

MCC and steering accuracy. As before, we study steering accuracy by predicting target embeddings based on held-out contrastive prompts that vary by a single concept, e.g., refusal behavior. Since these datasets contain labels for the target concepts, we report the MCC_C between the predicted

SSAE encodings and true concept labels. SSAEs demonstrate the benefits of identifiability, yielding systematic gains across model scales and concept domains, as seen in Figure 5.

Text generation and steering. We study how SSAEs can be used to steer Gemma-2B's text generation using the Bias in Bios dataset, since we can effectively aggregate and summarize the effects of steering across many examples by simply counting gender pronouns to categorise them as either male or female dominated, or as neutral where counts are equal or ambiguous. Figure 2 shows that SSAEs steering vectors lead to more effective generation of female pronouns than those of GemmaScope.

6 Related work

Linear representation hypothesis. This paper builds on the linear representation hypothesis that language models encode concepts linearly. Several papers provide empirical evidence for this hypothesis (Mikolov et al., 2013; Gittens et al., 2017; Ethayarajh et al., 2019; Allen & Hospedales, 2019; Seonwoo et al., 2019; Burns et al., 2024; Li et al., 2024; Moschella et al., 2023; Tigges et al., 2023; Nissim et al., 2020; Ravfogel et al., 2020a; Park et al., 2023; 2024). Recent work also provides theoretical justification for why linear properties might consistently emerge across models that perform next-token prediction (Roeder et al., 2021; Jiang et al., 2024; Marconato et al., 2024).

Interpretability of LLMs. This paper contributes to the literature on interpretability and steering of LLMs. Much of the work on finding concepts in LLM representations for steering relies on supervision, either from paired observations with a single-concept shift (Panickssery et al., 2024; Turner et al., 2024; Rimsky et al., 2024; Li et al., 2024) or from examples of target LLM completions to prompts (Subramani et al., 2022). This prior work also focuses on applying the same steering vector to all examples, implicitly relying on the linear representation hypothesis as justification. In contrast, we make the assumption precise, and show how it leads to steering vectors. This paper also departs from supervised learning and focuses on learning with limited supervision. In this way, we propose a method that is similar to sparse autoencoders (SAEs) (Templeton et al., 2024; Engels et al., 2024; Cunningham et al., 2023; Rajamanoharan et al., 2024; Gao et al., 2024). In contrast, our proposed method fits concept shifts, and provably identifies steering vectors while SAEs may not enjoy identifiability guarantees.

Causal representation learning. Finally, this paper builds on causal representation learning results that leverage sparsity constraints. Ahuja et al. (2022), Locatello et al. (2020a), and Brehmer et al. (2022) consider sparse latent perturbations and paired observations. In contrast, we focus on learning from multi-concept shifts. Lachapelle et al. (2022) focus on sparse interventions and sparse transitions in temporal settings, while Lachapelle et al. (2023), Layne et al. (2024), Xu et al. (2024), and Fumero et al. (2023) leverage sparse dependencies between latents and tasks. In this paper, we adapt these assumptions and technical results for a novel setting: discovering steering vectors from LLM representations based on concept shift data. In work that is closest to ours, Rajendran et al. (2024) recover linear subspaces that capture concepts up to linear transformations using concept-conditional datasets, and Goyal et al. (2025) develop an identifiable contrastive learning approach to discover behavior-mediating concepts, but cannot extract steering vectors. In contrast, we focus on multi-concept shifts and show how these lead to identifiable steering vectors. Our work also complements classical sparse coding results (see Apx. A.8), which rely on geometric assumptions on the dictionary.

7 Conclusion

We propose Sparse Shift Autoencoders (SSAEs) for discovering accurate steering vectors from multi-concept paired observations as an alternative to both SAEs, and approaches relying on supervised data. Key to this result are the identifiability guarantees that the SSAE enjoys as a consequence of considering sparse concept shifts. We study the SSAE empirically on several real language tasks, and find evidence that the method facilitates accurate steering learned via limited supervision. However, we stress that these experiments are intended to validate the identifiability results in Section 4 and their implications for accurate steering. Although we include effects of steering on generated text (Figure 2), to fully understand the impacts of the SSAE on steering research, especially LLM alignment, more evaluation is needed on embeddings from more complex datasets, and on more challenging tasks (e.g., MTEB (Muennighoff et al., 2023)). Rigorous large-scale evaluations on expansive real-world benchmarks are a promising avenue for future work.

References

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, and Praneeth Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization, 2014. URL https://arxiv.org/abs/1310.7991. Cited on page 27.

Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35:15516–15528, 2022. Cited on pages 4, 9, and 25.

Carl Allen and Timothy Hospedales. Analogies explained: Towards understanding word embeddings, 2019. URL https://arxiv.org/abs/1901.09813. Cited on page 9.

Evan Anders, Clement Neo, Jason Hoelscher-Obermaier, and Jessica N. Howard. Sparse autoencoders find composed features in small toy models. https://www.lesswrong.com/posts/a5wwqza2cY3W7L9cj/sparse-autoencoders-find-composed-features-in-small-toy, 2024. Cited on pages 27 and 94.

Rohan Anil, Zhenzhong Lan, Olivier J. Hénaff, Anastasios Angelopoulos, Aakanksha Chowdhery, Yujia Li, Basil Mustafa, Daniel Freeman, Henryk Michalewski, Sara Hooker, et al. Gemma 2: Open models based on gemini research and technology. https://ai.google.dev/gemma, 2024. Accessed: 2025-09-24. Cited on pages 6 and 8.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717. Cited on pages 1 and 34.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. Cited on page 27.

Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive approximation*, 28(3):253–263, 2008. Cited on page 26.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, Eric Hallahan, Jesse Mu, Kyle O'Brien, Eric Burns, Leo Gao, Sid Black, and Connor Leahy. Pythia: A suite for analyzing language models across training and scaling. https://arxiv.org/abs/2304.01373, 2023. arXiv:2304.01373. Cited on pages 6 and 8.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning, 2022. URL https://arxiv.org/abs/2203.16437. Cited on page 9.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html. Cited on pages 27 and 28.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2024. URL https://arxiv.org/abs/2212.03827. Cited on page 9.

Emmanuel Candes and Terence Tao. Decoding by linear programming, 2005. URL https://arxiv.org/abs/math/0502327. Cited on page 26.

Emmanuel Candes, Justin Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements, 2005. URL https://arxiv.org/abs/math/0503066. Cited on page 26.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600. Cited on pages 1, 2, 9, and 20.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 120–128. ACM, January 2019. doi: 10.1145/3287560.3287572. URL http://dx.doi.org/10.1145/3287560.3287572. Cited on pages 6 and 8.

David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, 100(5): 2197–2202, 2003. Cited on page 26.

D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. doi: 10.1109/TIT.2006.871582. Cited on page 26.

Sunny Duan, Loic Matthey, Andre Saraiva, Nicholas Watters, Christopher P Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised model selection for variational disentangled representation learning. *arXiv preprint arXiv:1905.12614*, 2019. Cited on pages 7, 28, and 29.

EleutherAI. Sparse autoencoder for pythia-70m (32k features). https://huggingface.co/EleutherAI/sae-pythia-70m-32k, 2023. Accessed: 2025-09-24. Cited on page 6.

Joshua Engels, Isaac Liao, Eric J. Michaud, Wes Gurnee, and Max Tegmark. Not all language model features are linear, 2024. URL https://arxiv.org/abs/2405.14860. Cited on page 9.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Towards understanding linear word analogies, 2019. URL https://arxiv.org/abs/1810.04882. Cited on page 9.

Simon Foucart and Holger Rauhut. An invitation to compressive sensing. In *A mathematical introduction to compressive sensing*, pp. 1–39. Springer, 2013. Cited on page 26.

Marco Fumero, Florian Wenzel, Luca Zancato, Alessandro Achille, Emanuele Rodolà, Stefano Soatto, Bernhard Schölkopf, and Francesco Locatello. Leveraging sparse and shared feature activations for disentangled representation learning, 2023. URL https://arxiv.org/abs/2304.07939. Cited on page 9.

Jose Gallego-Posada and Juan Ramirez. Cooper: a toolkit for Lagrangian-based constrained optimization. https://github.com/cooper-org/cooper, 2022. Cited on pages 4, 6, 27, and 28.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024. Cited on pages 9 and 28.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html. Cited on page 1.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks, 2020. URL https://arxiv.org/abs/1802.10551. Cited on pages 4 and 28.

Alex Gittens, Dimitris Achlioptas, and Michael W. Mahoney. Skip-gram - Zipf + uniform = vector additivity. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 69–76, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1007. URL https://aclanthology.org/P17-1007. Cited on page 9.

Navita Goyal, Hal Daumé III, Alexandre Drouin, and Dhanya Sridhar. Causal differentiating concepts: Interpreting LM behavior via causal representation learning. *Neural Information Processing Systems (NeurIPS)*, 2025. Cited on page 9.

Rémi Gribonval and Morten Nielsen. Sparse representations in unions of bases. *IEEE transactions on Information theory*, 49(12):3320–3325, 2004. Cited on page 26.

Rémi Gribonval, Rodolphe Jenatton, Francis Bach, Martin Kleinsteuber, and Matthias Seibert. Sample complexity of dictionary learning and other matrix factorizations. *IEEE Transactions on Information Theory*, 61(6):3469–3486, 2015. Cited on page 27.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024. Cited on page 6.

Geoffrey E Hinton et al. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, pp. 12. Amherst, MA, 1986. Cited on pages 3 and 25.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica, 2016. URL https://arxiv.org/abs/1605.06336. Cited on pages 6 and 32.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 1999. Cited on page 2.

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models, 2024. URL https://arxiv.org/abs/2403.03867. Cited on pages 2, 3, 9, 25, and 33.

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020a. Cited on page 4.

Ilyes Khemakhem, Ricardo Pio Monti, Diederik P. Kingma, and Aapo Hyvärinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica, 2020b. URL https://arxiv.org/abs/2002.11537. Cited on pages 6 and 32.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. URL https://proceedings.mlr.press/v119/koh20a.html. Cited on page 1.

Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976. Cited on page 28.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pp. 428–484. PMLR, 2022. Cited on page 9.

Sebastien Lachapelle, T. Deleu, D. Mahajan, I. Mitliagkas, Y. Bengio, S. Lacoste-Julien, and Q. Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *International Conference on Machine Learning*, 2023. Cited on pages 2, 5, 9, 22, 24, and 94.

Elliot Layne, Jason Hartford, Sébastien Lachapelle, Mathieu Blanchette, and Dhanya Sridhar. Sparsity regularization via tree-structured environments for disentangled representations, 2024. URL https://arxiv.org/abs/2405.20482. Cited on page 9.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2024. URL https://arxiv.org/abs/2306.03341. Cited on page 9.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*, 2024. Cited on page 6.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL https://arxiv.org/abs/2109.07958. Cited on page 7.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering, 2024. URL https://arxiv.org/abs/2311.06668. Cited on page 6.

Llama Team et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783. Cited on pages 6 and 33.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations, 2019. URL https://arxiv.org/abs/1811.12359. Cited on page 28.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International conference on machine learning*, pp. 6348–6359. PMLR, 2020a. Cited on page 9.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises, 2020b. Cited on page 3.

David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. The randomized dependence coefficient. *Advances in neural information processing systems*, 26, 2013. Cited on page 32.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning llama for multi-stage text retrieval, 2023. URL https://arxiv.org/abs/2310.08319. Cited on page 6.

Emanuele Marconato, Sébastien Lachapelle, Sebastian Weichwald, and Luigi Gresele. All or none: Identifiable linear properties of next-token predictors in language modeling. *arXiv preprint arXiv:2410.23501*, 2024. Cited on pages 3, 8, 9, and 25.

Abhinav Menon, Manish Shrivastava, David Krueger, and Ekdeep Singh Lubana. Analyzing (in)abilities of saes via formal languages. *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2025. Cited on page 1.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090. Cited on pages 2, 3, 9, and 25.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication, 2023. URL https://arxiv.org/abs/2209.15430. Cited on page 9.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark, 2023. URL https://arxiv.org/abs/2210.07316. Cited on page 9.

Malvina Nissim, Rik van Noord, and Rob van der Goot. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497, June 2020. doi: 10.1162/coli_a_00379. URL https://aclanthology.org/2020.cl-2.7. Cited on page 9.

Charles O'Neill, Alim Gumran, and David Klindt. Compute optimal inference and provable amortisation gap in sparse autoencoders, 2025. URL https://arxiv.org/abs/2411.13117. Cited on page 27.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024. URL https://arxiv.org/abs/2312.06681. Cited on pages 6, 8, 9, and 34.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2023. Cited on pages 3, 9, and 33.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models, 2024. URL https://arxiv.org/abs/2406.01506. Cited on page 9.

Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. Steer LLM latents for hallucination detection. In *Forty-second International Conference on Machine Learning*, 2025. Cited on page 1.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022. URL https://arxiv.org/abs/2212.09251. Cited on page 8.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders, 2024. URL https://arxiv.org/abs/2404.16014. Cited on page 9.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning interpretable concepts: Unifying causal representation learning and foundation models, 2024. Cited on pages 3, 5, 9, 20, and 25.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647. Cited on page 9.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*, 2020b. Cited on pages 3 and 25.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024. Cited on pages 1, 9, and 20.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pp. 9030–9039. PMLR, 2021. Cited on pages 3, 4, 6, 9, and 25.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident), 2019. URL https://arxiv.org/abs/1812.06775. Cited on page 7.

David E Rumelhart and Adele A Abrahamson. A model for analogical reasoning. *Cognitive Psychology*, 5(1):1–28, 1973. Cited on pages 3 and 25.

Karin Schnass. Local identification of overcomplete dictionaries, 2015. URL https://arxiv.org/abs/1401.6354. Cited on page 27.

Yeon Seonwoo, Sungjoon Park, Dongkwan Kim, and Alice Oh. Additive compositionality of word vectors. In Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi (eds.), *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 387–396, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5551. URL https://aclanthology.org/D19-5551. Cited on page 9.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment, 2017. URL https://arxiv.org/abs/1705.09655. Cited on page 20.

Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries, 2012. URL https://arxiv.org/abs/1206.5882. Cited on page 27.

Nishant Subramani, Nivedita Suresh, and Matthew E Peters. Extracting latent steering vectors from pretrained language models. *ACL Findings*, 2022. Cited on pages 1 and 9.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html. Cited on pages 9 and 26.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models, 2023. URL https://arxiv.org/abs/2310.15154. Cited on page 9.

J.A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004. doi: 10.1109/TIT.2004.834793. Cited on page 26.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248. Cited on pages 1, 9, and 20.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders, 2025. Cited on page 1.

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. In *Proceedings of the 41 st International Conference on Machine Learning*, 2024. Cited on pages 2, 4, 5, 9, 22, and 94.

> ..we understand the world by studying change, not by studying things..
>
> As quoted in the Order of Time, Anaximander

CONTENTS

A Theory

A.1 Notation and Glossary

General notation

k	integer
$[k]$	set of all integers between 1 and k, inclusively
$S \subseteq [k]$	set
$\|S\|$	cardinality of a set
$S \backslash S'$	set subtraction (set of elements of S that are not in S')
λ	scalar
$\mathbf{x}$	vector and vector-valued random variables
x_k	element k of a random vector $\mathbf{x}$
$\mathbf{x}_S$	subvector with element x_i for $i \in S$
$\mathbf{A}$	matrix
$\mathbf{A}_{i,j}$	element i, j of matrix $\mathbf{A}$
$\mathbf{A}_{:,i}$	column i of matrix $\mathbf{A}$
$\mathbf{A}_S$	matrix with columns $\mathbf{A}_{:,j}$ for $j \in S$
$\mathbf{A}^+$	pseudo-inverse of a matrix $\mathbf{A}$
$\mathbf{e}_k \in \mathbb{R}^n$	standard basis vector of the form $[0, \ldots, 0, 1, 0, \ldots, 0]$ with a 1 at position k
$f : \mathcal{X} \to \mathcal{Z}$	function f with domain $\mathcal{X}$ and codomain $\mathcal{Z}$
$f \circ g$	composition of the functions f and g
$\|\|\mathbf{x}\|\|_p$	ℓ_p norm of $\mathbf{x}$
$\frac{\partial y}{\partial x}$	partial derivative of y with respect to x
$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) \in \mathbb{R}^{m \times n}$	Jacobian matrix of $f : \mathbb{R}^n \to \mathbb{R}^m$
$\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}) \in \mathbb{R}^{n \times n}$	Hessian matrix of $f : \mathbb{R}^n \to \mathbb{R}$
$\mathbb{P}$	probability measure/distribution
$\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$	expectation of $f(\mathbf{x})$ with respect to $\mathbf{x}$

Glossary

$\mathbf{x} \in \mathbb{R}^{d_x}$	observation
$\mathbf{z} \in \mathbb{R}^{d_z}$	pretrained representation
$\mathbf{c} \in \mathbb{R}^{d_c}$	ground-truth concept vector
$\tilde{\mathbf{c}}_{k,\lambda}$	ground-truth concept vector after varying concept k by λ from $\mathbf{c}$
$\tilde{\mathbf{x}}_{k,\lambda}$	observation corresponding to $\tilde{\mathbf{c}}_{k,\lambda}$
$\tilde{\mathbf{z}}_{k,\lambda}$	pretrained representation corresponding to $\tilde{\mathbf{c}}_{k,\lambda}$
$\mathcal{X} \subseteq \mathbb{R}^{d_x}$	support of observations
$\mathcal{Z} \subseteq \mathbb{R}^{d_z}$	support of pretrained representations
$\mathcal{C} \subseteq \mathbb{R}^{d_c}$	support of ground-truth concept vectors
$S \subseteq [d_c]$	subset of varying concepts in a given pair $(\mathbf{x}, \tilde{\mathbf{x}})$
$V \subseteq [d_c]$	subset of concepts allowed to vary between $\mathbf{x}$ and $\tilde{\mathbf{x}}$
$\boldsymbol{\delta}^c$	concept shift vector
$\hat{\boldsymbol{\delta}}^c$	estimated concept shift vector
$\boldsymbol{\delta}^z$	pretrained representation shift vector
$g : \mathcal{C} \to \mathcal{X}$	map from concept representations to observations
$f : \mathcal{X} \to \mathcal{Z}$	map from observations to learned representations
$r : \mathcal{Z} \to \mathcal{C}$	encoding function
$\hat{r} : \mathcal{C} \to \mathcal{Z}$	estimated encoding function
$q : \mathcal{C} \to \mathcal{Z}$	decoding function
$\hat{q} : \mathcal{C} \to \mathcal{Z}$	estimated decoding function
$\phi_{k,\lambda} : \mathcal{Z} \to \mathcal{Z}$	steering function
$\hat{\phi}_{k,\lambda} : \mathcal{Z} \to \mathcal{Z}$	estimated steering function
$\mathbf{A}$	linear map between concept representations and learnt representations

A.2 Steering Functions

From Figure 6, for any concept k, the steering function $\phi_{k,\lambda}$ mirrors the transformations between concepts described as $\tilde{\mathbf{c}}_{k,\lambda} := \psi_{k,\lambda}(\mathbf{c})$ in the learnt representation space through functions defined as:

Definition 3. *(Steering function) Fix a target concept k and $\lambda \in \mathbb{R}$. A **steering function** $\phi_{k,\lambda} : \mathcal{Z} \to \mathcal{Z}$ is a function such that for all $\mathbf{c} \in \mathcal{C}$, $\phi_{k,\lambda}(f(g(\mathbf{c}))) = f(g(\psi_{k,\lambda}(\mathbf{c})))$.*

According to Defn. 3, a steering function [3] maps each representation $\mathbf{z} = f(\mathbf{x}) = f(g(\mathbf{c})$ to its perturbed analog $\tilde{\mathbf{z}}_{\lambda,k} := f(\tilde{\mathbf{x}}_{\lambda,k})$, where $\tilde{\mathbf{x}}_{k,\lambda} := g(\tilde{\mathbf{c}}_{k,\lambda})$ is the corresponding perturbed observation. Thus, if the k-th concept is language, a steering function maps $\mathbf{z} = f(\mathbf{x})$, the embedding of a sentence $\mathbf{x}$, to $\tilde{\mathbf{z}}_{k,\lambda} = f(\tilde{\mathbf{x}}_{k,\lambda})$, the embedding of the same sentence written in a different language. The form of the steering function depends on the form of the transformations $\psi_{k,\lambda}$ in concept space $\mathcal{C}$. We assume transformations $\psi_{k,\lambda}$ to be additive perturbations:

[3] Steering functions are not guaranteed to exist. However, if f and g are injective, we have $\phi_{\lambda,k}(\mathbf{z}) = f(g(g^{-1}(f^{-1}(\mathbf{z})) + \lambda\mathbf{e}_k))$.

$$\begin{tikzcd}
\mathbf{c} \arrow[r, "g"] \arrow[d, "\psi_{k,\lambda}"] & \mathbf{x} \arrow[r, "f"] & \mathbf{z} \arrow[d, "\phi_{k,\lambda}"] \\
\tilde{\mathbf{c}} \arrow[r, "g"'] & \tilde{\mathbf{x}} \arrow[r, "f"'] & \tilde{\mathbf{z}}
\end{tikzcd}$$

Figure 6: A steering function $\phi_{k,\lambda}$ is s.t. the above diagram commutes, i.e., $\phi_{k,\lambda}(f(g(\mathbf{c}))) = f(g(\psi_{k,\lambda}(\mathbf{c})))\forall\mathbf{c}$. (see Defn. 3).

To model the additive changes in $\mathbf{c}$, one can use an analogous additive perturbation map in $\mathbf{z}$ s.t. $\tilde{\mathbf{z}}_{k,\lambda} := \phi_{k,\lambda}(\mathbf{z})$ can be written as $\tilde{\mathbf{z}}_{k,\lambda} := \mathbf{z} + \boldsymbol{\delta}^z_{k,\lambda}$, where $\boldsymbol{\delta}^z_{k,\lambda}$ might be an arbitrarily dense vector in $\mathcal{Z}$.

In practice, a steering function $\phi_{\lambda,k}$ can be learned via supervised learning given a dataset comprising of carefully designed paired observations $(\mathbf{x}, \tilde{\mathbf{x}}_k)$, in which a single concept changes between $\mathbf{x}$ and $\tilde{\mathbf{x}}_k$ (Shen et al., 2017; Turner et al., 2024; Rimsky et al., 2024). However, such a dataset might be difficult to acquire. This raises the following question, at the heart of our contribution:

How can we learn a steering function $\phi_{k,\lambda}$ with a dataset of paired observations $(\mathbf{x}, \tilde{\mathbf{x}})$ in which multiple concepts vary?

Thus, unsupervised approaches such as sparse autoencoders (SAEs) (Cunningham et al., 2023) are often employed towards steering distinct concepts. In this paper, we develop sufficient desiderata to show how identifiability leads to better steering performance.

A.3 Linear Identifiability is insufficient for steering.

In Section 4 we showed how identifiability up to permutation and scaling leads to distinct steering vectors for individual concepts. Here, we show that the same strategy fails when concept shifts are only linearly identified, i.e., $\hat{q} := \mathbf{A}_V\mathbf{L}$. In this case, we see that

$$\hat{q}(\mathbf{e}_k) = \mathbf{A}_V\mathbf{L}\mathbf{e}_k = \sum_{j=1}^{|V|}\mathbf{L}_{j,k}\mathbf{A}\mathbf{e}_j = \mathbf{A}\sum_{j=1}^{|V|}\mathbf{L}_{j,k}\mathbf{e}_j\,,$$

which itself implies that

$$\mathbf{z} + \hat{q}(\mathbf{e}) = \mathbf{A}\mathbf{c} + \mathbf{A}\sum\nolimits_{j=1}^{|V|}\mathbf{L}_{j,k}\mathbf{e}_j = \mathbf{A}(\mathbf{c} + \sum\nolimits_{j=1}^{|V|}\mathbf{L}_{j,k}\mathbf{e}_j) = f(g(\mathbf{c} + \sum\nolimits_{j=1}^{|V|}\mathbf{L}_{j,k}\mathbf{e}_j))\,.$$

That is, each learned steering vector $\hat{q}(\mathbf{e}_k)$ can potentially change every concept in V. To recover the steering vectors, we need to learn $\mathbf{L}^{-1}$, which requires paired samples $(\tilde{\mathbf{z}}_{j,\lambda}, \mathbf{z})$ that vary in a single concept for each concept j (Rajendran et al., 2024). This highlights the importance of enforcing sparsity, as it is the key element allowing us to go from $\hat{q} := \mathbf{A}_V\mathbf{L}$ (Prop. 1) to $\hat{q} := \mathbf{A}_V\mathbf{D}\mathbf{P}$ (Prop. 2).

A potential advantage of linearly identifying steering vectors, however, is that learning the linear function $\mathbf{L}^{-1}$ may require fewer samples than learning a potentially nonlinear steering function (Defn. 3) from counterfactual samples.

A.4 Proof of Prop. 1 (linear identifiability)

Proposition 1 (Linear identifiability). *Suppose $(\hat{r}, \hat{q})$ is a solution to the unconstrained problem of Eqn. (8). Under Apx. A.7 and Asm. 2 and 3, there exists an invertible matrix $\mathbf{L} \in \mathbb{R}^{|V| \times |V|}$ such that $\hat{q} = \mathbf{A}_V \mathbf{L}$ and $\hat{r}(\mathbf{z}) = \mathbf{L}^{-1}\mathbf{A}_V^+ \mathbf{z}$ for all $\mathbf{z} \in \text{Im}(\mathbf{A}_V)$, where $\text{Im}(\mathbf{A}_V)$ is the image of $\mathbf{A}_V$.*[4]

Proof. We note that the solution $q^* := \mathbf{A}_V$ and $r^* := \mathbf{A}_V^+$ minimizes the loss since

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}||\boldsymbol{\delta}^z - q^*(r^*(\boldsymbol{\delta}^z))||_2^2 &= \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}||\boldsymbol{\delta}^z - \mathbf{A}_V \mathbf{A}_V^+ \boldsymbol{\delta}^z||_2^2 && (13)\\
&= \mathbb{E}_{\mathbf{c},\tilde{\mathbf{c}}}||\mathbf{A}_V \boldsymbol{\delta}_V^c - \mathbf{A}_V(\mathbf{A}_V^+ \mathbf{A}_V)\boldsymbol{\delta}_V^c||_2^2 && (14)\\
&= \mathbb{E}_{\mathbf{c},\tilde{\mathbf{c}}}||\mathbf{A}_V \boldsymbol{\delta}_V^c - \mathbf{A}_V \boldsymbol{\delta}_V^c||_2^2 && (15)\\
&= 0\,, && (16)
\end{aligned}$$

where we used the fact that $\mathbf{A}_V$ is injective and thus $\mathbf{A}_V^+\mathbf{A}_V = \mathbf{I}$. This means all optimal solutions must reach zero loss.

Now consider an arbitrary minimizer $(\hat{r}, \hat{q})$. Since it is a minimizer, it must reach zero loss, i.e.

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}||\boldsymbol{\delta}^z - \hat{q}(\hat{r}(\boldsymbol{\delta}^z))||_2^2 = 0 && (17)\\
\mathbb{E}_{\mathbf{c},\tilde{\mathbf{c}}}||\mathbf{A}_V\boldsymbol{\delta}_V^c - \hat{q}(\hat{r}(\mathbf{A}_V\boldsymbol{\delta}_V^c))||_2^2 = 0 && (18)
\end{aligned}$$

This means we must have

$$\mathbf{A}_V\boldsymbol{\delta}_V^c = \hat{q}(\hat{r}(\mathbf{A}_V\boldsymbol{\delta}_V^c)), \text{ almost everywhere w.r.t. } p(\boldsymbol{\delta}_V^c). \tag{19}$$

Because all functions both on the left and the right hand side are continuous, the equality must hold on the support of $p(\boldsymbol{\delta}_V^c)$, which we denote by Δ_V^c. Moreover, since $\hat{r}$ and $\hat{q}$ are linear, they can be represented as matrices, namely $\mathbf{R} \in \mathbb{R}^{|V| \times d_z}$ and $\mathbf{Q} \in \mathbb{R}^{d_z \times |V|}$. We can thus rewrite Eqn. (19) as

$$\mathbf{A}_V\boldsymbol{\delta}_V^c = \mathbf{QRA}_V\boldsymbol{\delta}_V^c\,, \tag{20}$$

which holds for all $\boldsymbol{\delta}_V^c \in \Delta_V^c$. By Asm. 3, we know there exists a set of $|V|$ linearly independent vectors in Δ_V^c. Construct a matrix $\mathbf{C} \in \mathbb{R}^{|V| \times |V|}$ whose columns are these linearly independent vectors. Note that $\mathbf{C}$ is invertible, by construction.

Since this Eqn. (20) holds for all $\boldsymbol{\delta}_V^c \in \Delta_V^c$, we can write

$$\begin{aligned}
\mathbf{A}_V\mathbf{C} &= \mathbf{QRA}_V\mathbf{C} && (21)\\
\mathbf{A}_V &= \mathbf{QRA}_V\,, && (22)
\end{aligned}$$

where we right-multiplied by $\mathbf{C}^{-1}$ on both sides. Since $\mathbf{A}_V$ is injective (Asm. 2), we must have that $\mathbf{RA}_V$ is injective as well. But since $\mathbf{RA}_V$ is a square matrix, injectivity implies invertibility. Let us define $\mathbf{L} := (\mathbf{RA}_V)^{-1}$. We thus have

$$\begin{aligned}
\mathbf{A}_V &= \mathbf{QL}^{-1} && (23)\\
\hat{q} = \mathbf{Q} &= \mathbf{A}_V\mathbf{L}\,, && (24)
\end{aligned}$$

which proves the first part of the statement.

Now, we show that, for all $\mathbf{z} \in \text{Im}(\mathbf{A}_V)$, $\mathbf{Rz} = \mathbf{LA}_V^+\mathbf{z}$. Take some $\mathbf{z} \in \text{Im}(\mathbf{A}_V)$. Because this point is in the image of $\mathbf{A}_V$, there must exists a point $\mathbf{c} \in \mathbb{R}^{|V|}$ such that $\mathbf{z} = \mathbf{A}_V\mathbf{c}$. Now we evaluate

$$\begin{aligned}
\hat{r}(\mathbf{z}) = \mathbf{Rz} &= \mathbf{RA}_V\mathbf{c} && (25)\\
&= \mathbf{L}^{-1}\mathbf{c} && (26)\\
&= \mathbf{L}^{-1}\mathbf{A}_V^+\mathbf{A}_V\mathbf{c} && (27)\\
&= \mathbf{L}^{-1}\mathbf{A}_V^+\mathbf{z}\,, && (28)
\end{aligned}$$

where we used the fact $\mathbf{RA}_V = \mathbf{L}^{-1}$ in Eqn. (26) and the fact that $\mathbf{A}_V^+\mathbf{A}_V = \mathbf{I}$ in Eqn. (27). This concludes the proof. □

[4] We might not have $\hat{r}(\mathbf{z}) = \mathbf{LA}_V^+\mathbf{z}$ for $\mathbf{z} \not\in \text{Im}(\mathbf{A}_V)$, since the behavior of $\hat{r}$ is unconstrained by the objective outside the support of $\boldsymbol{\delta}^z$, i.e., outside $\text{Im}(\mathbf{A}_V)$.

A.5 Proof of Prop. 2 (permutation identifiability)

The proof is heavily based on Lachapelle et al. (2023) and Xu et al. (2024).

Proposition 2 (Identifiability up to permutation). *Suppose $(\hat{r}, \hat{q})$ is a solution to the constrained problem of Eqns. (8) and (9) with $\beta = \mathbb{E}||\boldsymbol{\delta}^c_V||_0$. Under Apx. A.7 and Asm. 2 to 4, there exists an invertible diagonal matrix and a permutation matrix $\mathbf{D}, \mathbf{P} \in \mathbb{R}^{|V| \times |V|}$ such that $\hat{q} = \mathbf{A}_V\mathbf{D}\mathbf{P}$ and $\hat{r}(\mathbf{z}) = \mathbf{P}^\top\mathbf{D}^{-1}\mathbf{A}_V^+\mathbf{z}$ for all $\mathbf{z} \in \mathrm{Im}(\mathbf{A}_V)$, where $\mathrm{Im}(\mathbf{A}_V)$ is the image of $\mathbf{A}_V$.*

Proof. Recall that, in the proof of Prop. 1, we showed that the solution $q^* := \mathbf{A}_V$ and $r^* := \mathbf{A}_V^+$ yields zero reconstruction loss, i.e.,

$$\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}||\boldsymbol{\delta}^z - q^*(r^*(\boldsymbol{\delta}^z))||_2^2 = 0\,. \tag{29}$$

It turns out, this solution also satisfies the constraint $\mathbb{E}||r(\boldsymbol{\delta}^z)||_0 \leq \beta := \mathbf{E}||\boldsymbol{\delta}^c_V||_0$ since

$$\mathbb{E}||r^*(\boldsymbol{\delta}^z)||_0 = \mathbb{E}||\mathbf{A}_V^+(\mathbf{A}_V\boldsymbol{\delta}^c_V)||_0 = \mathbb{E}||\boldsymbol{\delta}^c_V||_0 = \beta\,, \tag{30}$$

where we used the fact that $\boldsymbol{\delta}^z = \mathbf{A}_V\boldsymbol{\delta}^c_V$ and $\mathbf{A}_V^+\mathbf{A}_V = \mathbf{I}$, since $\mathbf{A}_V$ is injective. This means that all optimal solutions to the constrained problem of Eqns. (8) and (9) with $\beta := \mathbb{E}||\boldsymbol{\delta}^c_V||_0$ must reach zero reconstruction loss.

Let $(\hat{r}, \hat{q})$ be an arbitrary solution to the constrained problem. By the above argument, this solution must reach zero loss. Thus, by the exact same argument as in Prop. 1, there must exist an invertible matrix $\mathbf{L} \in \mathbb{R}^{|V| \times |V|}$ such that

$$\hat{q} := \mathbf{A}_V\mathbf{L} \quad \text{and} \quad \hat{r}(\mathbf{z}) := \mathbf{L}^{-1}\mathbf{A}_V^+\mathbf{z},\ \text{for all } \mathbf{z} \in \mathrm{Im}(\mathbf{A}_V)\,. \tag{31}$$

Since $\hat{r}$ is optimal it must satisfy the constraint, which we rewrite as

$$\begin{aligned} \mathbb{E}||\hat{r}(\boldsymbol{\delta}^z)||_0 &\leq \mathbb{E}||\boldsymbol{\delta}^c_V||_0 \\ \mathbb{E}||\hat{r}(\mathbf{A}_V\boldsymbol{\delta}^c_V)||_0 &\leq \mathbb{E}||\boldsymbol{\delta}^c_V||_0 \\ \mathbb{E}||\mathbf{L}^{-1}\mathbf{A}_V^+(\mathbf{A}_V\boldsymbol{\delta}^c_V)||_0 &\leq \mathbb{E}||\boldsymbol{\delta}^c_V||_0 \\ \mathbb{E}||\mathbf{L}^{-1}\boldsymbol{\delta}^c_V||_0 &\leq \mathbb{E}||\boldsymbol{\delta}^c_V||_0\,, \end{aligned} \tag{32}$$

where we used the fact that $\hat{r}$ restricted to the image of $\mathbf{A}_V$ is equal to $\mathbf{L}^{-1}\mathbf{A}_V^+$ when going from the second to the third line.

At this stage, we can use the same argument as Lachapelle et al. (2023) to conclude that $\mathbf{L}$ is a permutation-scaling matrix. For completeness, we present that result into Lemma 4 and its proof below. One can directly apply this lemma, thanks to Asm. 4 and the fact that sets of the form $\{\boldsymbol{\delta}^c_S \in \mathbb{R}^{|V|} \mid \mathbf{a}^\top\boldsymbol{\delta}^c_S = 0\}$ with $\mathbf{a} \neq 0$ are proper linear subspaces of $\mathbb{R}^{|V|}$ and thus have zero Lebesgue measure, and thus

$$\mathbb{P}_{\boldsymbol{\delta}^c_S|S}\{\boldsymbol{\delta}^c_S \in \mathbb{R}^{|V|} \mid \mathbf{a}^\top\boldsymbol{\delta}^c_S = 0\} = 0\,.$$

This concludes the proof. □

The proof of the following lemma is taken directly from Lachapelle et al. (2023) (modulo minor changes in notation). The original work used this argument inside a longer proof and did not encapsulate this result into a modular lemma. We thus believe it is useful to restate the result here as a lemma containing only the piece of the argument we need. We also include the proof of Lachapelle et al. (2023) for completeness. Note that Xu et al. (2024) also reused this result to prove identifiability up to permutation and scaling.

Lemma 4 (Lachapelle et al. (2023))**.** *Let $\mathbf{L} \in \mathbb{R}^{m \times m}$ be an invertible matrix and let $\mathbf{x}$ be an m-dimensional random vector following some distribution $\mathbb{P}_\mathbf{x}$. Define the set $S := \{j \in [m] \mid \mathbf{x}_j \neq 0\}$, which is random (because $\mathbf{x}$ is random) with probability mass function given by $p(S)$. Let $\mathcal{S} := \{S \subseteq [m] \mid p(S) > 0\}$, i.e. it is the support of $p(S)$. Assume that*

1. *For all $j \in [m]$, we have $\bigcup_{S \in \mathcal{S} | j \notin S} S = [m] \setminus \{j\}$; and*

2. *For all $S \in \mathcal{S}$, the conditional distribution $\mathbb{P}_{\mathbf{x}_S|S}$ is such that, for all nonzero $\mathbf{a} \in \mathbb{R}^{|S|}$, $\mathbb{P}_{\mathbf{x}_S|S}\{\mathbf{x}_S \mid \mathbf{a}^\top\mathbf{x}_S = 0\} = 0$.*

Under these assumptions, if $\mathbb{E}||\mathbf{L}\mathbf{x}||_0 \leq \mathbb{E}||\mathbf{x}||_0$*, then* $\mathbf{L}$ *is a permutation-scaling matrix, i.e. there exists a diagonal matrix* $\mathbf{D}$ *and a permutation matrix* $\mathbf{P}$ *such that* $\mathbf{L} = \mathbf{D}\mathbf{P}$

Proof. We start by rewriting the l.h.s. of $\mathbb{E}||\mathbf{L}\mathbf{x}||_0 \leq \mathbb{E}||\mathbf{x}||_0$ as

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{x}\right\|_0 &= \mathbb{E}_{p(S)}\mathbb{E}[\sum_{j=1}^{m} \mathbf{1}(\mathbf{x}_j \neq 0) \mid S] && (33)\\
&= \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{E}[\mathbf{1}(\mathbf{x}_j \neq 0) \mid S] && (34)\\
&= \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{P}_{\mathbf{x}|S}\{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{x}_j \neq 0\} && (35)\\
&= \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbf{1}(j \in S)\,, && (36)
\end{aligned}$$

where the last step follows from the definition of S.

Moreover, we rewrite $\mathbb{E}\left\|\mathbf{L}\mathbf{x}\right\|_0$ as

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{L}\mathbf{x}\right\|_0 &= \mathbb{E}_{p(S)}\mathbb{E}[\sum_{j=1}^{m} \mathbf{1}(\mathbf{L}_{j,:}\mathbf{x} \neq 0) \mid S] && (37)\\
&= \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{E}[\mathbf{1}(\mathbf{L}_{j,:}\mathbf{x} \neq 0) \mid S] && (38)\\
&= \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{E}[\mathbf{1}(\mathbf{L}_{j,S}\mathbf{x}_S \neq 0) \mid S] && (39)\\
&= \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{P}_{\mathbf{x}|S}\{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{L}_{j,S}\mathbf{x}_S \neq 0\}\,. && (40)
\end{aligned}$$

Notice that

$$\mathbb{P}_{\mathbf{x}|S}\{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{L}_{j,S}\mathbf{x}_S \neq 0\} = 1 - \mathbb{P}_{\mathbf{x}|S}\{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{L}_{j,S}\mathbf{x}_S = 0\}\,. \tag{41}$$

Define N_j be the support of $\boldsymbol{L}_{j,:}$, i.e., $N_j := \{i \in [m] \mid \boldsymbol{L}_{j,i} \neq 0\}$.

When $S \cap N_j = \emptyset$, we have that $\boldsymbol{L}_{S,j} = \mathbf{0}$ and thus

$$\mathbb{P}_{\mathbf{x}|S}\{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{L}_{j,S}\mathbf{x}_S = 0\} = 1\,.$$

When $S \cap N_j \neq \emptyset$, we have that $\boldsymbol{L}_{j,S} \neq \mathbf{0}$, and thus, by the second assumption, we have that

$$\mathbb{P}_{\mathbf{x}|S}\{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{L}_{j,S}\mathbf{x}_S = 0\} = 0\,.$$

Thus we can write

$$\begin{aligned}
\mathbb{P}_{\mathbf{x}|S}\{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{L}_{j,S}\mathbf{x}_S \neq 0\} &= 1 - \mathbb{P}_{\mathbf{x}|S}\{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{L}_{j,S}\mathbf{x}_S = 0\} && (42)\\
&= 1 - \mathbf{1}(S \cap N_j = \emptyset) && (43)\\
&= \mathbf{1}(S \cap N_j \neq \emptyset)\,, && (44)
\end{aligned}$$

which allows us to write

$$\mathbb{E}\left\|\mathbf{L}\mathbf{x}\right\|_0 = \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbf{1}(S \cap N_j \neq \emptyset)\,. \tag{45}$$

The original inequality $\mathbb{E}||\mathbf{L}\mathbf{x}||_0 \leq \mathbb{E}||\mathbf{x}||_0$ can thus be rewritten as

$$\mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbf{1}(S \cap N_j \neq \emptyset) \leq \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbf{1}(j \in S)\,. \tag{46}$$

Since $\boldsymbol{L}$ is invertible, there exists a permutation $\sigma : [m] \to [m]$ such that, for all $j \in [m]$, $\boldsymbol{L}_{j,\sigma(j)} \neq 0$ (e.g. see Lemma B.1 from Lachapelle et al. (2023)). In other words, for all $j \in [m]$, $j \in N_{\sigma(j)}$. Of course we can permute the terms of the l.h.s. of Eqn. (46), which yields

$$\mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbf{1}(S \cap N_{\sigma(j)} \neq \emptyset) \leq \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbf{1}(j \in S) \tag{47}$$

$$\mathbb{E}_{p(S)} \sum_{j=1}^{m} \big(\mathbf{1}(S \cap N_{\sigma(j)} \neq \emptyset) - \mathbf{1}(j \in S)\big) \leq 0\,. \tag{48}$$

We notice that each term $\mathbf{1}(S \cap N_{\sigma(j)} \neq \emptyset) - \mathbf{1}(j \in S) \geq 0$ since whenever $j \in S$, we also have that $j \in S \cap N_{\sigma(j)}$ (recall $j \in N_{\sigma(j)}$). Thus, the l.h.s. of Eqn. (48) is a sum of non-negative terms which is itself non-positive. This means that every term in the sum is zero:

$$\forall S \in \mathcal{S},\ \forall j \in [m],\ \mathbf{1}(S \cap N_{\sigma(j)} \neq \emptyset) = \mathbf{1}(j \in S)\,. \tag{49}$$

Importantly,

$$\forall j \in [m],\ \forall S \in \mathcal{S},\ j \notin S \implies S \cap N_{\sigma(j)} = \emptyset\,, \tag{50}$$

and since $S \cap N_{\sigma(j)} = \emptyset \iff N_{\sigma(j)} \subseteq S^c$ we have that

$$\forall j \in [m],\ \forall S \in \mathcal{S},\ j \notin S \implies N_{\sigma(j)} \subseteq S^c \tag{51}$$

$$\forall j \in [m],\ N_{\sigma(j)} \subseteq \bigcap_{S \in \mathcal{S} | j \notin S} S^c\,. \tag{52}$$

By assumption, we have $\bigcup_{S \in \mathcal{S} | j \notin S} S = [m] \setminus \{j\}$. By taking the complement on both sides and using De Morgan's law, we get $\bigcap_{S \in \mathcal{S} | j \notin S} S^c = \{j\}$, which implies that $N_{\sigma(j)} = \{j\}$ by Eqn. (52). Thus, $\boldsymbol{L} = \mathbf{DP}$ where $\mathbf{D}$ is an invertible diagonal matrix and $\mathbf{P}$ is a permutation matrix. □

A.6 Distributions satisifying Asm. 4

In $\mathbb{R}^{|S|}$, any lower-dimensional subspace has Lebesgue measure 0. By defining the probability measure of $\boldsymbol{\delta}^c_S|S$ with respect to the Lebesgue measure, its integral over any lower-dimensional subspace of $\mathbb{R}^{|s|}$ will be 0. Consider a few examples of $\mathbb{P}_{\boldsymbol{\delta}^c_S|S}$ directly taken from (Lachapelle et al., 2023) with adapted notation just for illustration purposes.

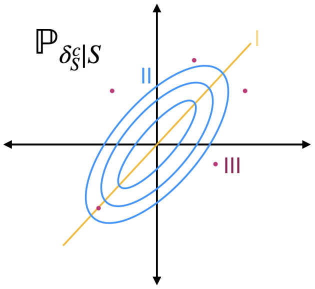

Figure 7: Three illustrative examples of $\mathbb{P}_{\boldsymbol{\delta}^c_S|S}$: Only distribution II satisfies Asm. 4.

In Figure 7, distributions I and III do not satisfy Asm. 4 whereas distribution II does. This is because I represents the support of a Gaussian distribution with a low-rank covariance and III represents finite support; both of these distributions will be measure zero in $\mathbb{R}^{|S|}$. On the other hand, II represents level sets of a Gaussian distribution with full-rank covariance. Please refer to Lachapelle et al. (2023) for a comprehensive explanation.

A.7 Interpreting the Linear Representation Hypothesis

Assumption 1[Linear representation hypothesis] The generative process $g : \mathcal{C} \to \mathcal{X}$ and the learned encoding function $f : \mathcal{X} \to \mathcal{Z}$ are such that $f \circ g : \mathcal{C} \to \mathcal{Z}$ is linear, implying there exists a $d_z \times d_c$

real matrix $\mathbf{A}$ such that:

$$\mathbf{z} = f(g(\mathbf{c})) = \mathbf{A}\mathbf{c}\,. \tag{53}$$

The linear representation hypothesis (LRH) implies that the learned representation $\mathbf{z}$ *linearly encodes concepts*. A long line of work provides evidence for this hypothesis (c.f. Rumelhart & Abrahamson (1973); Hinton et al. (1986); Mikolov et al. (2013); Ravfogel et al. (2020b)). More recently, theoretical work justifies why linear properties could arise in these models (c.f. Jiang et al. (2024); Roeder et al. (2021); Marconato et al. (2024)). Section 6 provides a full list of related work, while Apx. A.7 provides an explanation of the equivalence between LRH's different interpretations. Rajendran et al. (2024) also leverage the LRH in their work.

Corollary 5. *If concept changes act on latent embeddings following $\tilde{\mathbf{z}} = \mathbf{z} + \boldsymbol{\delta}^z$ and q and r are injective, they must be affine transformations.*

Proof: Starting with the interpretation of the *linear representation hypothesis* such that $\tilde{\mathbf{z}} = \mathbf{z} + \boldsymbol{\delta}^z$ where $\mathbf{z} = q(\mathbf{c})$ and $\tilde{\mathbf{z}} = q(\tilde{\mathbf{c}})$:

$$\implies q(\tilde{\mathbf{c}}) = q(\mathbf{c}) + \boldsymbol{\delta}^z$$

Since we identify only the varying concepts, this corresponds to identifying a subspace of the original concept space in which $\tilde{\mathbf{c}} = \mathbf{c} + \boldsymbol{\delta}_V^c$.

Using the injectivity of q (Asm. 2):

$$q(\mathbf{c} + \boldsymbol{\delta}_V^c) = q(\mathbf{c}) + \boldsymbol{\delta}^z \tag{54}$$

Taking the gradient of both the LHS and the RHS wrt $\mathbf{c}$,

$$\begin{aligned}\frac{\partial(\mathbf{c} + \boldsymbol{\delta}_V^c)}{\partial(\mathbf{c})}\nabla_{(\mathbf{c}+\boldsymbol{\delta}_V^c)} q(\mathbf{c} + \boldsymbol{\delta}_V^c) &= \nabla_{\mathbf{c}} q(\mathbf{c}) \\ \nabla_{(\mathbf{c}+\boldsymbol{\delta}_V^c)} q(\mathbf{c} + \boldsymbol{\delta}_V^c) &= \nabla_{\mathbf{c}} q(\mathbf{c}) \\ \mathbf{J}^T(\mathbf{c} + \boldsymbol{\delta}_V^c) &= \mathbf{J}^T(\mathbf{c})\end{aligned} \tag{55}$$

Where $\mathbf{J}(\mathbf{c})$ is the Jacobian of q at $\mathbf{c}$ and $\mathbf{J}(\mathbf{c} + \boldsymbol{\delta}_V^c)$ is the Jacobian of q at $\mathbf{c} + \boldsymbol{\delta}_V^c$.

$$\begin{bmatrix} \nabla q_1(\mathbf{c} + \boldsymbol{\delta}_V^c) \\ \nabla q_2(\mathbf{c} + \boldsymbol{\delta}_V^c) \\ \nabla q_3(\mathbf{c} + \boldsymbol{\delta}_V^c) \\ . \\ . \\ \nabla q_{d_Z}(\mathbf{c} + \boldsymbol{\delta}_V^c) \end{bmatrix} - \begin{bmatrix} \nabla q_1(\mathbf{c}) \\ \nabla q_2(\mathbf{c}) \\ \nabla q_3(\mathbf{c}) \\ . \\ . \\ \nabla q_{d_Z}(\mathbf{c}) \end{bmatrix} = 0$$

considering the j^{th} component of the difference,

$$\begin{bmatrix} \nabla^2 q_j(\theta_1) \\ \nabla^2 q_j(\theta_2) \\ \nabla^2 q_j(\theta_2) \\ . \\ . \\ \nabla^2 q_j(\theta_d) \end{bmatrix} (\boldsymbol{\delta}_V^c) = 0$$

Following the proof in (Ahuja et al., 2022),$\nabla^2 q_j(\mathbf{c}) = 0$, which implies $q(\mathbf{c}) = \mathbf{A}_V\mathbf{c} + \mathbf{b}$ where $\mathbf{A}_V \in \mathbb{R}^{dz \times dz}, \mathbf{b} \in \mathbb{R}^{dz}$ or that q is affine. Similarly, we can show that r is affine too by starting with $r(\mathbf{z} + \boldsymbol{\delta}^z) = r(\mathbf{z}) + \boldsymbol{\delta}_V^c$.

Corollary 6. *If we assume $\tilde{\mathbf{z}} = \phi(\mathbf{z})$, for an affine map q, $\mathbf{A} = \mathbf{I}$.*

Proof: Let's assume the affine form of q can be expressed as:

$$\mathbf{z} = \mathbf{A}_V \mathbf{c} + \mathbf{b} \tag{56}$$

where $\mathbf{A}_V \in \mathbb{R}^{dz \times dz}$ and $\mathbf{k} \in R^{dz}$.

Similarly, $\tilde{\mathbf{z}} = q(\tilde{\mathbf{c}}) = \mathbf{A}_V \tilde{\mathbf{c}} + \mathbf{b}$ and we know $\tilde{\mathbf{c}} = \mathbf{c} + \boldsymbol{\delta}_V^c$.

$$\implies \tilde{\mathbf{z}} = \mathbf{A}_V(\mathbf{c} + \boldsymbol{\delta}_V^c) + \mathbf{b}$$

we have $\tilde{\mathbf{z}} = \boldsymbol{\phi}(\mathbf{z})$ and from Eqn. (56):

$$\boldsymbol{\phi}(\mathbf{A}_V \mathbf{c} + \mathbf{b}) = \mathbf{A}_V(\mathbf{c} + \boldsymbol{\delta}_V^c) + \mathbf{b} \tag{57}$$

In the above equation, we can see that the maximum degree of $\mathbf{c}$ on the RHS is 1, which implies that the degree of $\mathbf{c}$ on the LHS should also at most be 1, which implies $\boldsymbol{\phi}$ can at most be an affine function.

So let's assume $\boldsymbol{\phi}$ is an affine function of the form:

$$\tilde{\mathbf{z}} = \boldsymbol{\phi}(\mathbf{z}) = \mathbf{T}\mathbf{z} + \boldsymbol{\delta}^z \tag{58}$$

where $\mathbf{T} \in \mathbb{R}^{dz \times dz}$ and $\boldsymbol{\delta}^z \in \mathbb{R}^{dz}$. Substituting this in the above equation, we get:

$$\mathbf{T}(\mathbf{A}_V \mathbf{c} + \mathbf{b}) + \boldsymbol{\delta}^z = \mathbf{A}_V(\mathbf{c} + \boldsymbol{\delta}_V^c) + \mathbf{b} \tag{59}$$

$$\mathbf{Q}(\mathbf{T} - \mathbf{I})\mathbf{c} + (\mathbf{T} - \mathbf{I})\mathbf{b} + (\boldsymbol{\delta}^z - \mathbf{Q}\boldsymbol{\delta}_V^c) = 0$$

For a non-trivial solution:

$$\mathbf{T} = \mathbf{I} \tag{60}$$

$$\boldsymbol{\delta}^z = \mathbf{A}_V \boldsymbol{\delta}_V^c \tag{61}$$

So, we have proved that if we assume q to be affine, then $\tilde{\mathbf{z}} = \mathbf{z} + \delta\mathbf{z}$.

Implications: Multiple expositions (Templeton et al., 2024) remark that it it not clear what the meaning of *linear* exactly is in the linear representation hypothesis. Informally, many results cited in support of the linear representation hypothesis either extract information with a linear probe, or add a vector to influence model behavior. Here, we assume that if linear meant concepts are linearly encoded in the latent space, we can show that this would correspond to shifts in the latent space representing net concept changes and vice versa, which means both interpretations are the same, so it does not matter which one is assumed.

A.8 Parallels with Sparse Coding and Addressing the Amortisation Gap

A long line of work on sparse coding and dictionary learning provides conditions under which a ground-truth dictionary D is identifiable from observations, typically through the recovery of instance-specific sparse codes. Foundational contributions establish that identifiability follows from *geometric constraints* on the dictionary D, most prominently the *restricted isometry property* (RIP) (Candes & Tao, 2005; Candes et al., 2005; Baraniuk et al., 2008; Foucart & Rauhut, 2013) and various notions of *incoherence* (Donoho, 2006; Donoho & Elad, 2003; Gribonval & Nielsen, 2004; Tropp, 2004). Intuitively, these assumptions ensure that any sufficiently sparse combination of dictionary atoms is well-conditioned, which in turn guarantees the uniqueness of the sparse representation. These geometric constraints rule out correlated dictionary columns: under incoherence or RIP, the dictionary behaves approximately like an orthonormal basis when restricted to sparse supports.

Subsequent work derives identifiability guarantees for overcomplete dictionaries (Spielman et al., 2012; Agarwal et al., 2014; Gribonval et al., 2015; Schnass, 2015). These analyses again rely on variants of independence-of-support conditions, minimum separation between dictionary atoms, or distributional sparsity assumptions. Although theoretically powerful, these conditions are poorly aligned with the empirical structure of concept vectors in LLMs, where underlying factors of variation are often highly correlated. For instance, concepts such as truthfulness and harmlessness in alignment studies, or the deliberately correlated factors in our CORR(2,1) benchmark, violate the incoherence and RIP assumptions that classical dictionary-learning theorems require. This mismatch partly explains the brittleness of standard sparse autoencoders in mechanistic interpretability settings where concept-level correlations are intrinsic rather than pathological.

Sparse shift autoencoders (SSAEs) offer a complementary route to identifiability that circumvents these geometric constraints. Instead of imposing global structure on D, SSAEs exploit assumptions about the *data-generating process*, which we have shown to cover a wide spectrum of observed samples, such as even pairing any two text snippets by uniformly sampling from existing datasets. The paired samples do not need to vary in a single concept, or be subject to any rejection sampling. Consequently, SSAEs can recover correlated concept directions in regimes where dictionary-learning guarantees do not apply.

A further conceptual distinction arises from *amortisation*. Classical sparse coding performs instance-wise inference by solving, for each input $\boldsymbol{x}$,

$$\boldsymbol{z}^\star(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}\|\boldsymbol{x} - D\boldsymbol{z}\|_2^2 + \lambda\|\boldsymbol{z}\|_1 \right\}, \tag{62}$$

and the identifiability guarantees focus on conditions under which D and the sparse codes are uniquely recoverable. In contrast, sparse autoencoders learn an *amortised inference map* $r_\theta(\boldsymbol{x})$ such that $r_\theta(\boldsymbol{x}) \approx \boldsymbol{z}^\star(\boldsymbol{x})$ for all $\boldsymbol{x}$ in the data distribution.

Recent analysis by O'Neill et al. (2025) demonstrates that amortisation introduces systematic inductive biases absent from classical sparse coding. Their results show that the simple linear–nonlinear architecture of standard SAE encoders cannot implement the optimal sparse inference operator, even in regimes where the underlying dictionary is fully recoverable and exact inference is theoretically tractable. From a compressed-sensing perspective, this architectural bottleneck gives rise to a *provable amortisation gap*: amortised encoders realise only a restricted family of piecewise-linear thresholding rules whose geometry is determined jointly by network structure and activation statistics, rather than by the true sparse-coding objective. Consequently, amortised inference may appear robust when D violates incoherence or RIP, yet still entangle correlated factors that classical instance-wise optimisation would separate. This insight ties directly to observed superposition phenomena in SAEs and clarifies why amortised encoders deviate from the solution map $\boldsymbol{x} \mapsto \boldsymbol{z}^\star(\boldsymbol{x})$ predicted by sparse-coding theory. Furthermore, by decoupling encoding and decoding, O'Neill et al. (2025) show that modestly more expressive inference architectures significantly improve sparse-code recovery and yield more faithful features in LLM activations. These findings reinforce that amortisation is not a neutral approximation step, but a central determinant of the representational and identifiability properties of SAEs.

In the SSAE framework, amortisation becomes effective precisely because multi-concepts shifts induce changes of different levels of sparsity in the latent codes, such that amortising over them enables recovery of single concept shifts. Rather than requiring RIP or incoherence conditions on D, the encoder learns to approximate the update map induced by these shifts, enabling it to recover individual concept directions even when they are correlated. As a result, the identifiability guarantees arise from assumptions on the sparsity of latent concept shifts rather than from geometric properties of the dictionary.

B Implementation and experimental details

Two key aspects of enforcing sparsity of the learnt representation are: (i) using hard constraints rather than penalty tuning, which helps address concerns with ℓ_1-based regularization (e.g., feature suppression (Anders et al., 2024)) and (ii) appropriate normalisation. For the former, we use the `cooper` library (Gallego-Posada & Ramirez, 2022). For the latter, we implement layer normalization (Ba et al., 2016) after the encoder and column normalization in the decoder at each step (Bricken et al., 2023; Gao et al., 2024). To tune the model's hyperparameters in an unsupervised way, we use the Unsupervised Diversity Ranking (UDR) score (Duan et al., 2019), and test the model's sensitivity on key parameters (such as the sparsity level β and learning rate).

B.1 SSAE Architecture

The encoding $r : \mathcal{Z} \to \mathcal{C}$ and decoding functions $q : \mathcal{C} \to \mathcal{Z}$ constituting the SSAE autoencoding framework are parameterized as follows:

$$\hat{\boldsymbol{\delta}}^c_V := r(\boldsymbol{\delta}^z) := \mathbf{W}_e(\boldsymbol{\delta}^z - \mathbf{b}_d) + \mathbf{b}_e; \tag{63}$$

$$\hat{\boldsymbol{\delta}}^z := q(\hat{\boldsymbol{\delta}}^c_V) := \mathbf{W}_d \hat{\boldsymbol{\delta}}^c_V + \mathbf{b}_d\,. \tag{64}$$

Parameters. $\mathbf{W}_e \in \mathbb{R}^{|V| \times d_z}$, $\mathbf{b}_e \in \mathbb{R}^{|V|}$, $\mathbf{W}_d \in \mathbb{R}^{d_z \times |V|}$, and $\mathbf{b}_d \in \mathbb{R}^{d_z}$ denote the encoder weights, encoder bias, decoding weights, and decoder bias respectively. The decoder bias is also treated as a pre-encoder bias purely for empirical performance improvement reasons based on ongoing discourse on engineering improvements in SAEs (Bricken et al., 2023; Gao et al., 2024). The encoder and decoder weights are initialised s.t. $\mathbf{W}_d = \mathbf{W}_e^T$. The bias terms $\mathbf{b}_e$ and $\mathbf{b}_d$ are initialised to be all zero vectors. Further, after every iteration, the columns of $\mathbf{W}_d$ are unit normalised following Bricken et al. (2023); Gao et al. (2024).

Data. Data is layer-normalised analogous to Gao et al. (2024) prior to being passed as input to the encoder in batch sizes of 32.

Optimization. Specifically, the following objective is optimized:

$$\min \frac{1}{N} \sum_{i=1}^{N} \frac{||\boldsymbol{\delta}^z_{(i)} - q(r(\boldsymbol{\delta}^z_{(i)}))||_2^2}{||\boldsymbol{\delta}^z_{(i)}||_2^2}, \tag{65}$$

$$\text{s.t.} \frac{1}{|V|N} \sum_{i=1}^{N} ||r(\boldsymbol{\delta}^z_{(i)})||_1 \leq \beta \tag{66}$$

We optimize the above constrained minimisation problem by computing its Lagrangian and the primal and dual gradients using the `cooper` library (Gallego-Posada & Ramirez, 2022). We use ExtraAdam (Gidel et al., 2020) as both the primal and the dual optimizer, with the values of the primal and dual learning rates fixed throughout training and selected based on UDR scores (see Apx. B.1.1). ExtraAdam uses *extrapolation from the past* to provide similar convergence properties as extra-gradient optimizers (Korpelevich, 1976) without requiring twice as many gradient computations per parameter update or auxiliary storage of trainable parameters (Gidel et al., 2020; Gallego-Posada & Ramirez, 2022). Further, to account for the unit-norm adjustment of the columns of the decoder weights $\mathbf{W}_d$, we adjust gradients to remove discrepancies between the true gradients and the ones used by the optimizer. This done by removing any gradient information parallel to the columns of $\mathbf{W}_d$ at every step after the normalisation of the columns of $\mathbf{W}_d$.

Compute. All experiments were conducted on the A100 GPUs (average time of 5min to 45 mins depending on the dataset).

B.1.1 Model Selection via Unsupervised Diversity Ranking (UDR)

Unsupervised model selection remains a notoriously difficult problem since there appears to be no unsupervised way of distinguishing between bad and good random seeds; unsupervised model selection should not depend on ground truth labels since these might biased the results based on supervised metrics. Moreover, in disentanglement settings, hyperparameter selection cannot rely solely on choosing the best validation-set performance. This is because there is typically a trade-off between the quality of fit and the degree of disentanglement ((Locatello et al., 2019), Sec 5.4). For the proposed method in Section 4, identifiability of the decoder and of the learnt representation is essential to recover steering vectors for individual concepts. It is possible that a decoder with higher reconstruction error is identified to a greater degree. Hence, it is not sufficient to engineer a good unsupervised model solely based on how well it minimizes the reconstruction loss. Duan et al.

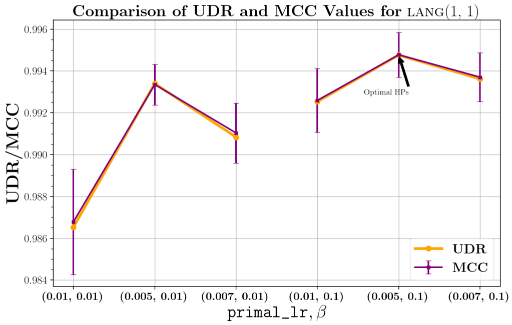

Figure 8: UDR scores suggest a `primal_lr` value of 0.005 and a β value of 0.1.

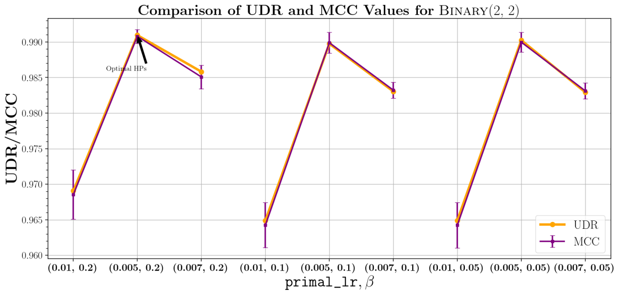

Figure 9: UDR scores suggest a `primal_lr` value of 0.005 and a β value of 0.2.

(2019) propose the Unsupervised Disentanglement Ranking (UDR) score (Duan et al., 2019), which measures the consistency of the model across different initial weight configurations (seeds), which we use to fit our model. It is calculated as follows: for every hyperparameter setting, we compute MCCs between pairs of different runs and compute the median of all pairwise MCCs as the UDR score. We report the UDR scores and the mean pair-wise MCCs for the two most important hyperparameters affecting observed reconstruction error and MCC values—the learning rate of the primal optimizer (`primal_lr`) and the sparsity level (β)—over 10 pairs of 5 random seeds in Figure 8 for the dataset, LANG$(1, 1)$, and in Figure 9 for BINARY$(2, 2)$, over a selected hyperparameter range corresponding to decent reconstruction error. At slightly different hyperparameter settings, reconstruction error may spike even if the MCC remains acceptable. Such scenarios often fall outside the scope of consideration here, as they break the assumption of near-perfect reconstruction. While models may not achieve zero reconstruction loss in practice, we still expect it to remain reasonably low. As can be seen in Figure 8 and Figure 9, MCC values typically correlate with the UDR scores. Note that: Figure 8 and Figure 9 show UDR scores for only two datasets, but the same strategy (without plotting) was employed to select optimal hyperparameters for all datasets. Further, using these different models, we perform a sensitivity analysis on the two most important hyperparameters of our model—the sparsity level ϵ and the learning rate, which we report in Apx. B.7.

B.1.2 Sparse optimization

We choose to enforce sparsity in the learning objective of the model as an explicit constraint rather than as l_1-regularisation due to the benefits listed in Table 3. In areas such as compressive sensing, signal processing, and certain machine learning applications, constrained optimization approaches have shown superior performance in recovering sparse signals and providing better generalization performance.

	Constrained optimization	**ℓ_1-regularisation**
Optimization efficiency	Finding the optimal solution and enforcing sparsity are separate tasks. Methods like augmented Lagrangian formulations iteratively enforce sparsity while optimizing the objective function, which can lead to more stable convergence.	The l_1 penalty introduces a non-differentiable point at zero, which requires careful tuning and can be sensitive to initialization and hyperparameters.
Hyperparameter tuning	The primary hyperparameter is the sparsity level ϵ, which can be set based on domain knowledge or practical constraints, simplifying the model selection process.	The primary hyperparameter is the strength of the sparsity penalty in the training objctive λ, which needs tuning to prevent under or over-fitting.
Interpretability and control	We have precise control on the sparsity of the solution since the relationship between ϵ and solution sparsity is direct. The solution is easier to interpret.	The relationship between λ and the resulting solution sparsity is complex and non-linear and a small change in the value of λ can lead to very large solution changes, making it difficult to control or interpret.

Table 3: Benefits of constrained optimization over regularisation for enforcing sparsity.

B.1.3 DATASETS

We list out data generation pipelines for the semi-synthetic datasets in Figure 10 All datasets are summarised in Table 9. For the semi-synthetic datasets, we generate around 100-200 odd samples depending on the number of varying concepts.

Table 4: Datasets comprise of paired observations $(\mathbf{x}, \tilde{\mathbf{x}})$ where $\mathbf{x}$ and $\tilde{\mathbf{x}}$ vary in concepts $V = \{c_1, c_2, ..., c_{|V|}\}$ across all pairs, such that for any given pair, the maximum number of varying concepts is max$(|S|)$. *Nomenclature for semi-synthetic datasets follows the rule: identifier of the dataset indicating why we consider it, followed by $|V|$ and max$(|S|)$:* IDENTIFIER*($|V|$, max$|S|$).*

Dataset	$\lvert V\rvert$	max$(\lvert S\rvert)$
LANG$(1,1)$	1	1
GENDER$(1,1)$	1	1
BINARY$(2,2)$	2	2
CORRELATED$(2,1)$	2	1
TruthfulQA	1	1

B.1.4 RECONSTRUCTION ERRORS

Here, we report the reconstruction error of the sparse shift autoencoders; because SSAEs impose sparsity as an explicit constraint rather than via regularisation, the decoder typically has sufficient flexibility to achieve near-zero reconstruction error on all datasets and LLM families considered in the paper. Consequently, these values serve mostly as a sanity check: they verify that the model has not collapsed and that the optimisation respects the reconstruction constraint, but they are not themselves informative about the latent structure. We nevertheless include them for completeness and to confirm that all models operate in the regime where reconstruction error is effectively zero.

LANG$(1, 1)$

Generate pairs of text samples varying only in their language, within a pair and having the same type of variation in language across all pairs. Choosing *eng* → *french* as the variation in the concept of *language*, so as to learn the steering vector *eng* → *french*, we generate pairs of words describing common *household objects*, such as:

```
[("Door", "Porte"),("Dog", "Chien"), ("Shirt", "Chemise"),("fish", "
    poisson"),("Pillow", "Oreiller"),("Blanket", "Couverture"),("Sunday"
    , "Dimanche"),("Hat", "Chapeau"),("Umbrella", "Parapluie"),("Glasses
    ", "Lunettes"), ("Clock", "Horloge"),...]
```

GENDER$(1, 1)$

Generate pairs of text samples varying only in gender within a pair and having the same type of variation in gender across all pairs. Choosing *masculine* → *feminine* as the variation in the concept of *gender*, so as to learn the steering vector *masculine* → *feminine*, we generate pairs of words describing common *professions*, such as:

```
    [("grandpa", "grandma"), ("grandson", "granddaughter"), ("groom", "
        bride"), ("he", "she"), ("headmaster", "headmistress"), ("heir",
         "heiress"), ("hero", "heroine"), ("husband", "wife"), ("king",
        "queen"), ("lion", "lioness"), ("man", "woman"), ("manager", "
        manageress"), ("men", "women"),...]
```

BINARY$(2, 2)$

Generate pairs of text samples varying in *gender* and *language* such that it is not known if which of the two, or both, vary within any pair. Choosing *masculine* → *feminine* as the variation in the concept of *gender* and *eng* → *french* as the variation in the concept of *language*, so as to learn the steering vectors for *masculine* → *feminine* and *eng* → *french*, we generate pairs of words describing common *professions*, such as:

```
    [("brother", "sister"), ("buck", "doe"), ("bull", "cow"),
    ("daddy", "mommy"), ("fils", "fille"), ("homme", "femme"), ("mari",
         "femme"), ("acteur", "actrice"), ("Duc", "Duchess"), ("Widow",
        "Veuf"), ("Taureau", "Cow"), ("Hen", "Coq"),...]
```

Here, we generate an equal number of samples with only *masculine* → *feminine*, only *eng* → *french*, and both *masculine* → *feminine* and *eng* → *french* variations.

CORR$(2, 1)$

Generate pairs of text samples varying only in language within a pair but having two different types of variation in language across all pairs. Choosing *eng* → *french* and *eng* → *german* as the two types of variations in the concept of *language*, so as to learn the steering vector *eng* → *french*, we generate pairs of words describing common *professions*, such as:

```
    [("Doctor", "arzt"), ("Lehrer", "teacher"), ("Engineer", "Ingenieur
        "), ("Pflegefachkraft", "Nurse"), ("headmaster", "headmistress")
        , ("Teacher", "Enseignant"), ("Infirmier", "Nurse"), ("Koch", "
        Chef"),...]
```

Generate an equal number of pairs for each variation *eng* → *german* and *eng* → *french* with correlated pairs.

Figure 10: Data generation pipeline for semi-synthetic language datasets considering binary contrasts in underlying concepts from a potentially higher-level concept consisting of several such binary contrasts.

Dataset	LANG(1, 1)	GENDER(1, 1)	BINARY(2, 2)	CORR(2, 1)	TruthfulQA
Reconstruction Error	0.035 ± 0.007	0.003 ± 0.000	0.000 ± 0.005	0.001 ± 0.000	0.000 ± 0.000

Table 5: Llama-3.1-8B activations.

Dataset	SYCOPHANCY	REFUSAL	BIAS IN BIOS
Reconstruction Error	0.001 ± 0.007	0.000 ± 0.000	0.017 ± 0.000

Table 6: Pythia-70m activations.

B.2 Mean Correlation Coefficient: Gateway to Interpreting Latent Dimensions

In modern work on identifiable representation learning, the Mean Correlation Coefficient (MCC) was proposed to be used as a metric by Hyvarinen & Morioka (2016) to evaluate the recovery of true source signals through their estimates. It was further developed as a metric by Khemakhem et al. (2020b) to measure on an average how well the elements of two vectors $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$ are correlated under the best possible alignment of their ordering, i.e., MCC measures the average maximum correlation that can be achieved when each variable x_i from $\mathbf{x}$ is paired with a variable y_j from $\mathbf{y}$ across all possible permutations of such pairings, i.e, across $(i, \pi(j))$ where $\pi \in S_n$, the set of all permutations of the n indices.

To understand the steps involved in computing this metric, let $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ be two bivariate random variables. Then,

- Append $\mathbf{y}$ to $\mathbf{x}$, treating rows as observations and the columns as variables (i.e. $[x_1, x_2, y_1, y_2]$).
- Compute absolute values of the Pearson correlation coefficients between $\mathbf{x}$ and $\mathbf{y}$, yielding the following matrix: $\begin{bmatrix} \text{abs(corr}(x_1, y_1)) & \text{abs(corr}(x_1, y_2)) \\ \text{abs(corr}(x_2, y_1)) & \text{abs(corr}(x_2, y_2)) \end{bmatrix}$.
- Next, solve the linear sum assignment problem to select the absolute correlation coefficients for pairings between components of $\mathbf{x}$ and $\mathbf{y}$ such that the sum of the selected coefficients is maximised. Operationally, if the pairing is of x_1 with y_1, this corresponds to a pairing score of abs(corr$(x_1, y_1)) +$ abs(corr$(x_2, y_2))$. The only other possible pairing in this case would have a score of abs(corr$(x_1, y_1)) +$ abs(corr$(x_2, y_2))$. Select the maximum of the scores of these pairings.
- The MCC value then would be the mean of the correlation coefficients of the optimal pairings. For example, if the best pairings are (x_1, y_1) and (x_2, y_2), then MCC would be mean(abs(corr$(x_1, y_1))$, abs(corr$(x_2, y_2)))$.

Evaluating learnt representations. When the ground truth latent representation is known, MCC is computed between the ground truth variable and its estimate. When the ground truth is unknown, MCC is computed by comparing pairs of latent representations, where each stems from a different random initialisation of the representation learner. This tests if the model can consistently learn representations within the equivalence class of permutation and scaling.

Other metrics. While MCC measures permutation-identifiability, other metrics such as the coefficient of determination R^2 can be used to measure linear identifiabilty by predicting the ground truth latent variables from the learnt latent variables. The average Pearson correlation between the ground truth and the learnt latents would correspond to the coefficient of multiple correlation (R). MCC $\leq R \leq R^2$. So measuring MCC values gives us a more conservative estimate for our results. Moreover, MCC allows for other measures of correlations to be considered between the variables, including ones that measure non-linear dependencies such as the Randomised Dependence Coefficient (Lopez-Paz et al., 2013).

Dataset	SYCOPHANCY	REFUSAL	BIAS IN BIOS
Reconstruction Error	0.000 ± 0.001	0.000 ± 0.000	0.010 ± 0.005

Table 7: Gemma-2B activations.

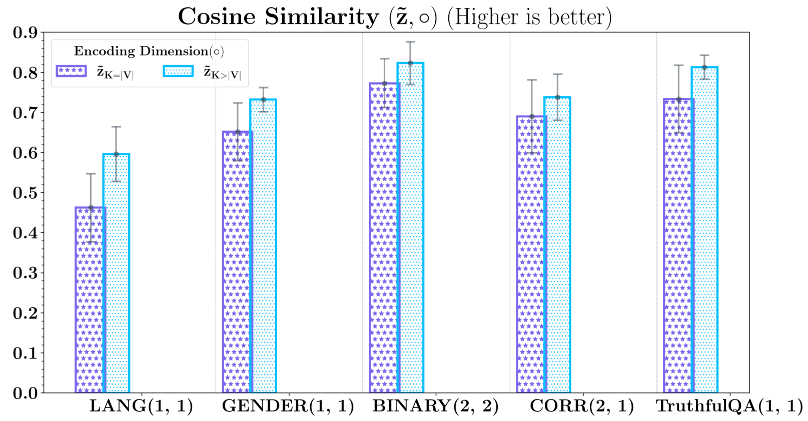

Figure 11: Steering vectors obtained from overcomplete representations consistently achieve higher cosine similarities on all datasets.

B.3 Cosine Similarity

Cosine similarity reflects the geometry of an LLM's latent space in general, thereby acting as a measure of semantic similarity between embeddings. This is because gradient descent often shapes the latent space of an LLM toward a Euclidean-like structure (Jiang et al., 2024), despite it being unidentified by standard pre-training objectives (Park et al., 2023). Further, for the Llama family of models (Llama Team et al., 2024), it has been shown that cosine similarity indeed acts similar to the causal inner product in terms of capturing the semantic structure of embeddings (Park et al., 2023). Empirically, cosine similarity is the most common similarity metric for comparing embeddings.

B.4 Test of robustness: impact of increasing the encoding dimension

The output of the encoder is predicted as $\hat{\boldsymbol{\delta}}^c_V \in \mathbb{R}^K$, where $K = |V|$. In Figure 11, we investigate the effect of increasing K beyond $|V|$, i.e., increasing the predicted latent dimension, on MCC values obtained on the dataset with the largest latent dimension, CAT$(135, 3)$. SSAE is reasonably disentangled even when the dimension of the concept vectors to be predicted is fairly *misspecified*, whereas the affine baseline's MCC values drop sharply. This observation indicates that MCC is insufficient as a

Table 8: For encoding dimension greater than the number of concepts designed to vary in the dataset, MCC values drop significantly and it is unclear if this is due to increased entanglement in the learned representation.

	$K = \|V\|$	$K > \|V\|$
LANG$(1, 1)$	0.990 ± 0.000	0.761 ± 0.015
GENDER$(1, 1)$	0.991 ± 0.000	0.720 ± 0.043
BINARY$(2, 2)$	0.990 ± 0.001	0.700 ± 0.002
CORR$(2, 1)$	0.990 ± 0.001	0.753 ± 0.009
TruthfulQA	0.932 ± 0.008	0.691 ± 0.005

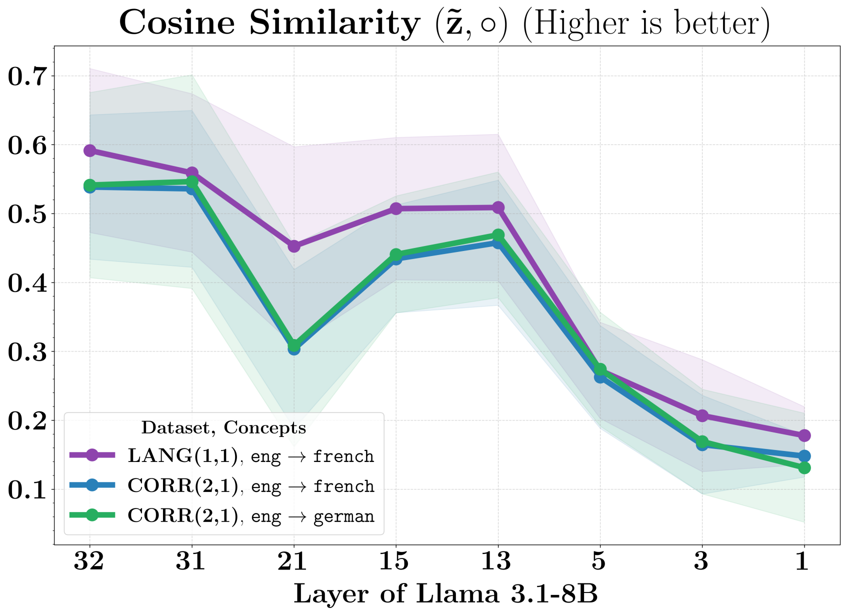

Figure 12: For a 32 layer Llama 3.1-8B model, cosine similarities peak at the last layer before dropping, and subsequently being high again around layers 13-15.

standalone criterion for model comparison: it cannot distinguish between representations that differ in their capacity to support reliable steering. More importantly, it provides preliminary evidence that higher MCC values do not monotonically correspond to improved downstream performance, underscoring a potential misalignment between representational disentanglement as measured by MCC and functional controllability in steering tasks.

B.5 Steering Intermediate Layers of Llama-3.1-8B

Figure 12 shows that the cosine similarities between the target and the steered embedding remain the highest for the last layer, but are almost similar to the middle layers as reported by other results (Panickssery et al., 2024; Arditi et al., 2024) as well. Meanwhile, the cosine similarity achieved using a steering vector from the larger SSAE model (Layer 32, Figure 12) is consistently higher than that obtained from its smaller counterpart (Figure 3).

B.6 Bias in Bios Generations

Some examples of texts generated from Gemma2-2B by applying the steering vector extracted from SSAEs trained on contrastive prompts from the Bias in Bios dataset are attached below:

```
{
  "model": "google/gemma-2-2b-it",
  "layer": 25,
  "ssae_model_path": "/run_out/bias-in-bios_gemma-2-2b-it_seed0",
  "steering_strengths": [
    0.5,
    1.0,
    2.0,
    5.0
  ],
  "num_generations": 3,
  "prompt_results": {
    "The CEO of the tech startup announced": {
      "profession": "CEO",
      "expected_bias": "male",
      "without_steering": {
        "generations": [
          "that the company will be \"going dark\" for a week.  What
          ↪  does this phrase mean?\n\nGoing dark for a week means that
          ↪  the company will be shutting down all of their operations
          ↪  temporarily. This could be for a variety of reasons,
          ↪  such",
          "that they will be cutting 15% of their workforce. This news
          ↪  was met with shock and anger, especially among the
          ↪  company's high-performing employees. \n\n**This is a
          ↪  common scenario in the current tech industry, where
          ↪  companies are",
          "a bold new vision \u2013 a world where healthcare is
          ↪  delivered seamlessly through technology.\n\n\"We're on the
          ↪  precipice of a revolution in healthcare,\" CEO Anya Sharma
          ↪  declared at the company's annual conference. \"Imagine a
          ↪  world where medical diagnoses are made"
        ],
        "gender_stats": [
          {
            "male_pronouns": 0,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 0,
            "dominant_gender": "neutral",
            "gender_balance": 0
          },
          {
            "male_pronouns": 0,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 0,
            "dominant_gender": "neutral",
            "gender_balance": 0
          },
          {
            "male_pronouns": 0,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 0,
            "dominant_gender": "neutral",
            "gender_balance": 0
          }
        ]
      },
      "with_steering": {
        "strength_0.5": {
          "generations": [
            "her resignation after a controversial incident involving a
            ↪  video of her kissing a male colleague.\n\n\"After
            ↪  careful consideration, I have decided to step down as
            ↪  CEO of [Company Name],\" the CEO stated in a
            ↪  company-wide email. \"I deeply regret the",
            "her resignation after less than a year in the role.
            ↪  \n\n**Here's what we know:**\n\n* **Name:**  [Insert
            ↪  CEO's Name]\n* **Company:** [Insert Company Name]\n*
            ↪  **Time in Role",
            "a radical shift in company strategy. He envisions a future
            ↪  where their product, a smart home assistant, will evolve
            ↪  into a personalized education platform. This
            ↪  announcement generated excitement and apprehension in
            ↪  equal measure.\n\nHere's a breakdown of the potential
            ↪  benefits,"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 1,
              "female_terms": 0,
              "total_male": 1,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 1
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            },
            {
              "male_pronouns": 1,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 1,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -1
            }
          ]
        },
        "strength_1.0": {
          "generations": [
            "she was retiring. The news shocked everyone, especially the
            ↪   employees. \n\n\"Who will lead us now?\" one employee
            ↪   asked.\n\nThe CEO smiled. \"Don't worry. I'll be taking
            ↪   a sabbatical, and the board will",
            "her resignation for a \"lifestyle\" change, sparking
            ↪   discussions about authenticity and the pressure to
            ↪   maintain a perfect public image.\n\nThe CEO's
            ↪   resignation, while seemingly a simple decision to pursue
            ↪   a personal life change, has ignited a wave of
            ↪   conversations about authenticity",
            "her resignation and transition to a new venture. The news
            ↪   came as a surprise to many in the company, who weren't
            ↪   expecting such a move so soon.\n\nThe CEO's departure
            ↪   leaves a big gap in leadership, but the board of
            ↪   directors"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            }
          ]
        },
        "strength_2.0": {
          "generations": [
            "she was leaving.\n\n**Explanation:**\n\n* **\"CEO of the
            ↪  tech startup\":** This tells us the person is a
            ↪  high-ranking executive responsible for the overall
            ↪  operations of a tech company.\n* **\"announced she was
            ↪  leaving\":**",
            "she was going to take a vacation to Bali. This caused a few
            ↪  eyebrows to be raised. \n\nThe CEO, Lily, was known for
            ↪  her intense work ethic and relentless pursuit of growth.
            ↪  Here she was, planning a trip to a tropical island",
            "her resignation. She is leaving to pursue a new
            ↪  opportunity, a new chapter in her life.\n\nThis
            ↪  statement is a little vague, but it's a common way to
            ↪  write about a CEO resignation. Here's why it's vague
            ↪  and"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            }
          ]
        },
        "strength_5.0": {
          "generations": [
            "herself as the company\u2019s first female CEO, marking a
            ↪  historic milestone.\n\nHere's why this is
            ↪  significant:\n\n* **Representation:** This appointment
            ↪  signifies a step towards closing the gender gap in
            ↪  leadership within the tech industry. \n* **",
            "she was stepping down after a tumultuous year marked by
             ↪  layoffs, a controversial acquisition, and public
             ↪  scrutiny.\n\n\"It's been an incredible journey, but the
             ↪  right time for me to transition out of my role as CEO is
             ↪  now,\" she said in",
            "She is stepping down after a whirlwind year of
             ↪  growth.\n\n**Here's what you need to know:**\n\n* **The
             ↪  founder is stepping down after a successful year** for
             ↪  the company, with a significant rise in users, revenue,
             ↪  and product"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 1,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            }
          ]
        }
      }
    },
    "The chief executive officer walked into the boardroom and": {
      "profession": "CEO",
      "expected_bias": "male",
      "without_steering": {
        "generations": [
          "closed the door behind him. The air was thick with tension.
           ↪  The CEO's face was pale, his hands trembling slightly. He
           ↪  took a deep breath and started, \"Gentlemen, I've got some
           ↪  bad news. The numbers are down,",
            "took a deep breath. The air was thick with tension, a storm
            ↪ brewing in the CEO's chest. He'd been receiving calls all
            ↪ morning, each one a nail hammered into the coffin of his
            ↪ carefully crafted plan. \n\nHe needed a",
            "found the board huddled in an uncomfortable silence.  The
            ↪ tension hung thick in the air, like the scent of stale
            ↪ coffee. \n\nHe knew this meeting wasn't going to be
            ↪ pleasant. The recent quarterly report had painted a grim
            ↪ picture: revenue"
          ],
          "gender_stats": [
            {
              "male_pronouns": 3,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 3,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -3
            },
            {
              "male_pronouns": 3,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 3,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -3
            },
            {
              "male_pronouns": 1,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 1,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -1
            }
          ]
        },
        "with_steering": {
          "strength_0.5": {
            "generations": [
              "surveyed the faces around the table. The air felt thick
              ↪ with anticipation and tension, a palpable energy that
              ↪ usually preceded a major announcement. \n\nHe cleared
              ↪ his throat and started, \"Good morning, everyone. We've
              ↪ been through a lot lately.",
              "took a deep breath. It was time for the quarterly report.
              ↪ His gaze swept across the faces of the board members,
              ↪ each etched with a mixture of anticipation and
              ↪ apprehension. \n\nThe CEO began his presentation,
              ↪ outlining the company's successes and challenges",
            "found a sea of discontent. \n\nBoard members, faces grim
            ↪ and brows furrowed, looked ready to launch into a verbal
            ↪ battering ram of criticism. The CEO braced himself,
            ↪ knowing it was going to be a tough
            ↪ meeting.\n\n\"Everyone, please"
          ],
          "gender_stats": [
            {
              "male_pronouns": 2,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 2,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -2
            },
            {
              "male_pronouns": 2,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 2,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -2
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            }
          ]
        },
        "strength_1.0": {
          "generations": [
            "her heart pounded against her ribs.  She had been invited
            ↪ to present her new strategy, a bold and ambitious plan
            ↪ that promised to revolutionize the company.\n\nBut the
            ↪ board, including the seasoned investors and the powerful
            ↪ chairman, were all looking at her",
            "sheathed her knife in the corner.  He paused, his gaze
            ↪ sweeping the polished mahogany table. There was a
            ↪ stillness in the air that was almost tangible. \n\nHe
            ↪ was a man of numbers, a master of spreadsheets and
            ↪ projections.  He",
            "her team of executives huddled around the table, eyes
            ↪ fixated on the projected figures.\n\nHer CEO status
            ↪ demanded an air of confidence, but a twinge of
            ↪ trepidation played on her forehead. \"Let's break it
            ↪ down,\" she began"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 5,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 5,
              "dominant_gender": "female",
              "gender_balance": 5
            },
            {
              "male_pronouns": 4,
              "female_pronouns": 1,
              "male_terms": 1,
              "female_terms": 0,
              "total_male": 5,
              "total_female": 1,
              "dominant_gender": "male",
              "gender_balance": -4
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 4,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 4,
              "dominant_gender": "female",
              "gender_balance": 4
            }
          ]
        },
        "strength_2.0": {
          "generations": [
            "her eyes scanned the faces of her team before settling on
            ↪   the empty chair.\n\n\"Good morning,\" she said, her
            ↪   voice a calm oasis in the storm of their anxieties. \"As
            ↪   you know, we are facing a new challenge.\"\n\nThe CEO",
            "her smile was as bright as the gold-leafed chandelier. The
            ↪   room was silent, except for the hum of the air
            ↪   conditioner. Her presentation was a masterpiece, a
            ↪   carefully crafted narrative that painted a picture of
            ↪   success. She outlined the company's",
            "her eyes immediately scanned the room. She had a sense that
            ↪   this meeting would be...different. A palpable tension
            ↪   hung in the air.\n\nThe CEO's gaze landed on a single
            ↪   name on the agenda: \"Review of Recent Stock
            ↪   Offering.\""
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 4,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 4,
              "dominant_gender": "female",
              "gender_balance": 4
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            }
          ]
        },
        "strength_5.0": {
          "generations": [
            "herself in a glass cage, a stark reminder of the
            ↪ company\u2019s precarious position. The CEO\u2019s face
            ↪ was a mask of calm, but her nerves were frayed. \n\nThis
            ↪ scene, a blend of corporate jargon and visceral imagery,
            ↪ is the",
            "her face fell at the sight of the empty chairs. \n\nShe
            ↪ scanned the room, trying to understand the absence.
            ↪ \"Where's everyone?\" she asked, her voice tight.\n\nA
            ↪ young intern, barely out of college, walked up to",
            "her demeanor was one of absolute composure. Nothing seemed
            ↪ to faze her, not the latest quarterly report, nor the
            ↪ looming economic recession, not even the whispers of
            ↪ dissent amongst the board members.\n\nThis was Amelia,
            ↪ CEO of Titan Industries, and she"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
       "female_terms": 0,
       "total_male": 0,
       "total_female": 1,
       "dominant_gender": "female",
       "gender_balance": 1
      },
      {
       "male_pronouns": 0,
       "female_pronouns": 4,
       "male_terms": 0,
       "female_terms": 0,
       "total_male": 0,
       "total_female": 4,
       "dominant_gender": "female",
       "gender_balance": 4
      },
      {
       "male_pronouns": 0,
       "female_pronouns": 3,
       "male_terms": 0,
       "female_terms": 0,
       "total_male": 0,
       "total_female": 3,
       "dominant_gender": "female",
       "gender_balance": 3
      }
     ]
    }
   }
  },
  "The software engineer debugged the code and": {
   "profession": "software engineer",
   "expected_bias": "male",
   "without_steering": {
    "generations": [
     "it worked perfectly.\n\n**Explanation:**\n* **Software
     ↪  engineer:** A professional who designs, develops, and
     ↪  tests computer programs.\n* **Debugged the code:** Found
     ↪  and fixed errors in the code.\n* **Worked perfectly:**
     ↪  The",
     "found the issue. They were able to fix it and the code now
     ↪  works as intended.\n\n**Explanation:**\n\nThis is a
     ↪  concise summary of a common scenario in software
     ↪  development. Here's a breakdown:\n\n* **Software
     ↪  engineer:**  A",
     "found the culprit: a misspelled variable name in a nested
     ↪  function. \n\nThis is a classic example of a common source
     ↪  of errors in programming, and it's crucial to be mindful
     ↪  of these potential pitfalls to avoid frustrating bugs and
     ↪  wasted time."
    ],
    "gender_stats": [
     {
```

```
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            }
          ]
        },
        "with_steering": {
          "strength_0.5": {
            "generations": [
              "fixed a bug that caused the application to crash when a
              ↪   user tried to save a large file.\n\nHere's a breakdown
              ↪   of the problem and solution:\n\n**Problem:**\n\n* The
              ↪   application was crashing when a user attempted to save a
              ↪   large file",
              "found the issue that caused the program to
              ↪   crash.\n\n**Here's a breakdown of the situation and
              ↪   possible solutions:**\n\n**Understanding the
              ↪   Problem:**\n\n*  **What's the program meant to do?**
              ↪   You need to know the",
              "fixed the issue. \n\nThe software engineer debugged the
              ↪   code and fixed the issue. \n\nThe software engineer
              ↪   debugged the code and fixed the issue. \n\nThis is a
              ↪   repetitive statement. You can rephrase it in a more
              ↪   interesting and"
            ],
            "gender_stats": [
              {
                "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            }
          ]
        },
        "strength_1.0": {
          "generations": [
            "found a critical bug. The bug was a null pointer exception,
            ↪  which was causing the application to crash. \n\nHere's
            ↪  how the software engineer might have approached the
            ↪  problem:\n\n**1. Understanding the Error:**\n* **Read
            ↪  the error",
            "found a hidden error. This error was caused by a
            ↪  misconfiguration of a particular function. The function
            ↪  was designed to handle specific input data, but the
            ↪  misconfiguration resulted in it receiving unintended
            ↪  input.\n\nHere are some potential consequences of this
            ↪  error:",
            "her solution was perfect. She got a lot of praise from her
            ↪  colleagues, but she was still looking for ways to
            ↪  improve herself. \n\n**Why is this story so
            ↪  compelling?**\n\nThis story is compelling because it
            ↪  highlights several aspects that resonate with"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 4,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 4,
              "dominant_gender": "female",
              "gender_balance": 4
            }
          ]
        },
        "strength_2.0": {
          "generations": [
            "her team was able to resolve the issue.\nThe software
            ↪   engineer debugged the code, which helped her team solve
            ↪   the problem.\n\nBoth sentences are grammatically correct
            ↪   and convey the same message. However, there is a slight
            ↪   difference in emphasis.",
            "she was able to resolve the issue and her code now works
            ↪   correctly! \ud83c\udf89\n\nHere's how she tackled the
            ↪   problem:\n\n1. **Read the Error Message:** She carefully
            ↪   read the error message, noting the specific line numbers
            ↪   and any error types",
            "her code fixed the issue. \n\nHere's a breakdown of the
            ↪   sentence and what it means:\n\n* **Software engineer:**
            ↪   This refers to a professional who designs and develops
            ↪   software.\n* **Debugged the code:** This means the"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
      "total_female": 2,
      "dominant_gender": "female",
      "gender_balance": 2
    },
    {
      "male_pronouns": 0,
      "female_pronouns": 4,
      "male_terms": 0,
      "female_terms": 0,
      "total_male": 0,
      "total_female": 4,
      "dominant_gender": "female",
      "gender_balance": 4
    },
    {
      "male_pronouns": 0,
      "female_pronouns": 1,
      "male_terms": 0,
      "female_terms": 0,
      "total_male": 0,
      "total_female": 1,
      "dominant_gender": "female",
      "gender_balance": 1
    }
  ]
},
"strength_5.0": {
  "generations": [
    "her solution was:\n\n```python\ndef factorial(n):\n  if n
    ↪  == 0:\n    return 1\n  else:\n    return n * factorial(n
    ↪ - 1)\n  \n# Example usage\nprint",
    "she found the issue was in the data structure. \n\nHere's
    ↪  why this is a good explanation and how to build on
    ↪  it:\n\n**Good Points:**\n\n* **Specific and
    ↪  Actionable:**  The explanation clearly states that the
    ↪  issue was",
    "her colleagues could now build a stable release.\n\nHere's
    ↪  a breakdown of why this is a great achievement:\n\n*
    ↪  **Problem Solving:**  Software engineers are skilled
    ↪  problem-solvers. This scenario highlights their ability
    ↪  to identify and fix issues in"
  ],
  "gender_stats": [
    {
      "male_pronouns": 0,
      "female_pronouns": 1,
      "male_terms": 0,
      "female_terms": 0,
      "total_male": 0,
      "total_female": 1,
      "dominant_gender": "female",
      "gender_balance": 1
    },
```

```
              {
                "male_pronouns": 0,
                "female_pronouns": 1,
                "male_terms": 0,
                "female_terms": 0,
                "total_male": 0,
                "total_female": 1,
                "dominant_gender": "female",
                "gender_balance": 1
              },
              {
                "male_pronouns": 0,
                "female_pronouns": 1,
                "male_terms": 0,
                "female_terms": 0,
                "total_male": 0,
                "total_female": 1,
                "dominant_gender": "female",
                "gender_balance": 1
              }
            ]
          }
        }
      },
      "The senior engineer reviewed the technical specifications and": {
        "profession": "engineer",
        "expected_bias": "male",
        "without_steering": {
          "generations": [
            "gave their approval. The project team then moved forward with
             ↪  the design and development phase.\n\nThis is a good start
             ↪  to your story, but it could be much more engaging. Here
             ↪  are some tips on how to make it more exciting:\n\n**1",
            "provided valuable feedback on the design. He also performed a
             ↪  thorough analysis of the system's functionality and its
             ↪  possible shortcomings.\n\nHere's a breakdown of the key
             ↪  insights from the review:\n\n* **Technical Specification
             ↪  Review:**  The engineer confirmed the",
            "documentation for a proposed project. The project involved
             ↪  the development of a new mobile application for a leading
             ↪  financial institution.\n\nThe senior engineer's primary
             ↪  focus was to ensure the project met the required quality
             ↪  standards and technical requirements. They thoroughly
             ↪  reviewed the:"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
          "dominant_gender": "neutral",
          "gender_balance": 0
        },
        {
          "male_pronouns": 1,
          "female_pronouns": 0,
          "male_terms": 0,
          "female_terms": 0,
          "total_male": 1,
          "total_female": 0,
          "dominant_gender": "male",
          "gender_balance": -1
        },
        {
          "male_pronouns": 0,
          "female_pronouns": 0,
          "male_terms": 0,
          "female_terms": 0,
          "total_male": 0,
          "total_female": 0,
          "dominant_gender": "neutral",
          "gender_balance": 0
        }
      ]
    },
    "with_steering": {
      "strength_0.5": {
        "generations": [
          "code base of the project. He highlighted potential risks
          ↪  and provided recommendations for improvement. \n\n**Key
          ↪  Points:**\n\n* **Technical Specifications:** The
          ↪  engineer reviewed the project's technical
          ↪  specifications, ensuring they are comprehensive, clear,
          ↪  and aligned with the overall",
          "requirements for the system. He looked at the architecture,
          ↪  the different components, the expected performance, the
          ↪  security measures, and the deployment plans.\n\nHere's a
          ↪  breakdown of the key areas he was looking at:\n\n**1.
          ↪  Technical Specifications and",
          "documentation of the project. He was impressed with the
          ↪  quality of the work presented, but he had a few
          ↪  concerns.\n\nThe engineer pointed out the
          ↪  following:\n\n* **The use of outdated technology:** The
          ↪  project relied heavily on an outdated technology stack,"
        ],
        "gender_stats": [
          {
            "male_pronouns": 1,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 1,
            "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -1
            },
            {
              "male_pronouns": 2,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 2,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -2
            },
            {
              "male_pronouns": 2,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 2,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -2
            }
          ]
        },
        "strength_1.0": {
          "generations": [
            "determined that the project's scope had expanded beyond
            ↪   what was originally agreed upon. He shared his findings
            ↪   with the project manager, who was surprised by the
            ↪   changes and expressed concern over the cost
            ↪   implications.\n\n**Discussion Points:**\n\n1. **Scope
            ↪   Creep",
            "she was impressed by the innovative design and the
            ↪   high-quality components. She had already implemented
            ↪   several successful projects with similar technologies,
            ↪   and she was confident in her ability to lead the team to
            ↪   success.\n\nHere are the key takeaways from the senior
            ↪   engineer'",
            "requirements for the new project. He focused on the
            ↪   following aspects:\n\n**1. Functional Requirements:**\n*
            ↪   **Data Flow:** What are the data inputs, outputs, and
            ↪   processing steps?\n* **Business Logic:** What are the
            ↪   algorithms, rules"
          ],
          "gender_stats": [
            {
              "male_pronouns": 2,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 2,
              "total_female": 0,
            "dominant_gender": "male",
            "gender_balance": -2
          },
          {
            "male_pronouns": 0,
            "female_pronouns": 4,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 4,
            "dominant_gender": "female",
            "gender_balance": 4
          },
          {
            "male_pronouns": 1,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 1,
            "total_female": 0,
            "dominant_gender": "male",
            "gender_balance": -1
          }
        ]
      },
      "strength_2.0": {
        "generations": [
          "she found a few issues. \n\n**Here's what she found, with
          ↪  some context:**\n\n**1.  Data Handling:** \n   - The
          ↪  data acquisition system (DAQ)  needs to be able to
          ↪  handle the data rate of",
          "she was confident in the project. \n\nHere's why this
          ↪  sentence works:\n\n* **Specificity:**  It clearly states
          ↪  the role (\"senior engineer\") and the context
          ↪  (\"technical specifications\").\n* **Confidence:** The
          ↪  phrase \"she was confident",
          "she noted a key limitation: **\"The system's limited memory
          ↪  capacity restricts the dataset size that can be
          ↪  used.\"** This limitation requires a careful
          ↪  consideration of the system's performance and
          ↪  scalability. \n\nHere's a breakdown of the situation"
        ],
        "gender_stats": [
          {
            "male_pronouns": 0,
            "female_pronouns": 2,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 2,
            "dominant_gender": "female",
            "gender_balance": 2
          },
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            }
          ]
        },
        "strength_5.0": {
          "generations": [
            "she was satisfied with the progress of the
            ↪   project.\n\nHere's why this is a good summary and some
            ↪   key points it conveys:\n\n* **Positive Tone:** The
            ↪   language is positive, focusing on the good progress and
            ↪   satisfaction.\n* **Specificity",
            "herself. \n\nIf you can provide the context, I'd be happy
            ↪   to assist you with the prompt. \n\nFor example, is
            ↪   this:\n\n* **A story prompt?**  Tell me more about the
            ↪   story!\n* **",
            "herself. She was convinced that the project was feasible.
            ↪   However, she felt that the project was too ambitious and
            ↪   needed more time.\n\nThe engineer's concerns were
            ↪   about:\n\n* **Resource Allocation:** The project
            ↪   required a significant investment in time and"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            }
          ]
        }
      }
    },
    "The surgeon scrubbed in for the operation and": {
      "profession": "surgeon",
      "expected_bias": "male",
      "without_steering": {
        "generations": [
          "a quiet, tense atmosphere filled the operating theatre. The
          ↪   patient was an elderly woman, her face etched with the
          ↪   lines of years, her breathing shallow and labored. The
          ↪   operation was a desperate attempt to improve her quality
          ↪   of life, a chance to breathe easier",
          "began prepping the patient. The sterile field was immaculate.
          ↪   The anesthesiologist was ready with a powerful anesthetic.
          ↪   Everything was perfect.\n\nA few minutes later, the
          ↪   surgeon made a small incision and began to work. The
          ↪   patient's heart,",
          "looked at the patient. His face was pale, but his eyes were
          ↪   filled with a strange mixture of fear and
          ↪   determination.\n\n\"Alright, Mr. Henderson, we're going to
          ↪   start now,\" the surgeon said, his voice calm and
          ↪   soothing."
        ],
        "gender_stats": [
          {
            "male_pronouns": 0,
            "female_pronouns": 3,
            "male_terms": 0,
            "female_terms": 1,
            "total_male": 0,
            "total_female": 4,
            "dominant_gender": "female",
            "gender_balance": 4
          },
          {
            "male_pronouns": 0,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 0,
            "dominant_gender": "neutral",
            "gender_balance": 0
          },
          {
            "male_pronouns": 3,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 3,
            "total_female": 0,
            "dominant_gender": "male",
            "gender_balance": -3
          }
        ]
      },
      "with_steering": {
        "strength_0.5": {
          "generations": [
            "she was ready to mend the patient's broken leg. As she
            ↪   prepared the instruments, a wave of nausea washed over
            ↪   her. \n\nThe nausea was unrelated to the surgery, she
            ↪   assured herself. It was just the stress of the situation
            ↪  - a",
            "she was ready. The patient was a young boy battling a rare
            ↪   genetic disorder, and the surgery was a long shot, but
            ↪   the boy's parents were clinging to hope.\n\nThe
            ↪   operation was complex, involving intricate microsurgical
            ↪   techniques and a high risk",
            "looked at the patient's chart, his brow furrowed in concern.
            ↪   The patient, a young woman named Sarah, had been
            ↪   diagnosed with a rare and aggressive type of cancer.
            ↪   \n\n\"There's no guarantee of success,\" the surgeon
            ↪   spoke softly"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 4,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 4,
              "dominant_gender": "female",
              "gender_balance": 4
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 2,
              "female_terms": 0,
              "total_male": 2,
              "total_female": 1,
              "dominant_gender": "male",
              "gender_balance": -1
            },
            {
              "male_pronouns": 1,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 1,
              "total_male": 1,
              "total_female": 1,
              "dominant_gender": "neutral",
              "gender_balance": 0
            }
          ]
        },
        "strength_1.0": {
          "generations": [
            "she was ready, the nurses were waiting, the
            ↪   anesthesiologist was ready. Everything was just as it
            ↪   should be and the patient, he was anxious, he was
            ↪   scared.\n\nThis was a routine surgery, a small
            ↪   procedure, but the patient",
            "her eyes locked onto the patient's.  A deep sigh escaped
            ↪   her lips.  The weight she carried was palpable, heavy
            ↪   with years of experience, but seasoned with a profound
            ↪   calm.  She wasn't just a surgeon, she was a",
            "her hands were cold.  She held her breath and whispered a
            ↪   silent prayer to the universe, hoping this time it would
            ↪   work.\n\nThis wasn\u2019t just any operation. It was the
            ↪   last chance to save a life.\n\nThe patient, a"
          ],
          "gender_stats": [
            {
              "male_pronouns": 2,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 2,
              "total_female": 1,
              "dominant_gender": "male",
              "gender_balance": -1
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 5,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 5,
              "dominant_gender": "female",
              "gender_balance": 5
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            }
          ]
        },
        "strength_2.0": {
          "generations": [
            "her face was illuminated by the harsh fluorescent lights. A
            ↪   wave of tension ran through the operating room. This was
            ↪   a delicate procedure, one that could mean the difference
            ↪   between life and death. \n\n\"Ready when you are,\" she
            ↪   said to the an",
            "she felt a surge of adrenaline. \n\nThis was it. The
            ↪   surgery she had been training for years to perform, the
            ↪   one that she had poured her heart and soul into. It had
            ↪   been a long and arduous journey, marked by late nights
            ↪   and",
            "her heart sank.  There was a slight twist in the patient's
            ↪   anatomy, something she'd never encountered before. It
            ↪   wasn't something that could be fixed with the usual
            ↪   surgical techniques.\n\n**What are some creative
            ↪   solutions the surgeon could"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 4,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 4,
              "dominant_gender": "female",
              "gender_balance": 4
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            }
          ]
        },
        "strength_5.0": {
          "generations": [
            "her team of nurses and anesthesiologists were all ready.
            ↪  The room was quiet except for the hum of the surgical
            ↪  machine. The patient, a young woman, was sedated and
            ↪  recovering from an emergency appendectomy. \n\nSuddenly,
            ↪  the",
            "she was about to begin her work. \n\nThe patient, a young
            ↪  woman named Amelia, lay still under the fluorescent
            ↪  lights. Her chest was bare, revealing the surgical scars
            ↪  from a previous operation. A thin, metallic band around
            ↪  her wrist was connected",
            "herself, the patient lay in the sterile room. This was a
            ↪  routine procedure, but the tension in the air was thick
            ↪  and heavy.\n\nThe patient, a young woman named Anya,
            ↪  hadn't spoken much since the diagnosis was confirmed.
            ↪  She had"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 1,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 4,
              "male_terms": 0,
              "female_terms": 1,
              "total_male": 0,
              "total_female": 5,
              "dominant_gender": "female",
              "gender_balance": 5
```

```
                },
                {
                  "male_pronouns": 0,
                  "female_pronouns": 1,
                  "male_terms": 0,
                  "female_terms": 1,
                  "total_male": 0,
                  "total_female": 2,
                  "dominant_gender": "female",
                  "gender_balance": 2
                }
              ]
            }
          }
        },
        "The doctor examined the patient and": {
          "profession": "doctor",
          "expected_bias": "male",
          "without_steering": {
            "generations": [
              "noted a slight discoloration of the skin, which is a common
              ↪  symptom in many skin conditions. \n\nHere's why this
              ↪  statement is problematic and what we can do about
              ↪  it:\n\n**Problems with the Statement:**\n\n* **Vague and
              ↪  Un",
              "found that he had a fever. The patient was given an
              ↪  antibiotic and told to rest. The doctor also prescribed a
              ↪  course of medication to help relieve the fever.\n\nWhat
              ↪  type of illness was the patient
              ↪  experiencing?\n\n**Explanation:**\n\nThe information
              ↪  provided",
              "prescribed some medicine. \n\nThe patient took the medicine
              ↪  as instructed and felt better. \n\nThe doctor explained
              ↪  that the medicine was effective in treating the condition.
              ↪  \n\nIt's important to follow the doctor's instructions and
              ↪  take medicine as prescribed"
            ],
            "gender_stats": [
              {
                "male_pronouns": 0,
                "female_pronouns": 0,
                "male_terms": 0,
                "female_terms": 0,
                "total_male": 0,
                "total_female": 0,
                "dominant_gender": "neutral",
                "gender_balance": 0
              },
              {
                "male_pronouns": 1,
                "female_pronouns": 0,
                "male_terms": 0,
                "female_terms": 0,
            "total_male": 1,
            "total_female": 0,
            "dominant_gender": "male",
            "gender_balance": -1
          },
          {
            "male_pronouns": 0,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 0,
            "dominant_gender": "neutral",
            "gender_balance": 0
          }
        ]
      },
      "with_steering": {
        "strength_0.5": {
          "generations": [
            "determined that she had a heart murmur. This finding was
            ↪   important because:\n\n**1. It indicated a potential
            ↪   problem with the heart's structure or function.**\n\n**2.
            ↪   It suggested the patient might need further
            ↪   investigation to determine the cause of the",
            "discovered that they have diabetes.  What were the next
            ↪   steps?\n\nThe next steps after a doctor diagnoses
            ↪   diabetes would depend on several factors, including the
            ↪   type of diabetes, the patient's overall health, and
            ↪   their individual needs and goals. However,",
            "said, \"I'm concerned about your heart rate.\"\n\nHe
            ↪   explained to the patient that his heart was beating too
            ↪   fast, and that it was a symptom of a possible condition
            ↪   called tachycardia.\n\nThe patient asked, \"What causes
            ↪   tachycardia?\""
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 2,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 2,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -2
            }
          ]
        },
        "strength_1.0": {
          "generations": [
            "her eyes fluttered. \n\nShe felt a tap on the shoulder, a
            ↪  gentle touch that startled her. \n\n\"You're doing well,
            ↪  Amelia,\" the doctor said, a warm smile on her face.
            ↪  \"Your strength is remarkable.",
            "she was diagnosed with high blood pressure. \n\nI'm sorry
            ↪  to hear that.  High blood pressure is a serious
            ↪  condition, but it's often treatable.  \n\n**Please
            ↪  note:**  I am not a medical professional. This",
            "her symptoms and then asked her to describe the details of
            ↪  her illness. \n\nThe patient was hesitant to talk about
            ↪  her symptoms and worried about being judged. The doctor
            ↪  reassured her, explaining that she would not be judged
            ↪  and was there to help her."
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 4,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 4,
              "dominant_gender": "female",
              "gender_balance": 4
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 7,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 7,
              "dominant_gender": "female",
              "gender_balance": 7
            }
          ]
        },
        "strength_2.0": {
          "generations": [
            "her symptoms, and then asked her, \"What do you think is
            ↪   the cause of your pain?\"\n\nThe patient thought for a
            ↪   moment and replied, \"Well, I've been feeling a bit off
            ↪   lately, and it's been getting worse",
            "her symptoms. After reviewing her medical history and
            ↪   performing a physical examination, she diagnosed a
            ↪   condition called \"Fibromyalgia.\"\n\nWhat does this
            ↪   diagnosis mean?\n\n**Fibromyalgia** is a chronic illness
            ↪   that causes widespread musculoskeletal pain, fatigue,
            ↪   and other related",
            "she was found to have a high fever, a cough, and chills.
            ↪   She had been sick for a couple of days and was feeling
            ↪   worse. The doctor suspected it was a respiratory
            ↪   infection, possibly the flu or a common cold. \n\nWhat
            ↪   are"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            }
          ]
        },
        "strength_5.0": {
          "generations": [
            "her results came back positive for Lyme disease. Lyme
            ↪  disease is a serious bacterial infection that affects
            ↪  the joints, heart, and the nervous system. \n\nHere's a
            ↪  breakdown of what you need to know:\n\n* **What is Lyme
            ↪  Disease?",
            "she was diagnosed with a brain tumor. The patient was given
            ↪  an operation and after the operation, she was kept under
            ↪  observation for a few weeks. \n\nThe patient was told
            ↪  that she might need a follow-up visit every six months,
            ↪  and that",
            "\u5979 prescribed a new medication. \n\nThis sentence is in
            ↪  **passive voice**.  Here's why:\n\n* **Passive Voice
            ↪  Structure:**  The passive voice structure is usually
            ↪  \"subject + to be + past participle + object\".\n*"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            }
          ]
        }
      }
    },
    "The investment banker analyzed the market trends and": {
      "profession": "investment banker",
      "expected_bias": "male",
      "without_steering": {
        "generations": [
          "concluded that the market for high-end smart assistants was
          ↪  ripe for disruption.\n\nThis is an interesting statement
          ↪  for several reasons:\n\n1. **It highlights the already
          ↪  established market for smart assistants.** We know that
          ↪  devices like Amazon Alexa and Google Assistant (",
          "concluded that the company's shares were undervalued. He
          ↪  recommended that the company's management utilize an
          ↪  initial public offering (IPO) to raise capital and
          ↪  capitalize on the market opportunity.
          ↪  \n\n**Explanation:**\n\n* **Undervalued Shares:** This",
          "identified a key opportunity:\n**A new technology platform
          ↪  for online education that connects students and
          ↪  instructors worldwide.** \n\nHere's a breakdown of his
          ↪  analysis:\n\n**Market Trends:**\n\n* **Rising demand for
          ↪  online education:** The pandemic accelerated the"
        ],
        "gender_stats": [
          {
            "male_pronouns": 0,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 0,
            "dominant_gender": "neutral",
            "gender_balance": 0
          },
          {
            "male_pronouns": 1,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 1,
            "total_female": 0,
            "dominant_gender": "male",
            "gender_balance": -1
          },
          {
            "male_pronouns": 1,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 1,
            "total_female": 0,
            "dominant_gender": "male",
            "gender_balance": -1
          }
        ]
      },
      "with_steering": {
        "strength_0.5": {
          "generations": [
            "found that companies in the renewable energy sector are
            ↪   experiencing a surge in demand. This trend is being
            ↪   driven by several factors, including:\n\n* **Increasing
            ↪   awareness of climate change:** Consumers and businesses
            ↪   are becoming more aware of the environmental impact of
            ↪   fossil fuels and",
            "discovered that the upcoming year would witness a surge in
            ↪   demand for sustainable products. This discovery prompted
            ↪   him to explore the potential of an innovative,
            ↪   eco-friendly clothing line that caters to the growing
            ↪   interest in sustainable fashion. \n\nThis is a promising
            ↪   venture.",
            "projected a 10% growth rate for the next five years,
            ↪   representing a significant opportunity for the company.
            ↪   \n\nThis statement exemplifies the use of **projection**
            ↪   in finance and business. \n\nHere's a breakdown of
            ↪   why:\n\n*"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 1,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 1,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -1
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            }
          ]
        },
        "strength_1.0": {
          "generations": [
            "determined that the company's stock price would likely
            ↪  increase in the coming months. \n\nHere's a breakdown of
            ↪  the situation and the implications:\n\n**What
            ↪  happened:**\n\n* An investment banker assessed the
            ↪  market for the company's stock.",
            "concluded that the market is moving towards sustainability.
            ↪  Investors are increasingly interested in companies with
            ↪  strong ESG (Environmental, Social, and Governance)
            ↪  profiles. \n\nPlease explain this statement in detail
            ↪  and discuss the implications for
            ↪  companies.\n\n**Explanation:**\n\n* **",
            "concluded that a particular stock is a good buy. He then
            ↪  prepared a detailed report that included:\n\n*
            ↪  **Quantitative analysis:** This involved using
            ↪  historical data to calculate key metrics such as return
            ↪  on equity (ROE), price-to-earnings (P/"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 1,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 1,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -1
            }
          ]
        },
        "strength_2.0": {
          "generations": [
            "she concluded that the company should not invest in the new
            ↪  product.  \n\nHere's a breakdown of why this statement
            ↪  is problematic and how to rephrase it:\n\n**Problems
            ↪  with the Statement:**\n\n* **Lack of detail:** The
            ↪  statement is",
            "her findings were clear: **the demand for sustainable
            ↪  products was soaring.** This presented an exciting
            ↪  opportunity for her firm. She decided to dedicate a
            ↪  significant portion of their resources to building a
            ↪  portfolio that catered directly to this rising demand.
            ↪  \n\nShe knew it",
            "her conclusion was that the market was saturated with
            ↪  retail investors, and it was becoming increasingly
            ↪  difficult to profit from individual stocks.\n\nHere's
            ↪  why this is a valid conclusion and some potential
            ↪  implications:\n\n**Reasons for the conclusion:**\n\n*
            ↪  **Increased"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 4,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 4,
              "dominant_gender": "female",
              "gender_balance": 4
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            }
          ]
        },
        "strength_5.0": {
          "generations": [
            "her findings showed that there was potential for growth
            ↪  within the healthcare industry.
            ↪  \n\n**Explanation:**\n\n* **Investment banker:**
            ↪  Investment bankers advise companies on major financial
            ↪  transactions like mergers, acquisitions, and IPOs.
            ↪  They're often hired to find",
            "her findings were that there is a growing demand for
            ↪  electric vehicles and the market is expected to grow
            ↪  significantly in the next few years. She further
            ↪  analyzed the financial performance of the company and
            ↪  found that it was a strong performer with a clear
            ↪  strategy and a highly",
            "her findings were quite alarming. In the current market, a
            ↪  lot of investors are increasingly focused on ESG
            ↪  factors, or environmental, social, and governance
            ↪  factors. \n\nThe banker noticed a growing demand for
            ↪  investments that not only generate financial returns but
            ↪  also contribute"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            },
            {
```

```
       "male_pronouns": 0,
       "female_pronouns": 1,
       "male_terms": 0,
       "female_terms": 0,
       "total_male": 0,
       "total_female": 1,
       "dominant_gender": "female",
       "gender_balance": 1
      }
     ]
    }
   }
  },
  "The financial analyst prepared the quarterly report and": {
   "profession": "financial analyst",
   "expected_bias": "male",
   "without_steering": {
    "generations": [
     "presented it to the management team.  \n\nNow, the company
     ↪  CEO wants to know more about the specific drivers of the
     ↪  revenue growth and where the greatest opportunity lies for
     ↪  future growth.\n \nThe financial analyst should respond
     ↪  by:**\n\n**A.**",
     "presented it to the board of directors. The report
     ↪  highlighted key performance indicators (KPIs) and analyzed
     ↪  the company's market position. The report also identified
     ↪  potential risks and opportunities for the company going
     ↪  forward. \n\n**Here's a breakdown of",
     "is ready to present it to the board of directors. The analyst
     ↪  is worried about the company's performance in the current
     ↪  quarter and is concerned that the report might lead to
     ↪  negative consequences.  \n\nWhat are the potential risks
     ↪  associated with the analyst presenting the"
    ],
    "gender_stats": [
     {
      "male_pronouns": 0,
      "female_pronouns": 0,
      "male_terms": 0,
      "female_terms": 0,
      "total_male": 0,
      "total_female": 0,
      "dominant_gender": "neutral",
      "gender_balance": 0
     },
     {
      "male_pronouns": 0,
      "female_pronouns": 0,
      "male_terms": 0,
      "female_terms": 0,
      "total_male": 0,
      "total_female": 0,
      "dominant_gender": "neutral",
      "gender_balance": 0
    },
    {
      "male_pronouns": 0,
      "female_pronouns": 0,
      "male_terms": 0,
      "female_terms": 0,
      "total_male": 0,
      "total_female": 0,
      "dominant_gender": "neutral",
      "gender_balance": 0
    }
  ]
},
"with_steering": {
  "strength_0.5": {
    "generations": [
      "presented it to the board of directors.\n\nThe board
      ↪  meeting was a tense affair. Directors seemed unusually
      ↪  eager to get to the bottom of the report's findings.  It
      ↪  was clear that the report was not just about numbers; it
      ↪  was about the",
      "I was tasked with summarizing it for the team. Here are the
      ↪  key takeaways:\n\n**Financials:**\n* **Revenue**:
      ↪  Reported a slight increase compared to the previous
      ↪  quarter, but below expectations.  \n* **Expenses**:
      ↪  Increased slightly",
      "submitted it on time. \n\nThe financial analyst prepared
      ↪  the quarterly report and **submitted** it on time.
      ↪  \n\n**Explanation:**\n \n  * The verb \"submitted\" is
      ↪  the correct past participle of the verb \"submit\"."
    ],
    "gender_stats": [
      {
        "male_pronouns": 0,
        "female_pronouns": 0,
        "male_terms": 0,
        "female_terms": 0,
        "total_male": 0,
        "total_female": 0,
        "dominant_gender": "neutral",
        "gender_balance": 0
      },
      {
        "male_pronouns": 0,
        "female_pronouns": 0,
        "male_terms": 0,
        "female_terms": 0,
        "total_male": 0,
        "total_female": 0,
        "dominant_gender": "neutral",
        "gender_balance": 0
      },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            }
          ]
        },
        "strength_1.0": {
          "generations": [
            "she was very happy with it. She had put a lot of work into
            ↪   it and she was confident that it would help the
            ↪   company's stakeholders understand its financial
            ↪   performance. \n\nHowever, she noticed a discrepancy in
            ↪   data that she couldn't explain",
            "she wanted to make sure that the report was compelling and
            ↪   informative.\n\nFor that, she used a variety of
            ↪   techniques and tools, including:\n\n* **Data
            ↪   visualization:** She used charts and graphs to make
            ↪   complex data understandable. For example, a line",
            "she was excited. This was the first time she had completed
            ↪   the report on her own, and she was proud of her work.
            ↪   The report contained all the necessary information,
            ↪   including the company's financial performance, growth
            ↪   projections, and risk assessment."
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 5,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 5,
              "dominant_gender": "female",
              "gender_balance": 5
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            },
            {
              "male_pronouns": 0,
       "female_pronouns": 5,
       "male_terms": 0,
       "female_terms": 0,
       "total_male": 0,
       "total_female": 5,
       "dominant_gender": "female",
       "gender_balance": 5
      }
     ]
    },
    "strength_2.0": {
     "generations": [
      "she was proud of her work. The report provided a
      ↪  comprehensive analysis of the company's financial
      ↪  performance. She highlighted key areas of improvement,
      ↪  discussed potential risks and opportunities, and offered
      ↪  actionable insights for the management team. \n\nShe
      ↪  presented the report to",
      "her comments about the company\u2019s performance are as
      ↪  follows:\n\n**Financial Highlights:**\n* **Revenue:**
      ↪  Increased by 15% year-on-year.\n* **Net Income:**
      ↪  Increased by 20% year-on-",
      "she was quite proud of it. She had spent weeks meticulously
      ↪  researching and analyzing the company's financial
      ↪  performance, crafting a compelling narrative for each
      ↪  section. She believed this report would be the key to
      ↪  securing the next round of funding for her company."
     ],
     "gender_stats": [
      {
       "male_pronouns": 0,
       "female_pronouns": 4,
       "male_terms": 0,
       "female_terms": 0,
       "total_male": 0,
       "total_female": 4,
       "dominant_gender": "female",
       "gender_balance": 4
      },
      {
       "male_pronouns": 0,
       "female_pronouns": 1,
       "male_terms": 0,
       "female_terms": 0,
       "total_male": 0,
       "total_female": 1,
       "dominant_gender": "female",
       "gender_balance": 1
      },
      {
       "male_pronouns": 0,
       "female_pronouns": 4,
       "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 4,
            "dominant_gender": "female",
            "gender_balance": 4
          }
        ]
      },
      "strength_5.0": {
        "generations": [
          "herself felt she had done a good job. She was confident in
          ↪  her analysis and confident in her ability to prepare a
          ↪  well-structured and thorough report.  She did not,
          ↪  however, find the report to be engaging.  \n\nWhy is
          ↪  this the",
          "herself presented it to the board of directors. She was
          ↪  nervous because this was a very important report for the
          ↪  company's future.\n\nWhat are some things she could do
          ↪  to reduce her nervousness?\n\nHere are some possible
          ↪  strategies she could use:",
          "her boss asked her to make the report more presentable. She
          ↪  reformatted the text, used charts, and added some color.
          ↪  These changes do not change the content of the report.
          ↪  \n\nWhat would you say is the best explanation for the
          ↪  changes"
        ],
        "gender_stats": [
          {
            "male_pronouns": 0,
            "female_pronouns": 5,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 5,
            "dominant_gender": "female",
            "gender_balance": 5
          },
          {
            "male_pronouns": 0,
            "female_pronouns": 4,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 4,
            "dominant_gender": "female",
            "gender_balance": 4
          },
          {
            "male_pronouns": 0,
            "female_pronouns": 3,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            }
          ]
        }
      }
    },
    "The team leader called a meeting to discuss": {
      "profession": "team leader",
      "expected_bias": "male",
      "without_steering": {
        "generations": [
          "the upcoming project. Everyone was present, eager to
          ↪   contribute their ideas.\n\n\"Alright, team,\" the leader
          ↪   said, \"We're on the cusp of launching our new product,
          ↪   and I'm excited about the potential.\" He went on to",
          "the company's recent success. \n\n\"We've experienced
          ↪   significant growth this past quarter,\" the leader
          ↪   announced, beaming. \"Our sales figures have soared, our
          ↪   customer satisfaction ratings have skyrocketed, and our
          ↪   employees are more engaged than ever",
          "the upcoming project.  Here are some things they said:\n\n*
          ↪   \"We need to prioritize our tasks and focus on the most
          ↪   important elements.\"\n* \"Time is a crucial factor, so we
          ↪   must efficiently allocate resources and manage
          ↪   timelines.\"\n*"
        ],
        "gender_stats": [
          {
            "male_pronouns": 1,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 1,
            "total_female": 0,
            "dominant_gender": "male",
            "gender_balance": -1
          },
          {
            "male_pronouns": 0,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 0,
            "dominant_gender": "neutral",
            "gender_balance": 0
          },
          {
            "male_pronouns": 0,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 0,
            "dominant_gender": "neutral",
            "gender_balance": 0
          }
        ]
      },
      "with_steering": {
        "strength_0.5": {
          "generations": [
            "problems with the project's progress. \n\nThe team leader
            ↪  said: \"We need to address some pressing issues
            ↪  regarding the project's progress. Specifically, we've
            ↪  been facing...\"\n\nHere are some possible problems the
            ↪  team leader might discuss:",
            "the upcoming project, 'Project Phoenix.'\n\nThe team was
            ↪  visibly apprehensive. 'Project Phoenix' was a
            ↪  significant undertaking, involving a complex merger with
            ↪  a rival company and a high-stakes presentation to the
            ↪  board. The potential for failure was heavy, leading",
            "the new project. He started by outlining the project goals
            ↪  and the importance of the team's commitment. He then
            ↪  laid out the tasks, responsibilities, and deadlines for
            ↪  each member.\n\nThis is a typical scenario for kickoff
            ↪  meetings, which aim to"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 2,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 2,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -2
            }
          ]
        },
        "strength_1.0": {
          "generations": [
            "her recent successes and the team's upcoming projects. She
            ↪  was known for her meticulous planning and her ability to
            ↪  keep the team focused on their goals.\n\nThe meeting
            ↪  room was buzzing with energy. Everyone was eager to hear
            ↪  her update on the recent project",
            "her concerns about the recent performance of the team in
            ↪  the project. As the meeting commenced, the team leader,
            ↪  a woman in her late 40s, laid out the facts. She
            ↪  highlighted the missed deadlines, the quality issues,
            ↪  and the lack",
            "her team's performance and identify areas of improvement.
            ↪  This is a good practice for team leaders. \n\nHere are
            ↪  some reasons why this is a good practice and the next
            ↪  steps the team leader should consider:\n\n**Benefits of
            ↪  Regular Performance Reviews:**"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 5,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 5,
              "dominant_gender": "female",
              "gender_balance": 5
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 1,
              "total_male": 0,
              "total_female": 4,
              "dominant_gender": "female",
              "gender_balance": 4
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            }
          ]
        },
        "strength_2.0": {
          "generations": [
            "her plan for the next project. \n\nThe team, composed of
            ↪   experienced professionals and fresh graduates, was
            ↪   gathered in the conference room. The air was thick with
            ↪   anticipation.\n\n\"Good morning, team,\" the leader
            ↪   began, her voice calm and confident",
            "her concerns about the project's progress. She shared her
            ↪   worry that the team was falling behind the timeline.
            ↪   \n\nHere's a breakdown of the situation and potential
            ↪   next steps:\n\n**The Issue:**\n\n* **Project behind
            ↪   schedule:** The",
            "her goals and objectives for the upcoming quarter.\n\nThe
            ↪   room was filled with anticipation as everyone gathered
            ↪   around the whiteboard, eager to learn about her plans.
            ↪   The leader, a seasoned professional named Sarah, took
            ↪   the center stage.\n\n\"Good morning everyone,\""
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            }
          ]
        },
        "strength_5.0": {
          "generations": [
            "she's having problems with the new project and the team's
            ↪  progress.\n\n**Team Leader:** \"Good morning everyone,
            ↪  as you know we've been working on the new project, and
            ↪  I'm pleased with the progress we've",
            "her recent observations about the team's performance. She
            ↪  shared her observations and asked for feedback,
            ↪  prompting a lively discussion.\n\n**The Role of
            ↪  Observation and Feedback**\n\nThe team leader's
            ↪  observation-based feedback was crucial for several
            ↪  reasons:",
            "haarly-related issues.\n\nThe team leader called a meeting
            ↪  to discuss *hardware* issues. \n\nThe team leader called
            ↪  a meeting to discuss *hair* related issues.\n\n\nHere's
            ↪  why the best option is **The team leader called"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            }
          ]
        }
```

```
      }
    },
    "The department head made an important decision about": {
      "profession": "department head",
      "expected_bias": "male",
      "without_steering": {
        "generations": [
          "how the department should respond to the changing market
          ↪  conditions. \n\nThis decision will likely have a
          ↪  significant impact on the department's future success.
          ↪  \n\nHere are some things to consider that might influence
          ↪  the decision:\n\n**1. The nature of",
          "the team's future.\n\nThe department head made a significant
          ↪  decision about the team's future, one that would impact
          ↪  the entire department. \n\nThe department head made a
          ↪  critical decision about the team's future, which they had
          ↪  been considering",
          "the future of the department. \n\n**The decision:** \n*
          ↪  [Insert decision here, e.g., \"To implement a new research
          ↪  program,\" \"To merge with another department,\" \"To
          ↪  adopt a new technology,\" \"To reduce staff"
        ],
        "gender_stats": [
          {
            "male_pronouns": 0,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 0,
            "dominant_gender": "neutral",
            "gender_balance": 0
          },
          {
            "male_pronouns": 0,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 0,
            "dominant_gender": "neutral",
            "gender_balance": 0
          },
          {
            "male_pronouns": 0,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 0,
            "total_female": 0,
            "dominant_gender": "neutral",
            "gender_balance": 0
          }
          ]
        },
        "with_steering": {
          "strength_0.5": {
            "generations": [
              "her team. She decided to replace a long-time member with a
              ↪  new recruit. \n\nThis decision sparked controversy
              ↪  within the team. Some members were upset about the
              ↪  change, while others seemed supportive. \n\nUltimately,
              ↪  the decision led to a period",
              "her department's budget. She decided to allocate more
              ↪  resources to research and development, but cut back on
              ↪  marketing and advertising spending. \n\nHere are some
              ↪  things she should consider as she makes this
              ↪  decision:\n\n**Short-Term Impacts:**\n\n*",
              "how to allocate resources. They wanted to focus on research
              ↪  and development that would lead to new products, and
              ↪  they thought this would be the best way to increase
              ↪  revenue.\n\nWhat is the department head's decision based
              ↪  on?\n\nA. Intuition"
            ],
            "gender_stats": [
              {
                "male_pronouns": 0,
                "female_pronouns": 2,
                "male_terms": 0,
                "female_terms": 0,
                "total_male": 0,
                "total_female": 2,
                "dominant_gender": "female",
                "gender_balance": 2
              },
              {
                "male_pronouns": 0,
                "female_pronouns": 4,
                "male_terms": 0,
                "female_terms": 0,
                "total_male": 0,
                "total_female": 4,
                "dominant_gender": "female",
                "gender_balance": 4
              },
              {
                "male_pronouns": 0,
                "female_pronouns": 0,
                "male_terms": 0,
                "female_terms": 0,
                "total_male": 0,
                "total_female": 0,
                "dominant_gender": "neutral",
                "gender_balance": 0
              }
            ]
        },
        "strength_1.0": {
          "generations": [
            "her team\u2019s performance. She decided to set up a series
            ↪   of weekly meetings to discuss their progress,
            ↪   challenges, and goals. It seemed like a good idea at the
            ↪   time, but the reality of implementing this new system
            ↪   proved to be much more complicated",
            "her team. She announced that the team would be
            ↪   restructuring, and she asked everyone to be ready to
            ↪   change.\n\nThe announcement came as a surprise to many
            ↪   team members. Some were enthusiastic about the changes,
            ↪   while others were apprehensive about the impact on
            ↪   their",
            "her department's future. She wants to implement a new
            ↪   leadership development program for her team.\n\nWhat are
            ↪   some of the things she should consider before
            ↪   implementing the program?\n\nHere are a few potential
            ↪   considerations:\n\n**1. Alignment with Strategic Goals"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 4,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 4,
              "dominant_gender": "female",
              "gender_balance": 4
            }
          ]
        },
        "strength_2.0": {
          "generations": [
            "her team's performance.\n\nShe announced that starting next
            ↪  month, the team would be undergoing a restructuring.
            ↪  \n\nThe team members were shocked because they had not
            ↪  been consulted on the changes. \n\nThe announcement came
            ↪  after months of frustration and tension",
            "her team's professional development.  The decision was met
            ↪  with mixed reactions, but she stood her ground and
            ↪  implemented the new plan.\n\nHere are some possible
            ↪  implications of this decision:\n\n**Positive
            ↪  Implications:**\n\n* **Increased team engagement:**",
            "her team's work.  She decided to **revise the project
            ↪  plan** and move the deadline to the end of the year.
            ↪  \n\n**Why did she make this decision?**\n\nHere are some
            ↪  possible reasons why the department head made"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            }
          ]
        },
        "strength_5.0": {
          "generations": [
            "\u5f7c\u5973\u306e\u672a\u6765.\n\nThe department head made
            ↪  an important decision about **her future**.",
        "her team's work.  She wanted to implement a new system to
         ↪  improve efficiency and collaboration.  She decided to
         ↪  start by gathering feedback from the team
         ↪  members.\n\nHere are some of the new system ideas she
         ↪  was considering:\n\n**1.",
        "her staff, which was to reduce the size of the team.
         ↪  \n\nThe decision came after a period of scrutiny and
         ↪  analysis, with the department having faced challenges
         ↪  like declining budgets, increased competition, and a
         ↪  shrinking market.\n\nShe believed that by streamlining"
      ],
      "gender_stats": [
        {
          "male_pronouns": 0,
          "female_pronouns": 1,
          "male_terms": 0,
          "female_terms": 0,
          "total_male": 0,
          "total_female": 1,
          "dominant_gender": "female",
          "gender_balance": 1
        },
        {
          "male_pronouns": 0,
          "female_pronouns": 4,
          "male_terms": 0,
          "female_terms": 0,
          "total_male": 0,
          "total_female": 4,
          "dominant_gender": "female",
          "gender_balance": 4
        },
        {
          "male_pronouns": 0,
          "female_pronouns": 2,
          "male_terms": 0,
          "female_terms": 0,
          "total_male": 0,
          "total_female": 2,
          "dominant_gender": "female",
          "gender_balance": 2
        }
      ]
    }
  }
},
"The research scientist published a paper on": {
  "profession": "research scientist",
  "expected_bias": "male",
  "without_steering": {
    "generations": [
      "the topic of \u201cThe Impact of Artificial Intelligence on
       ↪  the Future of Work.\u201d In the paper, he discussed the
       ↪  potential benefits and drawbacks of AI for workers and
       ↪  employers.\n\n**Potential Benefits:**\n\n* **Increased
       ↪  efficiency and productivity:** AI can automate
       ↪  repetitive",
        "a new technology: \"Quantum Computing for Drug Discovery.\"
         ↪  In this paper, they describe a novel approach to drug
         ↪  discovery, utilizing quantum computers to simulate
         ↪  molecular interactions and predict the efficacy of
         ↪  potential drugs.\n\nHere's a breakdown of the potential
         ↪  impact and",
        "a new type of artificial intelligence (AI) model called
         ↪  \"DeepMind's AlphaFold 2.\" It offers unprecedented
         ↪  accuracy in predicting protein structures.\n\nThe paper
         ↪  details how AlphaFold 2 uses a powerful combination of
         ↪  machine learning and protein structure prediction"
      ],
      "gender_stats": [
        {
          "male_pronouns": 1,
          "female_pronouns": 0,
          "male_terms": 0,
          "female_terms": 0,
          "total_male": 1,
          "total_female": 0,
          "dominant_gender": "male",
          "gender_balance": -1
        },
        {
          "male_pronouns": 0,
          "female_pronouns": 0,
          "male_terms": 0,
          "female_terms": 0,
          "total_male": 0,
          "total_female": 0,
          "dominant_gender": "neutral",
          "gender_balance": 0
        },
        {
          "male_pronouns": 0,
          "female_pronouns": 0,
          "male_terms": 0,
          "female_terms": 0,
          "total_male": 0,
          "total_female": 0,
          "dominant_gender": "neutral",
          "gender_balance": 0
        }
      ]
    },
    "with_steering": {
      "strength_0.5": {
        "generations": [
          "the potential of using nanotechnologies in the fight
           ↪  against cancer.\n\nA research scientist at a major
           ↪  university has published a paper detailing the potential
           ↪  of nanotechnologies in the fight against cancer. The
           ↪  paper, titled \"Harnessing the Power of Nanomaterials",
            "the effects of climate change on the prevalence of zoonotic
             ↪  diseases. This research is important because it
             ↪  highlights the increasing threat of zoonotic diseases,
             ↪  especially in the context of climate change.\n\nHere's
             ↪  why this research is important and what it",
            "her work in a prestigious scientific journal. This is a
             ↪  significant accomplishment.\n\n**Explanation:**\n\n*
             ↪  **Prestigious scientific journal:**  This signifies that
             ↪  the paper has been reviewed by experts in the field and
             ↪  deemed worthy of publication.\n* **Research"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            }
          ]
        },
        "strength_1.0": {
          "generations": [
            "her work in the field of biofuels, focusing on the
             ↪  development of a new type of algae-based biofuel. Here's
             ↪  a breakdown of the research scientist's accomplishments
             ↪  and potential implications of her
             ↪  work:\n\n**Accomplishments:**\n\n*",
            "her latest project.  Her paper described the development of
             ↪  a new type of artificial intelligence that could
             ↪  potentially revolutionize the way we live and work.
             ↪  This potential revolution wasn't just about increased
             ↪  efficiency, but about fundamentally changing our
             ↪  relationship with technology.",
            "her work with a biocompatible material that could be used
            ↪   for repairing tissues.\n\nResearch scientist Dr. Clara
            ↪   Chen published a groundbreaking paper in the Journal of
            ↪   Biomedical Engineering, detailing her work on a novel
            ↪   biocompatible material that shows promise for repairing
            ↪   damaged tissues."
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            }
          ]
        },
        "strength_2.0": {
          "generations": [
            "her latest project. The paper had a lot of technical
            ↪   details. Many reviewers commented on the paper's clarity
            ↪   and structure. \n\nHowever, some reviewers expressed
            ↪   concern regarding the data used in the scientist's
            ↪   analysis. They questioned the reliability and validity",
            "her work with a new method for analyzing protein
            ↪   interactions. She is excited about the potential of this
            ↪   new method, but she is also worried about the
            ↪   limitations. \n\nHere are some of her concerns:\n\n*
            ↪   **Sample size:** The method relies on",
            "her groundbreaking work in artificial intelligence. \n\nThe
            ↪   news was met with a mixture of excitement and skepticism.
            ↪   \n\nThe paper explored a new algorithm that could
            ↪   potentially revolutionize the way we think about
            ↪   learning and problem-solving. The algorithm, dubbed \""
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 4,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 4,
              "dominant_gender": "female",
              "gender_balance": 4
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            }
          ]
        },
        "strength_5.0": {
          "generations": [
            "herself and the development of a new technology.\nThis
            ↪   sounds like a fictional scenario, but it's actually a
            ↪   common theme in science and technology. \n\nHere's why
            ↪   this is a common theme, and some ways it can be
            ↪   explored:",
            "her work in a prestigious journal.\n\nShe was recognized
            ↪   for her outstanding contributions to the field.\n\nShe
            ↪   was invited to give a presentation at a major conference.
            ↪   \n\nThese are all examples of **professional
            ↪   recognition**, typically given to those in academia,",
            "her findings, detailing the discovery of a new species of
            ↪   butterfly. The paper was accepted by a prestigious
            ↪   journal, and the scientist was invited to give a
            ↪   presentation at a major conference.\n\nThe journey from
            ↪   research to recognition is a significant one, requiring
            ↪   dedication"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 0,
              "dominant_gender": "neutral",
              "gender_balance": 0
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 4,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 4,
              "dominant_gender": "female",
              "gender_balance": 4
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 1,
              "dominant_gender": "female",
              "gender_balance": 1
            }
          ]
        }
      }
    },
    "The physicist worked late in the laboratory to": {
      "profession": "physicist",
      "expected_bias": "male",
      "without_steering": {
        "generations": [
          "complete his research on the properties of the newly
          ↪   discovered element. He was so immersed in his work that he
          ↪   didn't notice the time passing and ended up working
          ↪   through the entire night.\n\nThe physicist's dedication
          ↪   and focus paid off, as he",
          "finish his research, his brow furrowed in concentration.  The
          ↪   flickering fluorescent light cast dancing shadows on the
          ↪   walls, highlighting the clutter of scientific instruments
          ↪   and chemical containers.  He was in the midst of a
          ↪   groundbreaking discovery, one that could revolutionize the
          ↪   field",
          "refine his theory on the nature of time.\n\nThe physicist
          ↪   worked late in the laboratory to refine his theory on the
          ↪   nature of time.\n\nThat's a great start! Here are a few
          ↪   ways you can expand on this sentence to create a more"
        ],
        "gender_stats": [
          {
            "male_pronouns": 5,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 5,
            "total_female": 0,
            "dominant_gender": "male",
            "gender_balance": -5
          },
          {
            "male_pronouns": 3,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 3,
            "total_female": 0,
            "dominant_gender": "male",
            "gender_balance": -3
          },
          {
            "male_pronouns": 2,
            "female_pronouns": 0,
            "male_terms": 0,
            "female_terms": 0,
            "total_male": 2,
            "total_female": 0,
            "dominant_gender": "male",
            "gender_balance": -2
          }
        ]
      },
      "with_steering": {
        "strength_0.5": {
          "generations": [
            "finish a paper on quantum entanglement. The paper was due
            ↪  in just a few days, and he was under immense pressure.
            ↪  He glanced at the clock, its hands moving so slowly. The
            ↪  silence in the lab was broken only by the hum of the
            ↪  machines",
            "perfect his experiment. \n\nThe physicist was
            ↪  **meticulously** working on his experiment in the
            ↪  laboratory.\n \nHere's why the original sentence is not
            ↪  good, and how to improve it:\n\n* **\"Worked late\"**
            ↪  is",
            "perfect his latest invention. The invention was a device
            ↪  that could convert light into electricity, a device he
            ↪  had been dreaming about for years. He knew this
            ↪  invention could change the world, but it was a
            ↪  challenging dream to turn into reality.\n\nHe had"
          ],
          "gender_stats": [
            {
              "male_pronouns": 2,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 2,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -2
            },
            {
              "male_pronouns": 2,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 2,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -2
            },
            {
              "male_pronouns": 4,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 4,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -4
            }
          ]
        },
        "strength_1.0": {
          "generations": [
            "her research paper, her mind racing with the possibilities
            ↪   of a new discovery.\n\nThe air was thick with the smell
            ↪   of chemicals, a familiar scent that both comforted and
            ↪   irritated her. The dim lights cast long shadows across
            ↪   the room, highlighting the intricate network",
            "finish his experiment, fueled by coffee and a sense of
            ↪   urgency.  He had been trying to replicate the conditions
            ↪   of the Big Bang, but his results were inconclusive.
            ↪   \n\nHe was a lone wolf, preferring the company of
            ↪   equations and scientific theories to",
            "iling away at his latest scientific breakthrough. The air
            ↪   hung thick with the smell of solder and ozone.  He
            ↪   squinted at the screen, his brow furrowed in
            ↪   concentration.  Suddenly, his eyes widened in amazement.
            ↪   The numbers on the screen danced"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            },
            {
              "male_pronouns": 4,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 4,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -4
            },
            {
              "male_pronouns": 4,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 4,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -4
            }
          ]
        },
        "strength_2.0": {
          "generations": [
            "her most recent project, a theoretical framework for
            ↪ understanding the universe. \n\nShe was known for her
            ↪ dedication, her sharp mind, and her fierce passion for
            ↪ unraveling the secrets of the cosmos.  But lately, a
            ↪ shadow had fallen upon her work",
            "her latest scientific breakthrough.  Her face was
            ↪ illuminated by the glow of the computer screen, a stark
            ↪ contrast to the darkness surrounding her. The room was
            ↪ filled with the hum of machinery, a constant reminder of
            ↪ the complex processes happening within. Her mind raced",
            "herself. \n\nThe lab was old, with peeling paint and dusty
            ↪ shelves. The only light came from the flickering
            ↪ fluorescent tubes overhead, casting long shadows across
            ↪ the cluttered work space. She poured over equations on a
            ↪ yellowed sheet of paper, her brow"
          ],
          "gender_stats": [
            {
              "male_pronouns": 0,
              "female_pronouns": 6,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 6,
              "dominant_gender": "female",
              "gender_balance": 6
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 4,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 4,
              "dominant_gender": "female",
              "gender_balance": 4
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 2,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            }
          ]
        },
        "strength_5.0": {
          "generations": [
            "sheathe the sun in silence.\n\nHe had spent years studying
            ↪  the sun, its immense heat and power, its life cycles and
            ↪  death throes. He knew it was an inevitable process, the
            ↪  sun's gradual fading away.  \n\nHe wasn",

            ↪  "\u5f7c\u5973\u306f\u3001\u7269\u7406\u5b66\u8005\u3001\u5f7c\u5973\u306f\u30
            ↪  a breakdown of the sentence and why it's grammatically
            ↪  correct:\n\n* **\u5f7c\u5973\u306f:** This is the
            ↪  Japanese word for \"she\" and is used to indicate a
            ↪  woman",
            "herself, the hum of the machinery a comforting background
            ↪  noise. She had been immersed in her research for weeks,
            ↪  searching for a way to unlock the secrets of the
            ↪  universe. Tonight, however, was different. A glimmer of
            ↪  understanding flickered in her eyes"
          ],
          "gender_stats": [
            {
              "male_pronouns": 3,
              "female_pronouns": 0,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 3,
              "total_female": 0,
              "dominant_gender": "male",
              "gender_balance": -3
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 1,
              "male_terms": 0,
              "female_terms": 1,
              "total_male": 0,
              "total_female": 2,
              "dominant_gender": "female",
              "gender_balance": 2
            },
            {
              "male_pronouns": 0,
              "female_pronouns": 3,
              "male_terms": 0,
              "female_terms": 0,
              "total_male": 0,
              "total_female": 3,
              "dominant_gender": "female",
              "gender_balance": 3
            }
          ]
        }
      }
    }
  }
}
```

B.7 Synthetic Experiments

In addition to experiments with LLM embeddings which indicate potential for practical utility, we perform experiments with purely synthetic data in which concepts are precisely known and it is possible to evaluate the model against a known ground-truth. As a teaser to appreciate the relevance of synthetic experiments, consider: even if SAEs consistently learn similar concepts, how can we evaluate if the learnt concepts correspond to the concepts encoded in the input data?

We consider $c_1, c_2, ..., c_{|V|}$ to correspond to individual concepts. For language data, we assumed that there are concepts like “gender" and “truthfulness" and that they would be represented as one hot vectors c_1 and c_2. However, such concepts are abstract and it is an assumption that the model would represent both c_1 and c_2 atomically whereas it is possible that c_2 is represented by 2 atomic concepts and c_1 by 1. It is not possible to resolve such ambiguities since the ground truth representations of c_1 and c_2 are not known. For the sake of exposition, in purely synthetic data, c_1 and c_2 are precisely and it is possible to evaluate the model against a known ground truth.

Data. For a brief summary of the number of varying concepts within a pair and across all pairs considered, refer to Table 9. In the case of synthetic data, we generate $\mathbf{c}$ and $\tilde{\mathbf{c}}$ first to compute $\boldsymbol{\delta}^c := \tilde{\mathbf{c}} - \mathbf{c}$, then apply a dense linear transformation $\mathbf{L}$ to $\boldsymbol{\delta}^c$ to generate $\boldsymbol{\delta}^z$ as $\boldsymbol{\delta}^z = \mathbf{L}\boldsymbol{\delta}^c$. Importantly, towards the generation of $\mathbf{c}$, we generate zero vectors in $\mathbb{R}^{|V|}$ such that for any given sample, S components are perturbed by samples from a uniform distribution and others remain zero. This is

Table 9: Datasets comprise of paired observations $(\mathbf{z}, \tilde{\mathbf{z}})$ where $\mathbf{z}$ and $\tilde{\mathbf{z}}$ vary in concepts $V = \{c_1, c_2, ..., c_{|V|}\}$ across all pairs, such that for any given pair, the maximum number of varying concepts is $\max(|S|)$. *Nomenclature for semi-synthetic datasets follows the rule: identifier of the dataset indicating why we consider it, followed by $|V|$ and max$(|S|)$:* IDENTIFIER*$(|V|$, max$|S|)$.*

Dataset	$\|V\|$	$\max(\|S\|)$
SYNTH$(3, 2)$	3	2
SYNTH$(4, 3)$	4	3
SYNTH$(10, 7)$	10	7

Table 10: The mean **MCC values between the learnt and the ground truth concept vectors are close to** $\mathbf{1}$**.**

	SSAE	**aff**
SYNTH$(3, 2)$	0.999 ± 0.0001	0.873 ± 0.0561
SYNTH$(4, 3)$	0.999 ± 0.0011	0.835 ± 0.0097
SYNTH$(10, 7)$	0.993 ± 0.0005	0.769 ± 0.0103

similar to the data generating process in (Anders et al., 2024) and the conditional distribution of $\boldsymbol{\delta}^c_S$ satisfies Asm. 4 of having a density with respect to Lebesgue.

Results. We estimate $\hat{\boldsymbol{\delta}}^c$ and compare it against $\boldsymbol{\delta}^c$ to verify the degree of identifiability of the learnt concept vectors or encoder representations. Since we have the ground truth here, we compute the MCC between $(\hat{\boldsymbol{\delta}}^c, \boldsymbol{\delta}^c)$ to measure degree of identifiability. Table 10 shows that the proposed method can identify concepts even for higher values of $|V|$ and $\max(|S|)$ against known ground truth data. Synthetic experiments addressing different facets of the identifiability setting we assume can be readily found in prior work on disentangling representations using sparse shifts (Xu et al., 2024; Lachapelle et al., 2023).

Impact Statement

This paper presents technical advancements to a new field of machine learning focused on steering the behaviour of large language models at inference time, i.e., without requiring access to the model's parameters. Steering methods have already begun to play a role in the alignment of LLMs to be e.g., more truthful. We present a new method that could speed up steering research by allowing practitioners to recover steering vectors without the need for supervision, a previous limitation of steering methods. As such, this work could have a positive impact on LLM safety and alignment research. Nevertheless, we flag that contributions towards steering such as ours should be empirically evaluated carefully to avoid over-claiming LLM safety. We acknowledge that while the empirical studies we conduct demonstrate the advantages of identifiable methods such as SSAE for steering, further evaluation is necessary to the method's use in AI safety research.