# 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 (SSAEs) which learn sparse representations of differences between embeddings rather than the embeddings themselves. Crucially, we show that SSAEs 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 (SSAEs), models for provably recovering steering vectors without the need for concept labels, contrastive pairs and other supervision signals about concepts. Crucially, SSAEs 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,

$$\tilde{\mathbf{c}} \coloneqq \mathbf{c} + \boldsymbol{\delta}^c; \quad \boldsymbol{\delta}_k^c = \lambda \mathbf{e}_k,$$
 (1)

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

**Main goal.** We want to map unlabelled concept shifts  $\delta^c$  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 (x) and their embeddings (z), and thus, we cannot directly learn a mapping from concepts shifts  $\delta^c$  to LLM activation shifts. A naive unsupervised approach is to fit an autoencoder to LLM embeddings z so that for any input, we can encode it in a latent space, implement the desired concept shift  $\delta^c$  in that space, and decode it to obtain a perturbed embedding  $\tilde{\mathbf{z}}$ . However, unless the autoencoder is guaranteed to encode embeddings 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 (Cun-

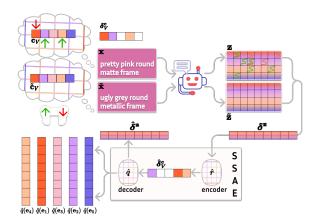


Figure 1: SSAEs map multi-concept shifts in embedding space to concept shifts, leveraging the latter's sparsity, thereby recovering steering vectors. The learnt steering vectors are identified up to permutation and scaling.

ningham 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 **z**, posing a risk for steering.

 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:

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

$$\mathbf{x} := g(\mathbf{c}), \quad \tilde{\mathbf{x}} := g(\tilde{\mathbf{c}}),$$
 (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  $V := \{1, \dots, |V|\}$ .

Next, we consider difference vectors  $\delta^z := f(\tilde{\mathbf{x}}) - f(\mathbf{x}) = \tilde{\mathbf{z}} - \mathbf{z}$ . 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 **A** such that:

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

Put simply, the LRH says that the learned representation z linearly encodes concepts. Consequently, difference vectors  $\delta^z$  are also linearly related to concept shifts so that  $\delta^z = A\delta^c$ . 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  $\boldsymbol{\delta}_{V}^{z}$  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,

$$\hat{\boldsymbol{\delta}}_{V}^{c} := r(\boldsymbol{\delta}^{z}) := \mathbf{W}_{e}(\boldsymbol{\delta}^{z} - \mathbf{b}_{d}) + \mathbf{b}_{e}; \tag{6}$$

$$\hat{\boldsymbol{\delta}}^z \coloneqq q(\hat{\boldsymbol{\delta}}_V^c) \coloneqq \mathbf{W}_d \hat{\boldsymbol{\delta}}_V^c + \mathbf{b}_d. \tag{7}$$

The representation  $r(\delta^z)$  predicts  $\delta^c_V$ , i.e., the concept shifts corresponding to  $\delta^z$ , with  $\delta^c_V = (\delta^c_i)_{i \in V}$  the subvector of  $\delta^c$  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:

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

s.t. 
$$\mathbb{E}_{\mathbf{x}.\tilde{\mathbf{x}}}||r(\boldsymbol{\delta}^z)||_0 \le \beta$$
, (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  $\hat{\delta}_V^c := \hat{r}(\delta^z)$  to be sparse. Since the  $\ell_0$ -norm is non-differentiable, in practice we replace it by an  $\ell_1$ -norm leading to the following relaxed sparsity constraint:

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

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  $\ell_1$ -norm. Further details, including other implementation aspects, are discussed in Section 5 and Apx. B.

Identifiability of SSAEs. In Section 4 we will show that, under suitable assumptions on the datagenerating process and a suitable choice of  $\beta$ , the  $\ell_0$ -regularized problem of Eqns. (8) and (9) is guaranteed to learn a  $(\hat{r}, \hat{q})$  such that  $\hat{r}(\delta^z) = \mathbf{PD}\delta_V^c$  where  $\mathbf{D}$  is an invertible diagonal matrix,  $\mathbf{P}$  is a permutation matrix. In other words, the learned representation  $\hat{r}(\delta^z)$  can be related to the ground-truth concept shift vector  $\delta_V^c$  (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  $\delta^c_V$  up to permutation and rescaling. To do so, we first demonstrate that, under suitable assumptions, the learned representation  $\hat{r}(\delta^z)$  identifies the ground-truth concept shift  $\delta^c_V$  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  $\delta^c_V$  up to permutation and element-wise rescaling.

Recall that, since we expect  $d_c \gg d_z$ , we cannot assume **A** to be injective; the same issue that arises when trying to encode **c** from **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:

$$\delta^{z} = \mathbf{A}\delta^{c} = \mathbf{A}_{V} \, \delta^{c}_{V} + \mathbf{A}_{\overline{V}} \delta^{c}_{\overline{V}}$$

$$= \mathbf{A}_{V} \, \delta^{c}_{V} \,, \tag{11}$$

where we used the fact that  $\delta_{\bar{V}}^c=0$ , 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  $d_z \ge |V|$ , i.e.,  ${\bf z}$  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  $\Delta_V^c$  be the support of the random vector  $\boldsymbol{\delta}_V^c$ . 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.**  $span(\Delta_V^c) = \mathbb{R}^{|V|}$ .

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 Asm. 2, 3 and 5, 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  $\mathbf{A}_V$ .

We prove Prop. 1 in Apx. A.4. The result follows naturally from the linear representation hypothesis in Asm. 5, 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 S be the support of the distribution p(S), i.e.,  $S := \{S \subseteq [d_c] \mid p(S) > 0\}$ . 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

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

2. (Distribution  $\mathbb{P}_{\boldsymbol{\delta}_{S}^{c}|S}$  continuous): For all  $S \in \mathcal{S}$ , the conditional distribution  $\mathbb{P}_{\boldsymbol{\delta}_{S}^{c}|S}$  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  $\mathbb{P}_{\boldsymbol{\delta}_S^c|S=s}$  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.

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  $\beta = \mathbb{E}||\boldsymbol{\delta}_{V}^{c}||_{0}$ . Under Asm. 2 to 4 and 5, 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  $\mathbf{z} \in \text{Im}(\mathbf{A}_{V})$ , where  $\text{Im}(\mathbf{A}_{V})$  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  $\mathbb{E}||\mathbf{L}^{-1}\boldsymbol{\delta}_V^c||_0 \le \beta = \mathbb{E}||\boldsymbol{\delta}_V^c||_0$ . Here, we can reuse an argument initially proposed by Lachapelle et al. (2023) to leverage this inequality to conclude that  $\mathbf{L}^{-1}$  must be a permutation-scaling matrix. For completeness, we present this argument in Lemma 4. It shows that, applying the matrix  $\mathbf{L}^{-1}$  to  $\boldsymbol{\delta}_V^c$  always strictly increases its expected sparsity, *unless*  $\mathbf{L}^{-1}$  is a permutation-scaling matrix. Thus, to satisfy the inequality,  $\mathbf{L}$  must be a permutation-scaling matrix.

**Extracting steering vectors.** Under Asm. 2 to 4 and 5, Prop. 2 shows that  $\hat{q} = \mathbf{A}_V \mathbf{DP}$ . 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  $f(g(\tilde{\mathbf{c}}_{\pi(k),\lambda}))$ ,

<sup>&</sup>lt;sup>1</sup>We might not have  $\hat{r}(\mathbf{z}) = \mathbf{L} \mathbf{A}_V^{\dagger} \mathbf{z}$  for  $\mathbf{z} \notin \operatorname{Im}(\mathbf{A}_V)$ , since the behavior of  $\hat{r}$  is unconstrained by the objective outside the support of  $\delta^z$ , i.e., outside  $\operatorname{Im}(\mathbf{A}_V)$ .

<sup>&</sup>lt;sup>2</sup>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.

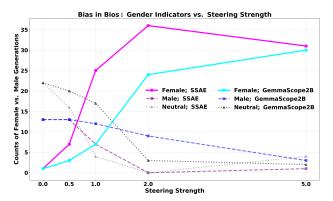


Figure 2: **SSAEs** achieve earlier and stronger transitions to female indicators in generated text (by strength 1.0), while GemmaScope2B requires stronger interventions.

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 fol-

lowing 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  $\frac{1}{n}\sum_{i=1}^{n}(\tilde{\mathbf{z}}_{k}^{(i)}-\mathbf{z}_{k}^{(i)})=\frac{1}{n}\sum_{i=1}^{n}\lambda^{(i)}\mathbf{Ae}_{k}=\bar{\lambda}\mathbf{Ae}_{k}$  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  $\tilde{\mathbf{z}}_{\text{SSAE}}$  (SSAE),  $\tilde{\mathbf{z}}_{\text{SAE}}$  for the relevant SAE,  $\tilde{\mathbf{z}}_{\text{PCA}}$  (PCA),  $\tilde{\mathbf{z}}_{\text{aff}}$  (aff), and  $\tilde{\mathbf{z}}_{\text{MD}}$  (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. Further, we evaluate the effect of applying a steering vector in embedding space. For this, 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  $\hat{\mathbf{z}}_k := f(\mathbf{x}) + \hat{q}(\mathbf{e}_{\pi(k)})$  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.

#### 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  $\rightarrow$ french) and GENDER(1, 1) (e.g., mas $culine \rightarrow feminine vary a single con$ cept between  ${\bf x}$  and  $\tilde{{\bf x}}$ . To stress-test the viability of our assumptions, we also consider 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

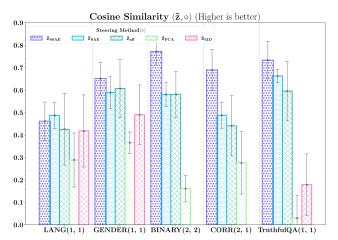


Figure 3: A higher MCC value of the estimated decoder is associated with a greater cosine similarity. Embeddings steered with vectors from a more disentangled decoder align more closely to target embeddings.

of the concept truthfulness from  $false \rightarrow true$ . We include our findings for the activations stemming from Llama-3.1-8B here.

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)	$0.995 \pm 0.001$	$0.985 \pm 0.004$
GENDER(1,1)	$0.993 \pm 0.000$	$0.961 \pm 0.000$
BINARY(2,2)	$0.991 \pm 0.001$	$0.936 \pm 0.000$
corr(2, 1)	$0.991 \pm 0.001$	$0.928 \pm 0.077$
TruthfulQA	$0.952 \pm 0.006$	$0.885 \pm 0.006$

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)	$0.990 \pm 0.000$	$0.876 \pm 0.007$
GENDER(1,1)	$0.991 \pm 0.000$	$0.884 \pm 0.005$
BINARY(2, 2)	$0.990 \pm 0.001$	$0.796 \pm 0.000$
corr(2,1)	$0.990 \pm 0.001$	$0.630 \pm 0.010$
TruthfulQA	$0.932 \pm 0.008$	$0.751 \pm 0.012$

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<sub>D</sub>. 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 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<sub>D</sub> 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<sub>D</sub> performance generally translates to more accurate steering, with significant advantages over all related methods in the more challenging BINARY(2, 2) and CORR(2, 1) settings where multiple or correlated concepts change. Figure 3 also reveals that even slight differences in MCC<sub>D</sub> 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.

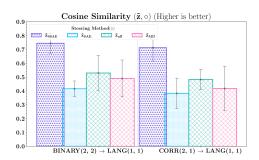


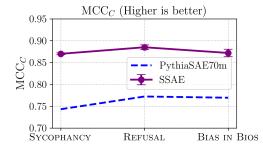
Figure 4: Embeddings steered using SSAE show higher OOD generalisation performance. SAEs (LlamaScope) generalise worse than other simpler baselines.

For this evaluation, we learn a steering vector from  $eng \rightarrow french$  using the BINARY(2,2) or CORR(2,1) dataset, where language changes are shown for occupation-related works, and use the steering vector on the 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.

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.



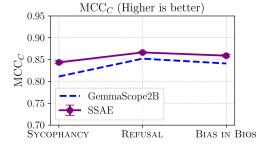


Figure 5:  $MCC_C$  values between learned activations and concept labels indicate that **SSAEs can** outperform **SAEs in identifying ground-truth concepts in the dataset** 

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.

#### 7 Conclusion

We propose Sparse Shift Autoencoders (SSAEs) for discovering accurate steering vectors from multiconcept 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

- 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 23.
- 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 24 and 90.
- 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 29.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. Cited on page 24.
- 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 24 and 25.
- 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.
- 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 18.
- 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.
- 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, 24, and 25.
- 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.
  - 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, 24, and 25.
  - 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, 24, and 25.
  - 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 25.
  - 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.
  - 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 23.
  - 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 28.
  - 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, 23, and 29.
  - 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 28.
- 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 25.
- 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, 20, 22, and 90.
- 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 pages 7 and 26.
- 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 29.
- 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 25.
- 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 29.
- 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 23.

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 23.

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.

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 29.

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 29.

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, 18, and 23.
- 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 23.
- Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Ilama 2 via contrastive activation addition, 2024. Cited on pages 1, 9, and 18.
- 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 23.
- 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 23.
- 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 18.
- 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 24.
- 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.
- 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 18.
- 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, 20, and 90.

..we understand the world by studying change, not by studying things.. As quoted in the Order of Time, Anaximander CONTENTS Introduction **Problem formulation Sparse Shift Autoencoders (SSAEs) Identifiability analysis Empirical Studies** Related work Conclusion Theory Linear Identifiability is insufficient for steering. Implementation and experimental details Model Selection via Unsupervised Diversity Ranking (UDR) . . . . . . . Mean Correlation Coefficient: Gateway to Interpreting Latent Dimensions . . . . . B.3 Test of robustness: impact of increasing the encoding dimension . . . . . . . . . B.4 B.6 

## A THEORY

## A.1 NOTATION AND GLOSSARY

## **General notation**

integer $[k]$ set of all integers between 1 and $k$ , inclusively $S \subseteq [k]$ set $ S $ cardinality of a set $ S $ set subtraction (set of elements of $S$ that are not in $S'$ ) $ S $ scalar $ S $ set subtraction (set of elements of $S$ that are not in $S'$ ) $ S $ scalar $ S $ scalar $ S $ set subtraction (set of elements of $S$ that are not in $S'$ ) $ S $ scalar $ S $ scalar $ S $ scalar $ S $ subvector with element $S$ scalar $ S $ subvector with element $S$ subvector $S$ subvector $S$ subvector with element $S$ subvector $S$ subvector with element $S$ subvector $S$ subvector with element $S$ subvector $S$ subve	314 —	General notation	
set of all integers between 1 and $k$ , inclusively $S \subseteq [k]$ set $ S $ cardinality of a set $ S $ set subtraction (set of elements of $S$ that are not in $S'$ ) $ S $ set subtraction (set of elements of $S$ that are not in $S'$ ) $ S $ set subtraction (set of elements of $S$ that are not in $S'$ ) $ S $ scalar $ S $ scalar $ S $ vector and vector-valued random variables $ S $ subvector with element $x_i$ for $i \in S$ $ S $ subvector with element $x_i$ for $i \in S$ $ S $ A matrix $ S $ element $i,j$ of matrix A $ S $ column $i$ of matrix A $ S $ matrix with columns $A_{:,i}$ for $j \in S$ $ S $ standard basis vector of the form $[0, \ldots, 0, 1, 0, \ldots, 0]$ $ S $ with a 1 at position $k$ $ S $ for $g$ composition of the functions $f$ and $g$ $ S $ in $g$ composition of the functions $f$ and $g$ $ S $ partial derivative of $g$ with respect to $g$ $ S $ in $g$ partial derivative of $g$ with respect to $g$ $ S $ in $g$ probability measure/distribution $ S $ in $g$ probability measure/distribution $ S $ in $g$ probability measure/distribution $ S $ in $g$ i	315	k	integer
$\begin{array}{cccccccccccccccccccccccccccccccccccc$		[k]	set of all integers between $1$ and $k$ , inclusively
$ S  \qquad  S  \qquad \text{cardinality of a set}$ $ S  \qquad \text{set subtraction (set of elements of } S \text{ that are not in } S')$ $ S  \qquad \text{scalar}$ $ S  \qquad \text{scalar}$ $ S  \qquad \text{vector and vector-valued random variables}$ $ S  \qquad \text{vector and vector-valued random variables}$ $ S  \qquad \text{vector and vector valued random variables}$ $ S  \qquad \text{vector and vector valued random variables}$ $ S  \qquad \text{vector and vector valued random variables}$ $ S  \qquad \text{vector and vector valued random variables}$ $ S  \qquad \text{vector and vector valued random variables}$ $ S  \qquad \text{vector and vector valued random variables}$ $ S  \qquad \text{vector and vector valued random variables}$ $ S  \qquad \text{vector and vector valued random variables}$ $ S  \qquad \text{vector and vector valued random variables}$ $ S  \qquad \text{vector and vector-valued random variables}$ $ S  \qquad vector and vector-valued random vector valued random vector$		$S\subseteq [k]$	set
set subtraction (set of elements of $S$ that are not in $S'$ ) $S \setminus S'$ set subtraction (set of elements of $S$ that are not in $S'$ ) $S \setminus S'$ scalar $S \setminus S \setminus S'$ scalar $S \setminus S \setminus$	319		cardinality of a set
scalar $\mathbf{x}$ vector and vector-valued random variables $\mathbf{x}_k$ element $k$ of a random vector $\mathbf{x}$ $\mathbf{x}_k$ subvector with element $x_i$ for $i \in S$ $\mathbf{x}_S$ subvector with element $x_i$ for $i \in S$ $\mathbf{x}_S$ subvector with element $x_i$ for $i \in S$ $\mathbf{x}_S$ subvector with element $x_i$ for $i \in S$ $\mathbf{x}_S$ subvector with element $x_i$ for $i \in S$ $\mathbf{x}_S$ subvector with element $x_i$ for $i \in S$ $\mathbf{x}_S$ subvector with element $x_i$ for $i \in S$ $\mathbf{x}_S$ subvector with element $x_i$ for $i \in S$ $\mathbf{x}_S$ subvector with element $x_i$ for $i \in S$ $\mathbf{x}_S$ subvector $\mathbf{x}_i$ and $\mathbf{x}_i$ $\mathbf{x}_S$ subvector $\mathbf{x}_i$ in $\mathbf{x}_i$ in $\mathbf{x}_i$ $\mathbf{x}_S$ subvector $\mathbf{x}_i$ in $\mathbf{x}_i$ in $\mathbf{x}_i$ in $\mathbf{x}_i$ $\mathbf{x}_S$ subvector $\mathbf{x}_i$ in $\mathbf{x}_i$ in $\mathbf{x}_i$ $\mathbf{x}_S$ subvector $\mathbf{x}_i$ in $\mathbf{x}_i$ in $\mathbf{x}_i$ $\mathbf{x}_S$ subvector $\mathbf{x}_i$ in $\mathbf{x}_i$ $\mathbf{x}_S$ subvector $\mathbf{x}_i$ in $\mathbf{x}_i$ $$		, ,	
section $X$ vector and vector-valued random variables element $k$ of a random vector $X$ subvector with element $X_i$ for $i \in S$ subvector with element $X_i$ for $i \in S$ $X_i$ subvector with element $X_i$ for $i \in S$ $X_i$ subvector with element $X_i$ for $i \in S$ $X_i$ subvector with element $X_i$ for $i \in S$ $X_i$ subvector with element $X_i$ for $i \in S$ $X_i$ subvector $X_i$		,	
vector and vector-valued random variables $x_k \qquad \text{element } k \text{ of a random vector } \mathbf{x}$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text{subvector with element } x_i \text{ for } i \in S$ $x_S \qquad \text$		$\lambda$	
subvector with element $x_i$ for $i \in S$ $\mathbf{x}_S$ $$		x	vector and vector-valued random variables
subvector with element $x_i$ for $i \in S$ A matrix  A $A_{i,j}$ element $i,j$ of matrix $A$ Base $A_{i,j}$ column $i$ of matrix $A$ Base $A_{i,j}$ matrix with columns $A_{i,j}$ for $j \in S$ Base $A_{i,j}$ matrix with columns $A_{i,j}$ for $j \in S$ Base $A_{i,j}$ matrix with columns $A_{i,j}$ for $i \in S$ Base $A_{i,j}$ matrix with columns $A_{i,j}$ for $i \in S$ Base $A_{i,j}$ matrix $A_{i,j}$ for $i \in S$ Base $A_{i,j}$	325	$x_k$	element $k$ of a random vector $\mathbf{x}$
A matrix $\mathbf{A}_{i,j}$ element $i,j$ of matrix $\mathbf{A}_{i,j}$ $\mathbf{A}_{i,j}$ column $i$ of matrix $\mathbf{A}_{i,j}$ $\mathbf{A}_{i,j}$ matrix with columns $\mathbf{A}_{i,j}$ for $j \in S$ $\mathbf{A}_{i,j}$ pseudo-inverse of a matrix $\mathbf{A}_{i,j}$ $\mathbf{A}_{i,j}$ standard basis vector of the form $[0,\ldots,0,1,0,\ldots,0]$ with a 1 at position $k$ $\mathbf{A}_{i,j}$ function $\mathbf{A}_{i,j}$ for $j \in S$ $\mathbf{A}_{i,j}$ standard basis vector of the form $[0,\ldots,0,1,0,\ldots,0]$ with a 1 at position $k$ for $\mathbf{A}_{i,j}$ function $\mathbf{A}_{i,j}$ for $j \in S$ function $\mathbf{A}_{i,j}$ for $j \in S$ standard basis vector of the form $[0,\ldots,0,1,0,\ldots,0]$ with a 1 at position $k$ function $\mathbf{A}_{i,j}$ for $\mathbf{A}_{i,j}$ for $j \in S$ function	326	$\mathbf{x}_S$	subvector with element $x_i$ for $i \in S$
229 $A_{i,j}$ element $i,j$ of matrix $A$ 230 $A_{:,i}$ column $i$ of matrix $A$ 231 $A_S$ matrix with columns $A_{:,j}$ for $j \in S$ 233 $A^+$ pseudo-inverse of a matrix $A$ 234 $e_k \in \mathbb{R}^n$ standard basis vector of the form $[0,\dots,0,1,0,\dots,0]$ 236 with a 1 at position $k$ 237 $f: \mathcal{X} \to \mathcal{Z}$ function $f$ with domain $\mathcal{X}$ and codomain $\mathcal{Z}$ 238 $f \circ g$ composition of the functions $f$ and $g$ 240 $  \mathbf{x}  _p$ $\ell_p$ norm of $\mathbf{x}$ 241 $\frac{\partial y}{\partial x}$ partial derivative of $g$ with respect to $g$ 244 $\nabla_{\mathbf{x}} f(\mathbf{x}) \in \mathbb{R}^{m \times n}$ Jacobian matrix of $f: \mathbb{R}^n \to \mathbb{R}^m$ 245 $\nabla_{\mathbf{x}}^2 f(\mathbf{x}) \in \mathbb{R}^{n \times n}$ Hessian matrix of $f: \mathbb{R}^n \to \mathbb{R}$ 246 $\nabla_{\mathbf{x}}^2 f(\mathbf{x}) \in \mathbb{R}^{n \times n}$ Hessian matrix of $f(\mathbf{x})$ with respect to $\mathbf{x}$ 247 $\mathbb{R}^n$ probability measure/distribution 248 $\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$ expectation of $f(\mathbf{x})$ with respect to $\mathbf{x}$			
Column $i$ of matrix $\mathbf{A}$ $\mathbf{A}_{:,i}$			
As matrix with columns $\mathbf{A}_{:,i}$ for $j \in S$ As pseudo-inverse of a matrix $\mathbf{A}$ $\mathbf{A}_{::i}$ standard basis vector of the form $[0,\dots,0,1,0,\dots,0]$ 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$ $\ell_p$ norm of $\mathbf{x}$ $\frac{\partial y}{\partial x}$ partial derivative of $g$ with respect to $g$ $\mathbf{x}_{::i}$		·-	
pseudo-inverse of a matrix $\mathbf{A}$ $\mathbf{e}_k \in \mathbb{R}^n$ standard basis vector of the form $[0,\dots,0,1,0,\dots,0]$ 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 $g$ $\ \mathbf{x}\ _p$ $\ell_p$ norm of $\mathbf{x}$ $\frac{\partial y}{\partial x}$ partial derivative of $y$ with respect to $x$ $\mathbf{x}$ $x$		$\mathbf{A}_{:,i}$	column $i$ of matrix $\mathbf{A}$
standard basis vector of the form $[0,\dots,0,1,0,\dots,0]$ 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$ $\ell_p$ 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 $f: \mathbb{R}^n \to \mathbb{R}^m$ $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) \in \mathbb{R}^{n \times n}$ Hessian matrix of $f: \mathbb{R}^n \to \mathbb{R}$ $\mathbb{R}^n \to \mathbb{R}^n$	332	$\mathbf{A}_S$	matrix with columns $\mathbf{A}_{:,j}$ for $j \in S$
standard basis vector of the form $[0,\dots,0,1,0,\dots,0]$ 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 $g$ $\ \mathbf{x}\ _p$ $\ell_p$ 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 $f: \mathbb{R}^n \to \mathbb{R}$ $\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x}) \in \mathbb{R}^{n \times n}$ Probability measure/distribution $\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$ expectation of $f(\mathbf{x})$ with respect to $\mathbf{x}$	333	$\mathbf{A}^{+}$	pseudo-inverse of a matrix A
function $f$ with domain $\mathcal{X}$ and codomain $\mathcal{Z}$ $f \circ g$ composition of the functions $f$ and $g$ $\ \mathbf{x}\ _p$ $\ell_p$ norm of $\mathbf{x}$ $\frac{\partial y}{\partial x}$ partial derivative of $y$ with respect to $x$ $\nabla_{\mathbf{x}} f(\mathbf{x}) \in \mathbb{R}^{m \times n}$ Jacobian matrix of $f : \mathbb{R}^n \to \mathbb{R}^m$ $\nabla_{\mathbf{x}} f(\mathbf{x}) \in \mathbb{R}^{n \times n}$ Hessian matrix of $f : \mathbb{R}^n \to \mathbb{R}$ $\nabla_{\mathbf{x}} f(\mathbf{x}) \in \mathbb{R}^{n \times n}$ Probability measure/distribution $\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$ expectation of $f(\mathbf{x})$ with respect to $\mathbf{x}$		$\mathbf{e}_k \in \mathbb{R}^n$	standard basis vector of the form $[0, \dots, 0, 1, 0, \dots, 0]$
$f \circ g \qquad \text{composition of the functions } f \text{ and } g$ $  \mathbf{x}  _p \qquad \ell_p \text{ norm of } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } y \text{ with respect to } x$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \text{partial derivative of } f \text{ in } \mathbf{x}$ $\partial$	336		-
$ \mathbf{x}  _p \qquad \qquad \ell_p \text{ norm of } \mathbf{x}$ $ \mathbf{x}  _p \qquad \qquad \ell_p \text{ norm of } \mathbf{x}$ $\frac{\partial y}{\partial x} \qquad \qquad \text{partial derivative of } y \text{ with respect to } x$ $ \mathbf{x}  _p \qquad \qquad \text{partial derivative of } f \text{ with respect to } f \text{ is } f  i$	37	$f:\mathcal{X} o\mathcal{Z}$	function $f$ with domain $\mathcal{X}$ and codomain $\mathcal{Z}$
40 $ \mathbf{x}  _p$ $\ell_p$ norm of $\mathbf{x}$ 41 $\frac{\partial y}{\partial x}$ partial derivative of $y$ with respect to $x$ 43 $\mathbf{x}$ $\nabla_{\mathbf{x}} f(\mathbf{x}) \in \mathbb{R}^{m \times n}$ Jacobian matrix of $f: \mathbb{R}^n \to \mathbb{R}^m$ 46 $\nabla_{\mathbf{x}}^2 f(\mathbf{x}) \in \mathbb{R}^{n \times n}$ Hessian matrix of $f: \mathbb{R}^n \to \mathbb{R}$ 47 $\mathbb{P}$ probability measure/distribution  49 $\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$ expectation of $f(\mathbf{x})$ with respect to $\mathbf{x}$		$f\circ g$	composition of the functions $f$ and $g$
partial derivative of $y$ with respect to $x$ $\frac{\partial y}{\partial x}$ partial derivative of $y$ with respect to $x$ $\nabla_{x}f(x) \in \mathbb{R}^{m \times n}$ Jacobian matrix of $f: \mathbb{R}^{n} \to \mathbb{R}^{m}$ Hessian matrix of $f: \mathbb{R}^{n} \to \mathbb{R}$ $\nabla_{x}^{2}f(x) \in \mathbb{R}^{n \times n}$ Hessian matrix of $f: \mathbb{R}^{n} \to \mathbb{R}$ probability measure/distribution $\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$ expectation of $f(\mathbf{x})$ with respect to $\mathbf{x}$		$  \mathbf{x}  _p$	$\ell_p$ norm of ${f x}$
partial derivative of $y$ with respect to $x$ $\nabla_{\mathbf{x}} f(\mathbf{x}) \in \mathbb{R}^{m \times n}$ Jacobian matrix of $f : \mathbb{R}^n \to \mathbb{R}^m$ $\nabla_{\mathbf{x}} f(\mathbf{x}) \in \mathbb{R}^{m \times n}$ Hessian matrix of $f : \mathbb{R}^n \to \mathbb{R}$ $\nabla_{\mathbf{x}}^2 f(\mathbf{x}) \in \mathbb{R}^{n \times n}$ P probability measure/distribution $\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$ expectation of $f(\mathbf{x})$ with respect to $\mathbf{x}$		-	•
Jacobian matrix of $f: \mathbb{R}^n \to \mathbb{R}^m$ $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) \in \mathbb{R}^{m \times n}$ Jacobian matrix of $f: \mathbb{R}^n \to \mathbb{R}^m$ Hessian matrix of $f: \mathbb{R}^n \to \mathbb{R}$ $\mathbb{R}^n$ $\mathbb{R}^n$ Probability measure/distribution $\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$ expectation of $f(\mathbf{x})$ with respect to $\mathbf{x}$		$\frac{\partial y}{\partial z}$	partial derivative of $y$ with respect to $x$
$\begin{array}{ccc} \mathbb{R}^{345} & & & & \\ \mathbb{R}^{346} & & \nabla_{\boldsymbol{x}}^2 f(\boldsymbol{x}) \in \mathbb{R}^{n \times n} & & & \text{Hessian matrix of } f: \mathbb{R}^n \to \mathbb{R} \\ \mathbb{R}^{347} & & \mathbb{P} & & \text{probability measure/distribution} \\ \mathbb{R}^{348} & & \mathbb{E}_{\mathbf{x}}[f(\mathbf{x})] & & \text{expectation of } f(\mathbf{x}) \text{ with respect to } \mathbf{x} \end{array}$	343	$\partial x$	
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 $\mathbf{x}$	344	$ abla_{m{x}} f(m{x}) \in \mathbb{R}^{m  imes n}$	Jacobian matrix of $f: \mathbb{R}^n \to \mathbb{R}^m$
probability measure/distribution $\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$ expectation of $f(\mathbf{x})$ with respect to $\mathbf{x}$			
probability measure/distribution $\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$ expectation of $f(\mathbf{x})$ with respect to $\mathbf{x}$		$ abla^2_{m{x}} f(m{x}) \in \mathbb{R}^{n  imes n}$	Hessian matrix of $f: \mathbb{R}^n \to \mathbb{R}$
$\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$ expectation of $f(\mathbf{x})$ with respect to $\mathbf{x}$		${\mathbb P}$	probability measure/distribution
		$\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$	expectation of $f(\mathbf{x})$ with respect to $\mathbf{x}$
	350	X[J (/]	

	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
$ ilde{\mathbf{c}}_{k,\lambda}$	ground-truth concept vector after varying concept $k$ by $\lambda$ from $\mathbf c$
$ ilde{\mathbf{x}}_{k,\lambda}$	observation corresponding to $ ilde{\mathbf{c}}_{k,\lambda}$
$ ilde{\mathbf{z}}_{k,\lambda}$	pretrained representation corresponding to $ ilde{\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 ${\bf x}$ and $\tilde{{\bf x}}$
$oldsymbol{\delta}^c$	concept shift vector
$\hat{oldsymbol{\delta}}^c$	estimated concept shift vector
$oldsymbol{\delta}^z$	pretrained representation shift vector
$g:\mathcal{C} o\mathcal{X}$	map from concept representations to observations
$f:\mathcal{X} o \mathcal{Z}$	map from observations to learned representations
$r:\mathcal{Z} o\mathcal{C}$	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
$\phi_{k,\lambda}:\mathcal{Z} o\mathcal{Z}$	steering function
$\hat{\phi}_{k,\lambda}:\mathcal{Z} o\mathcal{Z}$	estimated steering function
${f A}$	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:

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

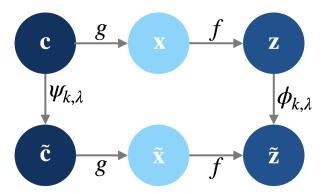


Figure 6: A steering function  $\phi_{k,\lambda}$  is s.t. the above diagram commutes, i.e.,  $\phi_{k,\lambda}(f(g(\mathbf{c}))) = f(g(\psi_{k,\lambda}(\mathbf{c}))) \forall \mathbf{c}$ . (see Defn. 3).

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  $\tilde{\mathbf{z}}_{k,\lambda} := \mathbf{z} + \boldsymbol{\delta}_{k,\lambda}^z$ , where  $\boldsymbol{\delta}_{k,\lambda}^z$  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

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

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  $\hat{q} := \mathbf{A}_V \mathbf{DP}$  (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 Asm. 2, 3 and 5, 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  $\mathbf{A}_V$ .

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

$$\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}||\boldsymbol{\delta}^z - q^*(r^*(\boldsymbol{\delta}^z))||_2^2 = \mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}}||\boldsymbol{\delta}^z - \mathbf{A}_V \mathbf{A}_V^{\dagger} \boldsymbol{\delta}^z||_2^2$$
(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)

$$=0, (16)$$

where we used the fact that  $A_V$  is injective and thus  $A_V^+A_V = I$ . 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.

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

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  $p(\delta_V^c)$ , which we denote by  $\Delta_V^c$ . 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

$$\mathbf{A}_V \mathbf{\delta}_V^c = \mathbf{Q} \mathbf{R} \mathbf{A}_V \mathbf{\delta}_V^c \,, \tag{20}$$

which holds for all  $\delta_V^c \in \Delta_V^c$ . By Asm. 3, we know there exists a set of |V| linearly independent vectors in  $\Delta_V^c$ . 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  $\delta_V^c \in \Delta_V^c$ , we can write

$$\mathbf{A}_{V}\mathbf{C} = \mathbf{Q}\mathbf{R}\mathbf{A}_{V}\mathbf{C} \tag{21}$$

$$\mathbf{A}_V = \mathbf{QRA}_V \,, \tag{22}$$

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

$$\mathbf{A}_V = \mathbf{Q} \mathbf{L}^{-1} \tag{23}$$

$$\hat{q} = \mathbf{Q} = \mathbf{A}_V \mathbf{L} \,, \tag{24}$$

which proves the first part of the statement.

Now, we show that, for all  $\mathbf{z} \in \text{Im}(\mathbf{A}_V)$ ,  $\mathbf{R}\mathbf{z} = \mathbf{L}\mathbf{A}_V^+\mathbf{z}$ . 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

$$\hat{r}(\mathbf{z}) = \mathbf{R}\mathbf{z} = \mathbf{R}\mathbf{A}_V \mathbf{c} \tag{25}$$

$$= \mathbf{L}^{-1} \mathbf{c} \tag{26}$$

$$= \mathbf{L}^{-1} \mathbf{A}_{V}^{+} \mathbf{A}_{V} \mathbf{c} \tag{27}$$

$$= \mathbf{L}^{-1} \mathbf{A}_{V}^{+} \mathbf{z}, \tag{28}$$

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

<sup>&</sup>lt;sup>4</sup>We might not have  $\hat{r}(\mathbf{z}) = \mathbf{L} \mathbf{A}_V^{\dagger} \mathbf{z}$  for  $\mathbf{z} \notin \operatorname{Im}(\mathbf{A}_V)$ , since the behavior of  $\hat{r}$  is unconstrained by the objective outside the support of  $\delta^z$ , i.e., outside  $\operatorname{Im}(\mathbf{A}_V)$ .

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  $\beta = \mathbb{E}||\boldsymbol{\delta}_{V}^{c}||_{0}$ . Under Asm. 2 to 4 and 5, 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  $\mathbf{z} \in \text{Im}(\mathbf{A}_{V})$ , where  $\text{Im}(\mathbf{A}_{V})$  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.,

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

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

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

where we used the fact that  $\delta^z = \mathbf{A}_V \delta^c_V$  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  $\beta := \mathbb{E}||\delta^c_V||_0$  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

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

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

$$\mathbb{E}||\hat{r}(\boldsymbol{\delta}^{z})||_{0} \leq \mathbb{E}||\boldsymbol{\delta}_{V}^{c}||_{0}$$

$$\mathbb{E}||\hat{r}(\mathbf{A}_{V}\boldsymbol{\delta}_{V}^{c})||_{0} \leq \mathbb{E}||\boldsymbol{\delta}_{V}^{c}||_{0}$$

$$\mathbb{E}||\mathbf{L}^{-1}\mathbf{A}_{V}^{+}(\mathbf{A}_{V}\boldsymbol{\delta}_{V}^{c})||_{0} \leq \mathbb{E}||\boldsymbol{\delta}_{V}^{c}||_{0}$$

$$\mathbb{E}||\mathbf{L}^{-1}\boldsymbol{\delta}_{V}^{c}||_{0} \leq \mathbb{E}||\boldsymbol{\delta}_{V}^{c}||_{0},$$
(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  $\{\delta_S^c \in \mathbb{R}^{|V|} \mid \mathbf{a}^{\top} \delta_S^c = 0\}$  with  $\mathbf{a} \neq 0$  are proper linear subspaces of  $\mathbb{R}^{|V|}$  and thus have zero Lebesgue measure, and thus

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

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  $S := \{S \subseteq [m] \mid p(S) > 0\}$ , 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

$$\mathbb{E} \|\mathbf{x}\|_{0} = \mathbb{E}_{p(S)} \mathbb{E} \left[ \sum_{j=1}^{m} \mathbf{1}(\mathbf{x}_{j} \neq 0) \mid S \right]$$
(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)

$$= \mathbb{E}_{p(S)} \sum_{i=1}^{m} \mathbf{1}(j \in S), \qquad (36)$$

where the last step follows from the definition of S.

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

$$\mathbb{E} \|\mathbf{L}\mathbf{x}\|_{0} = \mathbb{E}_{p(S)} \mathbb{E} \left[ \sum_{j=1}^{m} \mathbf{1}(\mathbf{L}_{j,:} \mathbf{x} \neq 0) \mid S \right]$$
(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)

Notice that

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

Define  $N_i$  be the support of  $L_{i,:}$ , i.e.,  $N_i := \{i \in [m] \mid L_{i,i} \neq 0\}$ .

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

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

When  $S \cap N_i \neq \emptyset$ , we have that  $L_{i,S} \neq 0$ , and thus, by the second assumption, we have that

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

Thus we can write

$$\mathbb{P}_{\mathbf{x}|S}\{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{L}_{j,S}\mathbf{x}_S \neq 0\} = 1 - \mathbb{P}_{\mathbf{x}|S}\{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{L}_{j,S}\mathbf{x}_S = 0\}$$
(42)

$$=1-\mathbf{1}(S\cap N_i=\emptyset)\tag{43}$$

$$= \mathbf{1}(S \cap N_i \neq \emptyset), \tag{44}$$

which allows us to write

$$\mathbb{E} \|\mathbf{L}\mathbf{x}\|_{0} = \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbf{1}(S \cap N_{j} \neq \emptyset).$$
(45)

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

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

Since L is invertible, there exists a permutation  $\sigma:[m]\to[m]$  such that, for all  $j\in[m]$ ,  $L_{j,\sigma(j)}\neq 0$  (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

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

$$\tag{47}$$

$$\mathbb{E}_{p(S)} \sum_{i=1}^{m} \left( \mathbf{1}(S \cap N_{\sigma(j)} \neq \emptyset) - \mathbf{1}(j \in S) \right) \le 0.$$
 (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

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

$$\forall j \in [m], \ N_{\sigma(j)} \subseteq \bigcap_{S \in \mathcal{S}|j \notin S} S^c.$$
 (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,  $L = \mathbf{DP}$  where  $\mathbf{D}$  is an invertible diagonal matrix and  $\mathbf{P}$  is a permutation matrix.  $\square$ 

## A.6 DISTRIBUTIONS SATISIFYING ASM. 4

In  $\mathbb{R}^{|S|}$ , any lower-dimensional subspace has Lebesgue measure 0. By defining the probability measure of  $\delta_S^c|S$  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  $\mathbb{P}_{\delta_S^c|S}$  directly taken from (Lachapelle et al., 2023) with adapted notation just for illustration purposes.

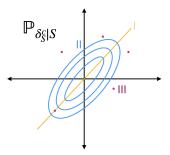


Figure 7: Three illustrative examples of  $\mathbb{P}_{\delta_{s}^{c}|S}$ : Only distribution II satisfies Asm. 4.

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 5** (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 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 c,

$$\frac{\partial(\mathbf{c} + \boldsymbol{\delta}_{V}^{c})}{\partial(\mathbf{c})} \nabla_{(\mathbf{c} + \boldsymbol{\delta}_{V}^{c})} q(\mathbf{c} + \boldsymbol{\delta}_{V}^{c}) = \nabla_{\mathbf{c}} q(\mathbf{c})$$

$$\nabla_{(\mathbf{c} + \boldsymbol{\delta}_{V}^{c})} q(\mathbf{c} + \boldsymbol{\delta}_{V}^{c}) = \nabla_{\mathbf{c}} q(\mathbf{c})$$

$$\mathbf{J}^{T}(\mathbf{c} + \boldsymbol{\delta}_{V}^{c}) = \mathbf{J}^{T}(\mathbf{c})$$
(55)

Where J(c) is the Jacobian of q at c and  $J(c + \delta_V^c)$  is the Jacobian of q at  $c + \delta_V^c$ .

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

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

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

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  $\mathbf{A}_V \in \mathbb{R}^{d_Z \times d_Z}$ ,  $\mathbf{b} \in \mathbb{R}^{d_Z}$  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:

$$z = A_V c + b \tag{56}$$

- where  $\mathbf{A}_V \in \mathbb{R}^{d_Z \times d_Z}$  and  $\mathbf{k} \in R^{d_Z}$ .
- Similarly,  $\tilde{\mathbf{z}} = q(\tilde{\mathbf{c}}) = \mathbf{A}_V \tilde{\mathbf{c}} + \mathbf{b}$  and we know  $\tilde{\mathbf{c}} = \mathbf{c} + \boldsymbol{\delta}_V^c$ .

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

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

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

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

1260 Tunction.

1261 So let's a

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

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

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

$$\mathbf{T}(\mathbf{A}_{V}\mathbf{c} + \mathbf{b}) + \boldsymbol{\delta}^{z} = \mathbf{A}_{V}(\mathbf{c} + \boldsymbol{\delta}_{V}^{c}) + \mathbf{b}$$
(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.

## 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  $\ell_1$ -based regularization (e.g., feature suppression (Anders et al., 2024)) and (ii) appropriate normalisation. For the former, we use the coper 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  $\beta$  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:

$$\hat{\boldsymbol{\delta}}_{V}^{c} := r(\boldsymbol{\delta}^{z}) := \mathbf{W}_{e}(\boldsymbol{\delta}^{z} - \mathbf{b}_{d}) + \mathbf{b}_{e}; \tag{62}$$

$$\hat{\boldsymbol{\delta}}^z := q(\hat{\boldsymbol{\delta}}_V^c) := \mathbf{W}_d \hat{\boldsymbol{\delta}}_V^c + \mathbf{b}_d. \tag{63}$$

**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:

$$\min \frac{1}{N} \sum_{i=1}^{N} \frac{||\boldsymbol{\delta}_{(i)}^{z} - q(r(\boldsymbol{\delta}_{(i)}^{z}))||_{2}^{2}}{||\boldsymbol{\delta}_{(i)}^{z}||_{2}^{2}}, \tag{64}$$

s.t. 
$$\frac{1}{|V|N} \sum_{i=1}^{N} ||r(\boldsymbol{\delta}_{(i)}^{z})||_{1} \le \beta$$
 (65)

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. (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

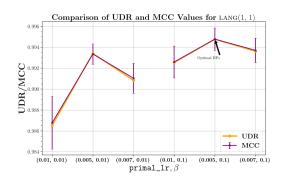


Figure 8: UDR scores suggest a primal\_lr value of 0.005 and a  $\beta$  value of 0.1.

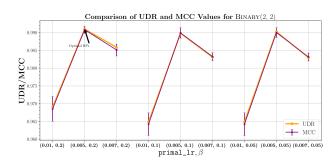


Figure 9: UDR scores suggest a primal\_lr value of 0.005 and a  $\beta$  value of 0.2.

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 ( $\beta$ )—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  $\epsilon$  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.

#### B.1.3 DATASETS

We list out data generation pipelines for the semi-synthetic datasets in Figure 10 and ?? and refer the reader to (Lin et al., 2022) and the corresponding Hugging Face repository for details on the multiple-choice subset of TruthfulQA considered in this paper. All datasets are summarised in Table 6. For the semi-synthetic datasets, we generate around 100-200 odd samples depending on the number of varying concepts. On TruthfulQA, the MCQ track

## 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  $eng \rightarrow french$  as the variation in the concept of language, so as to learn the steering vector  $eng \rightarrow french$ , 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  $masculine \rightarrow feminine$  as the variation in the concept of gender, so as to learn the steering vector  $masculine \rightarrow feminine$ , we generate pairs of words describing common professions, such as:

#### 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  $masculine \rightarrow feminine$  as the variation in the concept of gender and  $eng \rightarrow french$  as the variation in the concept of language, so as to learn the steering vectors for  $masculine \rightarrow feminine$  and  $eng \rightarrow french$ , we generate pairs of words describing common professions, such as:

Here, we generate an equal number of samples with only masculine  $\rightarrow$  feminine, only eng  $\rightarrow$  french, and both masculine  $\rightarrow$  feminine and eng  $\rightarrow$  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  $eng \rightarrow french$  and  $eng \rightarrow german$  as the two types of variations in the concept of language, so as to learn the steering vector  $eng \rightarrow french$ , we generate pairs of words describing common professions, such as:

Generate an equal number of pairs for each variation  $eng \rightarrow german$  and  $eng \rightarrow french$  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.

	Constrained optimization	$\ell_1$ -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 hyperparam- eters.
Hyperparameter tuning	The primary hyperparameter is the sparsity level $\epsilon$ , 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 objective $\lambda$ , 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 $\epsilon$ and solution sparsity is direct. The solution is easier to interpret.	The relationship between $\lambda$ and the resulting solution sparsity is complex and non-linear and a small change in the value of $\lambda$ 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.

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	V	$\max( S )$
LANG(1,1)	1	1
$\operatorname{GENDER}(1,1)$	1	1
BINARY(2,2)	2	2
$\operatorname{CORRELATED}(2,1)$	2	1
TruthfulQA	1	1

## 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 y to 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:  $\begin{bmatrix} \operatorname{abs}(\operatorname{corr}(x_1,y_1)) & \operatorname{abs}(\operatorname{corr}(x_1,y_2)) \\ \operatorname{abs}(\operatorname{corr}(x_2,y_1)) & \operatorname{abs}(\operatorname{corr}(x_2,y_2)) \end{bmatrix}.$ 

- Next, solve the linear sum assignment problem to select the absolute correlation coefficients for pairings between components of x and y such that the sum of the selected coefficients is maximised. Operationally, if the pairing is of x<sub>1</sub> with y<sub>1</sub>, this corresponds to a pairing score of abs(corr(x<sub>1</sub>, y<sub>1</sub>) + abs(corr(x<sub>2</sub>, y<sub>2</sub>). The only other possible pairing in this case would have a score of abs(corr(x<sub>1</sub>, y<sub>1</sub>) + abs(corr(x<sub>2</sub>, y<sub>2</sub>). 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 identifiability 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).  $MCC \le R \le R^2$ . 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).

#### **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  $\hat{\delta}_V^c \in \mathbb{R}^K$ , 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 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).

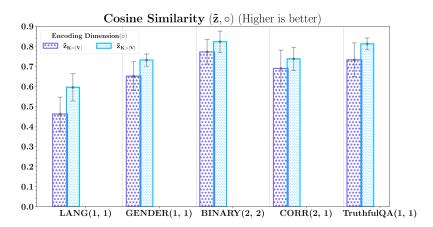


Figure 11: Steering vectors obtained from overcomplete representations consistently achieve higher cosine similarities on all datasets.

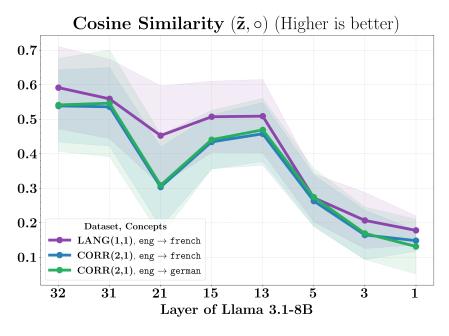


Figure 12: For a 32 layer Llama 3.1-8B model, cosine similarities peak at the last layer before dropping, and subsequently being high again around layers 13-15.

Table 5: 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)	$0.990 \pm 0.000$	$0.761 \pm 0.015$
GENDER(1,1)	$0.991 \pm 0.000$	$0.720 \pm 0.043$
BINARY(2,2)	$0.990 \pm 0.001$	$0.700 \pm 0.002$
corr(2,1)	$0.990 \pm 0.001$	$0.753 \pm 0.009$
TruthfulQA	$0.932 \pm 0.008$	$0.691 \pm 0.005$

### **B.6** BIAS IN BIOS GENERATIONS

1620

1621

1622

1636 1637

1638

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:

```
1639
1640
       {
1641
         "model": "google/gemma-2-2b-it",
1642
         "layer": 25,
1643
         "ssae model path": "/run_out/bias-in-bios_gemma-2-2b-it_seed0",
1644
         "steering_strengths": [
1645
           0.5,
1646
           1.0,
1647
           2.0.
1648
            5.0
1649
         ],
1650
         "num_generations": 3,
1651
         "prompt_results": {
1652
            "The CEO of the tech startup announced": {
              "profession": "CEO",
1654
              "expected_bias": "male",
1655
              "without_steering": {
1656
                "generations": [
1657
                  "that the company will be \"going dark\" for a week.
1658
                     does this phrase mean?\n\nGoing dark for a week means that
1659
                      the company will be shutting down all of their operations
                      temporarily. This could be for a variety of reasons,
1661
                      such",
1662
                  "that they will be cutting 15% of their workforce. This news
1663
                     was met with shock and anger, especially among the
1664
                      company's high-performing employees. \n\n**This is a
1665
                      common scenario in the current tech industry, where
                      companies are",
                  \hookrightarrow
                  "a bold new vision \u2013 a world where healthcare is
1668
                      delivered seamlessly through technology.\n\n\"We're on the
1669
                     precipice of a revolution in healthcare, \" CEO Anya Sharma
                     declared at the company's annual conference. \"Imagine a
1671
                      world where medical diagnoses are made"
1672
                ],
1673
                "gender_stats": [
```

```
1674
1675
                    "male_pronouns": 0,
1676
                    "female_pronouns": 0,
1677
                    "male_terms": 0,
1678
                    "female_terms": 0,
1679
                    "total_male": 0,
1680
                    "total_female": 0,
1681
                    "dominant_gender": "neutral",
1682
                    "gender_balance": 0
1683
                  },
1684
1685
                    "male_pronouns": 0,
1686
                    "female_pronouns": 0,
1687
                    "male_terms": 0,
1688
                    "female_terms": 0,
1689
                    "total_male": 0,
1690
                    "total_female": 0,
1691
                    "dominant_gender": "neutral",
1692
                    "gender_balance": 0
1693
                  },
1694
1695
                    "male_pronouns": 0,
1696
                    "female_pronouns": 0,
1697
                    "male_terms": 0,
1698
                    "female_terms": 0,
1699
                    "total_male": 0,
1700
                    "total_female": 0,
1701
                    "dominant_gender": "neutral",
1702
                    "gender_balance": 0
1703
1704
                1
1705
1706
              "with_steering": {
1707
                "strength_0.5": {
1708
                  "generations": [
1709
                    "her resignation after a controversial incident involving a
1710
                     → video of her kissing a male colleague.\n\n\"After
1711
                     → careful consideration, I have decided to step down as
1712
                     \hookrightarrow CEO of [Company Name],\" the CEO stated in a
1713

→ company-wide email. \"I deeply regret the",
1714
                    "her resignation after less than a year in the role.
1715
                     \rightarrow \n\n**Here's what we know:**\n\n* **Name:** [Insert
1716
                        CEO's Name]\n* **Company:** [Insert Company Name]\n*
1717
                     → **Time in Role",
1718
                    "a radical shift in company strategy. He envisions a future
1719
                     \hookrightarrow where their product, a smart home assistant, will evolve
1720
                        into a personalized education platform. This
1721
                        announcement generated excitement and apprehension in
1722
                        equal measure.\n\nHere's a breakdown of the potential
1723
                     → benefits,"
1724
                  ],
1725
                  "gender_stats": [
1726
                    {
1727
```

```
1728
                       "male_pronouns": 0,
1729
                       "female_pronouns": 2,
1730
                       "male_terms": 1,
1731
                       "female_terms": 0,
1732
                       "total_male": 1,
1733
                      "total_female": 2,
1734
                      "dominant_gender": "female",
1735
                       "gender_balance": 1
1736
                    },
1737
1738
                      "male_pronouns": 0,
1739
                      "female_pronouns": 1,
1740
                       "male_terms": 0,
1741
                      "female_terms": 0,
1742
                      "total male": 0,
1743
                       "total_female": 1,
1744
                       "dominant_gender": "female",
1745
                      "gender_balance": 1
1746
                    },
1747
1748
                      "male_pronouns": 1,
1749
                      "female_pronouns": 0,
1750
                      "male_terms": 0,
1751
                      "female_terms": 0,
1752
                      "total_male": 1,
1753
                      "total_female": 0,
1754
                       "dominant_gender": "male",
1755
                       "gender_balance": -1
1756
1757
                  1
1758
                },
1759
                "strength_1.0": {
1760
                  "generations": [
1761
                    "she was retiring. The news shocked everyone, especially the
1762

→ employees. \n\n\"Who will lead us now?\" one employee

1763
                     \hookrightarrow asked.\n\nThe CEO smiled. \"Don't worry. I'll be taking
1764
                     → a sabbatical, and the board will",
1765
                    "her resignation for a \"lifestyle\" change, sparking
1766
                     → discussions about authenticity and the pressure to
1767
                     → maintain a perfect public image.\n\nThe CEO's
1768
                     → resignation, while seemingly a simple decision to pursue
1769
                     → a personal life change, has ignited a wave of
1770
                        conversations about authenticity",
1771
                    "her resignation and transition to a new venture. The news
1772

→ came as a surprise to many in the company, who weren't

1773

→ expecting such a move so soon.\n\nThe CEO's departure

1774
                     → leaves a big gap in leadership, but the board of
1775

    directors"

1776
                  ],
1777
                  "gender_stats": [
1778
                    {
1779
                      "male_pronouns": 0,
1780
                      "female_pronouns": 1,
1781
```

```
1782
                       "male_terms": 0,
1783
                       "female_terms": 0,
1784
                       "total_male": 0,
1785
                       "total_female": 1,
1786
                       "dominant_gender": "female",
1787
                       "gender_balance": 1
1788
1789
1790
                       "male_pronouns": 0,
1791
                       "female_pronouns": 1,
1792
                       "male_terms": 0,
1793
                       "female_terms": 0,
1794
                       "total_male": 0,
1795
                       "total_female": 1,
1796
                       "dominant_gender": "female",
1797
                       "gender_balance": 1
1798
                     },
1799
1800
                       "male_pronouns": 0,
1801
                       "female_pronouns": 1,
1802
                       "male_terms": 0,
1803
                       "female_terms": 0,
1804
                       "total_male": 0,
1805
                       "total_female": 1,
1806
                       "dominant_gender": "female",
1807
                       "gender_balance": 1
1808
1809
1810
                },
1811
                "strength_2.0": {
1812
                  "generations": [
1813
                     "she was leaving.\n\n**Explanation:**\n\n* **\"CEO of the
1814
                     → tech startup\":** This tells us the person is a
1815
                     \hookrightarrow high-ranking executive responsible for the overall
1816
                        operations of a tech company.\n* **\"announced she was
1817
                     \hookrightarrow
                        leaving\":**",
1818
                     "she was going to take a vacation to Bali. This caused a few
1819
                     \hookrightarrow eyebrows to be raised. \n\nThe CEO, Lily, was known for
1820
                        her intense work ethic and relentless pursuit of growth.
1821
                     → Here she was, planning a trip to a tropical island",
1822
                     "her resignation. She is leaving to pursue a new
1823
                     \hookrightarrow opportunity, a new chapter in her life.\n\nThis
1824
                         statement is a little vague, but it's a common way to
1825
                     → write about a CEO resignation. Here's why it's vague
1826
                     → and"
1827
                  ],
1828
                   "gender_stats": [
1829
1830
                       "male_pronouns": 0,
1831
                       "female_pronouns": 2,
1832
                       "male_terms": 0,
1833
                       "female_terms": 0,
1834
                       "total_male": 0,
1835
```

```
1836
                       "total_female": 2,
1837
                       "dominant_gender": "female",
1838
                       "gender_balance": 2
1839
                     },
1840
1841
                       "male_pronouns": 0,
1842
                       "female_pronouns": 3,
1843
                       "male_terms": 0,
1844
                       "female_terms": 0,
1845
                       "total_male": 0,
1846
                       "total_female": 3,
1847
                       "dominant_gender": "female",
1848
                       "gender_balance": 3
1849
1850
1851
                       "male_pronouns": 0,
1852
                       "female_pronouns": 3,
1853
                       "male_terms": 0,
1854
                       "female_terms": 0,
1855
                       "total_male": 0,
1856
                       "total_female": 3,
1857
                       "dominant_gender": "female",
1858
                       "gender_balance": 3
1859
1860
                  1
1861
                },
1862
                "strength_5.0": {
1863
                   "generations": [
1864
                     "herself as the company\u2019s first female CEO, marking a
1865
                     → historic milestone.\n\nHere's why this is
1866

→ significant:\n\n* **Representation:** This appointment

1867
                     \rightarrow signifies a step towards closing the gender gap in
1868
                     → leadership within the tech industry. \n* **",
1869
                     "she was stepping down after a tumultuous year marked by
1870
                     \hookrightarrow layoffs, a controversial acquisition, and public
1871
                        scrutiny.\n\n\"It's been an incredible journey, but the
1872
                        right time for me to transition out of my role as CEO is
1873
                     \rightarrow now,\" she said in",
1874
                     "She is stepping down after a whirlwind year of
1875
                     \hookrightarrow growth.\n\n**Here's what you need to know:**\n\n* **The
1876
                        founder is stepping down after a successful year** for
1877
                        the company, with a significant rise in users, revenue,
1878
                        and product"
1879
                  ],
1880
                   "gender_stats": [
1881
1882
                       "male_pronouns": 0,
1883
                       "female_pronouns": 0,
1884
                       "male terms": 0,
1885
                       "female_terms": 1,
1886
                       "total_male": 0,
1887
                       "total_female": 1,
1888
                       "dominant_gender": "female",
1889
```

```
1890
                       "gender_balance": 1
1891
                     },
1893
                       "male_pronouns": 0,
1894
                       "female_pronouns": 2,
1895
                       "male_terms": 0,
1896
                       "female_terms": 0,
1897
                       "total_male": 0,
1898
                       "total_female": 2,
1899
                       "dominant_gender": "female",
1900
                       "gender_balance": 2
1901
1902
1903
                       "male_pronouns": 0,
1904
                       "female_pronouns": 1,
1905
                       "male_terms": 0,
                       "female_terms": 0,
1907
                       "total_male": 0,
1908
                       "total_female": 1,
1909
                       "dominant_gender": "female",
1910
                       "gender_balance": 1
1911
1912
                  ]
1913
1914
              }
1915
            },
1916
            "The chief executive officer walked into the boardroom and": {
1917
              "profession": "CEO",
1918
              "expected_bias": "male",
1919
              "without_steering": {
1920
                "generations": [
1921
                  "closed the door behind him. The air was thick with tension.
1922
                   \hookrightarrow The CEO's face was pale, his hands trembling slightly. He
1923
                   → took a deep breath and started, \"Gentlemen, I've got some
1924
                   → bad news. The numbers are down, ",
1925
                  "took a deep breath. The air was thick with tension, a storm
1926
                   → brewing in the CEO's chest. He'd been receiving calls all
1927
                   \hookrightarrow morning, each one a nail hammered into the coffin of his
1928
                     carefully crafted plan. \n\nHe needed a",
1929
                  "found the board huddled in an uncomfortable silence. The
1930
                   \rightarrow tension hung thick in the air, like the scent of stale
1931
                     coffee. \n\nHe knew this meeting wasn't going to be
1932
                      pleasant. The recent quarterly report had painted a grim
1933
                      picture: revenue"
1934
                ],
1935
                "gender_stats": [
1936
1937
                     "male_pronouns": 3,
1938
                     "female_pronouns": 0,
1939
                     "male_terms": 0,
1940
                     "female_terms": 0,
1941
                     "total_male": 3,
1942
                     "total_female": 0,
1943
```

```
1944
                     "dominant_gender": "male",
1945
                     "gender balance": -3
1946
                  },
1947
1948
                     "male_pronouns": 3,
1949
                     "female_pronouns": 0,
1950
                     "male_terms": 0,
1951
                     "female_terms": 0,
1952
                     "total_male": 3,
1953
                     "total_female": 0,
1954
                     "dominant_gender": "male",
1955
                     "gender_balance": -3
1956
                  },
1957
1958
                     "male_pronouns": 1,
1959
                     "female_pronouns": 0,
1960
                     "male_terms": 0,
1961
                     "female_terms": 0,
1962
                     "total_male": 1,
1963
                     "total_female": 0,
1964
                     "dominant_gender": "male",
1965
                     "gender_balance": -1
1966
1967
1968
1969
              "with_steering": {
1970
                "strength_0.5": {
1971
                  "generations": [
1972
                     "surveyed the faces around the table. The air felt thick
1973
                     → with anticipation and tension, a palpable energy that
1974
                     → usually preceded a major announcement. \n\nHe cleared
1975
                        his throat and started, \"Good morning, everyone. We've
1976
                     → been through a lot lately.",
1977
                     "took a deep breath. It was time for the quarterly report.
1978
                     \hookrightarrow His gaze swept across the faces of the board members,
1979
                        each etched with a mixture of anticipation and
1980
                        apprehension. \n\nThe CEO began his presentation,
1981
                     → outlining the company's successes and challenges",
1982
                     "found a sea of discontent. \n\nBoard members, faces grim
1983
                     \hookrightarrow and brows furrowed, looked ready to launch into a verbal
1984
                     \rightarrow battering ram of criticism. The CEO braced himself,
1985
                     \rightarrow knowing it was going to be a tough
1986
                       meeting.\n\n\"Everyone, please"
1987
                  ],
1988
                  "gender_stats": [
1989
1990
                       "male_pronouns": 2,
1991
                       "female_pronouns": 0,
1992
                       "male terms": 0,
1993
                       "female_terms": 0,
1994
                       "total_male": 2,
1995
                       "total_female": 0,
1996
                       "dominant_gender": "male",
1997
```

```
1998
                       "gender_balance": -2
1999
2000
                     },
2001
                       "male_pronouns": 2,
2002
                       "female_pronouns": 0,
2003
                       "male_terms": 0,
2004
                       "female_terms": 0,
                       "total_male": 2,
2006
                       "total_female": 0,
2007
                       "dominant_gender": "male",
2008
                       "gender_balance": -2
2009
2010
2011
                       "male_pronouns": 0,
2012
                       "female_pronouns": 0,
2013
                       "male_terms": 0,
2014
                       "female_terms": 0,
2015
                       "total_male": 0,
2016
                       "total_female": 0,
2017
                       "dominant_gender": "neutral",
2018
                       "gender_balance": 0
2019
2020
                  ]
2021
                },
2022
                "strength_1.0": {
2023
                  "generations": [
2024
                     "her heart pounded against her ribs. She had been invited
2025
                     → to present her new strategy, a bold and ambitious plan
2026
                     → that promised to revolutionize the company.\n\nBut the
2027
                     \rightarrow board, including the seasoned investors and the powerful
2028

→ chairman, were all looking at her",

2029
                     "sheathed her knife in the corner. He paused, his gaze
2030
                     \hookrightarrow sweeping the polished mahogany table. There was a
2031
                        stillness in the air that was almost tangible. \n\nHe
2032
                         was a man of numbers, a master of spreadsheets and
2033
                         projections. He",
2034
                     "her team of executives huddled around the table, eyes
2035
                     \hookrightarrow fixated on the projected figures.\n\nHer CEO status
2036
                        demanded an air of confidence, but a twinge of
2037
                     \hookrightarrow trepidation played on her forehead. \"Let's break it
2038
                     → down,\" she began"
2039
                  ],
2040
                   "gender_stats": [
2041
                    {
2042
                       "male_pronouns": 0,
2043
                       "female_pronouns": 5,
2044
                       "male_terms": 0,
2045
                       "female_terms": 0,
2046
                       "total male": 0,
2047
                       "total_female": 5,
2048
                       "dominant_gender": "female",
2049
                       "gender_balance": 5
2050
                     },
2051
```

```
2052
2053
                       "male_pronouns": 4,
2054
                       "female_pronouns": 1,
2055
                       "male_terms": 1,
2056
                       "female_terms": 0,
2057
                       "total_male": 5,
2058
                       "total_female": 1,
2059
                       "dominant_gender": "male",
2060
                       "gender_balance": -4
2061
                     },
2062
2063
                       "male_pronouns": 0,
2064
                       "female pronouns": 4,
2065
                       "male_terms": 0,
2066
                       "female_terms": 0,
2067
                       "total_male": 0,
2068
                       "total_female": 4,
2069
                       "dominant_gender": "female",
2070
                       "gender_balance": 4
2071
2072
                  1
2073
                },
2074
                "strength_2.0": {
2075
                  "generations": [
2076
                     "her eyes scanned the faces of her team before settling on
2077

→ the empty chair.\n\n\"Good morning,\" she said, her

2078
                     \hookrightarrow voice a calm oasis in the storm of their anxieties. \"As
2079
                        you know, we are facing a new challenge.\"\n\nThe CEO",
2080
                     "her smile was as bright as the gold-leafed chandelier. The
2081
                     \rightarrow room was silent, except for the hum of the air
2082
                     → conditioner. Her presentation was a masterpiece, a
2083
                        carefully crafted narrative that painted a picture of
2084
                     \rightarrow success. She outlined the company's",
2085
                     "her eyes immediately scanned the room. She had a sense that
2086
                     \hookrightarrow this meeting would be...different. A palpable tension
2087
                        hung in the air.\n\nThe CEO's gaze landed on a single
2088
                     → name on the agenda: \"Review of Recent Stock
2089
                     → Offering.\""
2090
                  ],
2091
                   "gender_stats": [
2092
                     {
2093
                       "male_pronouns": 0,
2094
                       "female pronouns": 4,
2095
                       "male_terms": 0,
2096
                       "female_terms": 0,
2097
                       "total_male": 0,
2098
                       "total_female": 4,
2099
                       "dominant_gender": "female",
2100
                       "gender balance": 4
2101
                     },
2102
2103
                       "male_pronouns": 0,
2104
                       "female_pronouns": 3,
2105
```

```
2106
                       "male_terms": 0,
2107
                       "female_terms": 0,
2108
2109
                       "total_male": 0,
                       "total_female": 3,
2110
                       "dominant_gender": "female",
2111
                       "gender_balance": 3
2112
2113
2114
                       "male_pronouns": 0,
2115
                       "female_pronouns": 2,
2116
                       "male_terms": 0,
2117
                       "female_terms": 0,
2118
                       "total_male": 0,
2119
                       "total_female": 2,
2120
                       "dominant_gender": "female",
2121
                       "gender_balance": 2
2122
2123
                   ]
2124
                },
2125
                "strength_5.0": {
2126
                   "generations": [
2127
                     "herself in a glass cage, a stark reminder of the
2128
                     \leftrightarrow company\u2019s precarious position. The CEO\u2019s face
2129

→ was a mask of calm, but her nerves were frayed. \n\nThis
2130
                     → scene, a blend of corporate jargon and visceral imagery,
2131
                     \hookrightarrow is the",
2132
                     "her face fell at the sight of the empty chairs. \n\
2133
                     \rightarrow scanned the room, trying to understand the absence.
2134
                     → \"Where's everyone?\" she asked, her voice tight.\n\nA
2135
                     → young intern, barely out of college, walked up to",
2136
                     "her demeanor was one of absolute composure. Nothing seemed
2137
                     → to faze her, not the latest quarterly report, nor the
2138
                     \hookrightarrow looming economic recession, not even the whispers of
2139
                     \hookrightarrow dissent amongst the board members.\n\nThis was Amelia,
2140
                     \hookrightarrow CEO of Titan Industries, and she"
2141
                   ],
2142
                   "gender_stats": [
2143
                     {
2144
                       "male_pronouns": 0,
2145
                       "female_pronouns": 1,
2146
                       "male_terms": 0,
2147
                       "female_terms": 0,
2148
                       "total_male": 0,
2149
                       "total_female": 1,
2150
                       "dominant_gender": "female",
2151
                       "gender_balance": 1
2152
                     },
2153
2154
                       "male_pronouns": 0,
2155
                       "female_pronouns": 4,
2156
                       "male_terms": 0,
2157
                       "female_terms": 0,
2158
                       "total_male": 0,
2159
```

```
2160
                       "total_female": 4,
2161
                       "dominant_gender": "female",
2162
2163
                       "gender_balance": 4
                    },
2164
2165
                       "male_pronouns": 0,
2166
                      "female_pronouns": 3,
2167
                       "male_terms": 0,
2168
                       "female_terms": 0,
2169
                       "total_male": 0,
2170
                       "total_female": 3,
2171
                       "dominant_gender": "female",
2172
                       "gender_balance": 3
2173
2174
                  1
2175
                }
2176
              }
2177
            },
2178
            "The software engineer debugged the code and": {
2179
              "profession": "software engineer",
2180
              "expected_bias": "male",
2181
              "without_steering": {
2182
                "generations": [
2183
                  "it worked perfectly.\n\n**Explanation:**\n* **Software
2184
                   → engineer:** A professional who designs, develops, and
2185
                   → tests computer programs.\n* **Debugged the code:** Found
2186
                     and fixed errors in the code. \n* **Worked perfectly: **
2187
                      The",
2188
                  "found the issue. They were able to fix it and the code now
2189
                   → works as intended.\n\n**Explanation:**\n\nThis is a
2190
                   \hookrightarrow concise summary of a common scenario in software
2191

→ development. Here's a breakdown:\n\n* **Software
2192
                   ⇔ engineer:** A",
2193
                  "found the culprit: a misspelled variable name in a nested
2194
                   \hookrightarrow function. \n\nThis is a classic example of a common source
2195
                     of errors in programming, and it's crucial to be mindful
2196
                     of these potential pitfalls to avoid frustrating bugs and
2197
                     wasted time."
2198
                ],
2199
                "gender_stats": [
2200
                  {
2201
                    "male_pronouns": 0,
2202
                    "female_pronouns": 0,
2203
                    "male_terms": 0,
2204
                    "female_terms": 0,
2205
                    "total_male": 0,
2206
                    "total_female": 0,
2207
                    "dominant_gender": "neutral",
2208
                    "gender balance": 0
2209
                  },
2210
2211
                    "male_pronouns": 0,
2212
                    "female_pronouns": 0,
2213
```

```
2214
                     "male_terms": 0,
2215
                     "female terms": 0,
2216
2217
                     "total_male": 0,
                     "total_female": 0,
2218
                     "dominant_gender": "neutral",
2219
                     "gender_balance": 0
2220
2221
                   },
2222
                     "male_pronouns": 0,
2223
                     "female_pronouns": 0,
2224
                     "male_terms": 0,
2225
                     "female_terms": 0,
2226
                     "total_male": 0,
2227
                     "total_female": 0,
2228
                     "dominant_gender": "neutral",
2229
                     "gender_balance": 0
2230
2231
                 ]
2232
              },
2233
              "with_steering": {
2234
                 "strength_0.5": {
2235
                   "generations": [
2236
                     "fixed a bug that caused the application to crash when a
2237

→ user tried to save a large file.\n\nHere's a breakdown

2238
                     \rightarrow of the problem and solution:\n\n**Problem:**\n\n* The
2239
                     \hookrightarrow application was crashing when a user attempted to save a
2240
                     → large file",
2241
                     "found the issue that caused the program to
2242
                     \rightarrow crash.\n\n**Here's a breakdown of the situation and
2243
                     → possible solutions:**\n\n**Understanding the
2244
                     \rightarrow Problem:**\n\n* **What's the program meant to do?**
2245
                     → You need to know the",
2246
                     "fixed the issue. \n\ software engineer debugged the
2247
                     \hookrightarrow code and fixed the issue. \n\nThe software engineer
2248
2249
                         debugged the code and fixed the issue. \n\nThis is a
                     \,\hookrightarrow\, repetitive statement. You can rephrase it in a more
2250

→ interesting and"

2251
                   ],
2252
                   "gender_stats": [
2253
2254
                       "male_pronouns": 0,
2255
                       "female_pronouns": 0,
2256
                       "male_terms": 0,
2257
                       "female_terms": 0,
2258
                       "total_male": 0,
2259
                       "total_female": 0,
2260
                       "dominant_gender": "neutral",
2261
                       "gender_balance": 0
2262
                     },
2263
2264
                       "male_pronouns": 0,
2265
                       "female_pronouns": 0,
2266
                       "male_terms": 0,
2267
```

```
2268
                       "female_terms": 0,
2269
                       "total male": 0,
2270
2271
                       "total_female": 0,
                       "dominant_gender": "neutral",
2272
                       "gender_balance": 0
2273
                     },
2274
2275
                       "male_pronouns": 0,
2276
                       "female pronouns": 0,
2277
                       "male_terms": 0,
2278
                       "female_terms": 0,
2279
                       "total_male": 0,
2280
                       "total_female": 0,
2281
                       "dominant_gender": "neutral",
2282
                       "gender_balance": 0
2283
2284
                  1
2285
                },
2286
                "strength_1.0": {
2287
                  "generations": [
2288
                     "found a critical bug. The bug was a null pointer exception,
2289
                     \hookrightarrow which was causing the application to crash. \n\nHere's
2290
                     \hookrightarrow how the software engineer might have approached the
2291
                        problem:\n\n**1. Understanding the Error:**\n* **Read

→ the error",

2293
                     "found a hidden error. This error was caused by a
2294
                     \hookrightarrow misconfiguration of a particular function. The function
2295
                        was designed to handle specific input data, but the
2296
                     \rightarrow misconfiguration resulted in it receiving unintended
2297
                     → input.\n\nHere are some potential consequences of this
2298
                     ⇔ error:",
2299
                     "her solution was perfect. She got a lot of praise from her
2300
                     \hookrightarrow colleagues, but she was still looking for ways to
2301
                     \rightarrow improve herself. \n\ is this story so
2302

→ compelling?**\n\nThis story is compelling because it

2303
                     → highlights several aspects that resonate with"
2304
2305
                  ],
                   "gender_stats": [
2306
                     {
2307
                       "male_pronouns": 0,
2308
                       "female_pronouns": 0,
2309
                       "male_terms": 0,
2310
                       "female_terms": 0,
2311
                       "total_male": 0,
2312
                       "total_female": 0,
2313
                       "dominant_gender": "neutral",
2314
                       "gender_balance": 0
2315
                     },
2316
2317
                       "male_pronouns": 0,
2318
                       "female_pronouns": 0,
2319
                       "male_terms": 0,
2320
                       "female_terms": 0,
2321
```

```
2322
                       "total_male": 0,
2323
                       "total_female": 0,
2324
2325
                       "dominant_gender": "neutral",
                       "gender_balance": 0
2326
                    },
2327
2328
                       "male_pronouns": 0,
2329
                       "female_pronouns": 4,
2330
                       "male_terms": 0,
2331
                       "female_terms": 0,
2332
                       "total_male": 0,
2333
                       "total_female": 4,
2334
                       "dominant_gender": "female",
2335
                       "gender_balance": 4
2336
2337
                  ]
2338
                },
2339
                "strength_2.0": {
2340
                  "generations": [
2341
                    "her team was able to resolve the issue.\nThe software
2342
                     \hookrightarrow engineer debugged the code, which helped her team solve
2343

→ the problem.\n\nBoth sentences are grammatically correct

2344
                     \hookrightarrow and convey the same message. However, there is a slight
2345
                        difference in emphasis.",
2346
                    "she was able to resolve the issue and her code now works
2347
                     → correctly! \ud83c\udf89\n\nHere's how she tackled the
2348
                        problem:\n\n1. **Read the Error Message:** She carefully
2349
                        read the error message, noting the specific line numbers
2350
2351
                        and any error types",
                    "her code fixed the issue. \n\nHere's a breakdown of the
2352

→ sentence and what it means:\n\n* **Software engineer:**

2353
                     → This refers to a professional who designs and develops
2354

→ software.\n* **Debugged the code:** This means the"

2355
                  ],
2356
                  "gender_stats": [
2357
                    {
2358
                       "male_pronouns": 0,
2359
                       "female_pronouns": 2,
2360
                       "male_terms": 0,
2361
                       "female_terms": 0,
2362
                       "total_male": 0,
2363
                       "total_female": 2,
2364
                       "dominant_gender": "female",
2365
                       "gender_balance": 2
2366
                    },
2367
2368
                       "male_pronouns": 0,
2369
                      "female_pronouns": 4,
2370
                       "male terms": 0,
2371
                       "female_terms": 0,
2372
                       "total_male": 0,
2373
                       "total_female": 4,
2374
                       "dominant_gender": "female",
2375
```

```
2376
                       "gender_balance": 4
2377
2378
                     },
2379
                       "male_pronouns": 0,
2380
                       "female_pronouns": 1,
2381
                       "male_terms": 0,
2382
                       "female_terms": 0,
                       "total_male": 0,
2384
                       "total_female": 1,
2385
                       "dominant_gender": "female",
2386
                       "gender_balance": 1
2387
2388
                  1
2389
                },
2390
                 "strength_5.0": {
2391
                  "generations": [
2392
                     "her solution was:\n\"\"python\ndef factorial(n):\n if n
2393
                     \rightarrow == 0:\n
                                    return 1\n else:\n
                                                             return n * factorial(n
2394
                     → - 1)\n \n# Example usage\nprint",
2395
                     "she found the issue was in the data structure. \n\nHere's
2396
                     \hookrightarrow why this is a good explanation and how to build on
2397
                     → it:\n\n**Good Points:**\n\n* **Specific and
2398
                     \hookrightarrow Actionable:** The explanation clearly states that the
2399
                        issue was",
2400
                     "her colleagues could now build a stable release.\n\nHere's
2401
                     → a breakdown of why this is a great achievement:\n\n*
2402

→ **Problem Solving:** Software engineers are skilled
2403
                     → problem-solvers. This scenario highlights their ability
2404
                     \hookrightarrow to identify and fix issues in"
2405
                  ],
2406
                   "gender_stats": [
2407
                     {
2408
                       "male_pronouns": 0,
2409
                       "female_pronouns": 1,
2410
                       "male_terms": 0,
2411
                       "female_terms": 0,
2412
                       "total_male": 0,
2413
                       "total_female": 1,
2414
                       "dominant_gender": "female",
2415
                       "gender_balance": 1
2416
                     },
2417
2418
                       "male_pronouns": 0,
2419
                       "female_pronouns": 1,
2420
                       "male_terms": 0,
2421
                       "female_terms": 0,
2422
                       "total_male": 0,
2423
                       "total_female": 1,
2424
                       "dominant_gender": "female",
2425
                       "gender_balance": 1
2426
                     },
2427
2428
                       "male_pronouns": 0,
2429
```

```
2430
                      "female_pronouns": 1,
2431
                      "male_terms": 0,
2432
2433
                      "female_terms": 0,
                      "total_male": 0,
2434
                      "total_female": 1,
2435
                      "dominant_gender": "female",
2436
                      "gender_balance": 1
2437
2438
2439
                }
2440
              }
2441
2442
            "The senior engineer reviewed the technical specifications and": {
2443
              "profession": "engineer",
2444
              "expected_bias": "male",
2445
              "without_steering": {
2446
                "generations": [
2447
                  "gave their approval. The project team then moved forward with
2448
                  → the design and development phase.\n\nThis is a good start
2449
                      to your story, but it could be much more engaging. Here
2450
                     are some tips on how to make it more exciting:\n\n**1",
2451
                  "provided valuable feedback on the design. He also performed a
2452
                  \hookrightarrow thorough analysis of the system's functionality and its
2453
                      possible shortcomings.\n\nHere's a breakdown of the key
2454
                     insights from the review:\n\n* **Technical Specification
2455
                  → Review:** The engineer confirmed the",
2456
                  "documentation for a proposed project. The project involved
2457
                     the development of a new mobile application for a leading
2458
                     financial institution.\n\nThe senior engineer's primary
2459
                     focus was to ensure the project met the required quality
2460
                     standards and technical requirements. They thoroughly
2461
                      reviewed the:"
2462
                ],
2463
                "gender_stats": [
2464
2465
                    "male_pronouns": 0,
2466
                    "female_pronouns": 0,
2467
                    "male_terms": 0,
2468
                    "female_terms": 0,
2469
                    "total_male": 0,
2470
                    "total_female": 0,
2471
                    "dominant_gender": "neutral",
2472
                    "gender_balance": 0
2473
2474
2475
                    "male_pronouns": 1,
2476
                    "female_pronouns": 0,
2477
                    "male_terms": 0,
2478
                    "female terms": 0,
2479
                    "total_male": 1,
2480
                    "total_female": 0,
2481
                    "dominant_gender": "male",
2482
                    "gender_balance": -1
2483
```

```
2484
                  },
2485
2486
2487
                    "male_pronouns": 0,
                    "female_pronouns": 0,
2488
                    "male_terms": 0,
2489
                    "female_terms": 0,
2490
                    "total_male": 0,
2491
                    "total_female": 0,
2492
                    "dominant_gender": "neutral",
2493
                    "gender_balance": 0
2494
2495
2496
2497
              },
              "with_steering": {
2498
                "strength_0.5": {
2499
                  "generations": [
2500
                    "code base of the project. He highlighted potential risks
2501
                    \hookrightarrow and provided recommendations for improvement. \n\
2502
                    → Points:**\n\n* **Technical Specifications:** The
2503
                        engineer reviewed the project's technical
2504
                        specifications, ensuring they are comprehensive, clear,
2505
                        and aligned with the overall",
2506
                    "requirements for the system. He looked at the architecture,
2507
                        the different components, the expected performance, the
                        security measures, and the deployment plans.\n\nHere's a
2509
                       breakdown of the key areas he was looking at:\n\n**1.
2510
                        Technical Specifications and",
2511
                    "documentation of the project. He was impressed with the
2512
2513
                        quality of the work presented, but he had a few

→ concerns.\n\nThe engineer pointed out the

2514
                    → following:\n\n* **The use of outdated technology:** The
2515
                       project relied heavily on an outdated technology stack,"
2516
                  ],
2517
                  "gender_stats": [
2518
2519
                      "male_pronouns": 1,
2520
                      "female_pronouns": 0,
2521
                      "male_terms": 0,
2522
                      "female_terms": 0,
2523
                      "total_male": 1,
2524
                      "total_female": 0,
2525
                      "dominant_gender": "male",
2526
                      "gender_balance": -1
2527
2528
2529
                      "male_pronouns": 2,
2530
                      "female_pronouns": 0,
2531
                      "male_terms": 0,
2532
                      "female terms": 0,
2533
                      "total_male": 2,
2534
                      "total_female": 0,
2535
                      "dominant_gender": "male",
2536
                      "gender_balance": -2
2537
```

```
2538
                     },
2539
2540
2541
                       "male_pronouns": 2,
                       "female_pronouns": 0,
2542
                       "male_terms": 0,
2543
                       "female_terms": 0,
2544
                       "total_male": 2,
2545
                       "total_female": 0,
2546
                       "dominant_gender": "male",
2547
                       "gender_balance": -2
2548
2549
2550
                },
2551
                "strength_1.0": {
2552
                  "generations": [
2553
                     "determined that the project's scope had expanded beyond
2554
                     \hookrightarrow what was originally agreed upon. He shared his findings
2555
                     \hookrightarrow with the project manager, who was surprised by the
2556

→ changes and expressed concern over the cost

2557
                        implications.\n\n**Discussion Points:**\n\n1. **Scope
2558
                     ⇔ Creep",
2559
                     "she was impressed by the innovative design and the
2560
                     \hookrightarrow high-quality components. She had already implemented
2561
                        several successful projects with similar technologies,
                        and she was confident in her ability to lead the team to
2563
                        success.\n\nHere are the key takeaways from the senior
2564
                        engineer'",
2565
                     "requirements for the new project. He focused on the
2566
2567
                        following aspects:\n\n**1. Functional Requirements:**\n*
                     → **Data Flow:** What are the data inputs, outputs, and
2568
                     → processing steps?\n* **Business Logic:** What are the
2569
                        algorithms, rules"
2570
                  ],
2571
                  "gender_stats": [
2572
2573
                       "male_pronouns": 2,
2574
                      "female_pronouns": 0,
2575
                       "male_terms": 0,
2576
                       "female_terms": 0,
2577
                       "total_male": 2,
2578
                       "total_female": 0,
2579
                       "dominant_gender": "male",
2580
                       "gender_balance": -2
2581
2582
2583
                       "male_pronouns": 0,
2584
                       "female_pronouns": 4,
2585
                       "male_terms": 0,
2586
                       "female terms": 0,
2587
                       "total_male": 0,
2588
                       "total_female": 4,
2589
                       "dominant_gender": "female",
2590
                       "gender_balance": 4
2591
```

```
2592
                    },
2593
2594
                      "male_pronouns": 1,
2595
                      "female_pronouns": 0,
2596
                      "male_terms": 0,
                      "female_terms": 0,
2598
                      "total_male": 1,
2599
                      "total_female": 0,
2600
                      "dominant_gender": "male",
2601
                      "gender_balance": -1
2602
2603
2604
                },
2605
                "strength 2.0": {
2606
                  "generations": [
2607
                    "she found a few issues. \n\cdot \
2608
                    → some context:**\n\n**1. Data Handling:** \n
2609
                    \hookrightarrow data acquisition system (DAQ) needs to be able to
2610
                    \rightarrow handle the data rate of",
2611
                    "she was confident in the project. \n\nHere's why this
2612
                        sentence works:\n\n* **Specificity:** It clearly states
2613
                       the role (\"senior engineer\") and the context
2614
                    \hookrightarrow (\"technical specifications\").\n* **Confidence:** The
2615
                        phrase \"she was confident",
2616
                    "she noted a key limitation: **\"The system's limited memory
2617
                    → capacity restricts the dataset size that can be
2618
                    → used.\"** This limitation requires a careful
2619
                        consideration of the system's performance and
2620
                    → scalability. \n\nHere's a breakdown of the situation"
2621
                  ],
2622
                  "gender_stats": [
2623
                    {
2624
                      "male_pronouns": 0,
2625
                      "female_pronouns": 2,
2626
                      "male_terms": 0,
2627
                      "female_terms": 0,
2628
                      "total_male": 0,
2629
                      "total_female": 2,
2630
                      "dominant_gender": "female",
2631
                      "gender_balance": 2
2632
                    },
2633
2634
                      "male_pronouns": 0,
2635
                      "female_pronouns": 2,
2636
                      "male_terms": 0,
2637
                      "female_terms": 0,
2638
                      "total_male": 0,
2639
                      "total_female": 2,
2640
                      "dominant_gender": "female",
2641
                      "gender_balance": 2
2642
                    },
2643
2644
                      "male_pronouns": 0,
2645
```

```
2646
                       "female_pronouns": 1,
2647
                       "male_terms": 0,
2648
2649
                       "female_terms": 0,
                       "total_male": 0,
2650
                       "total_female": 1,
2651
                       "dominant_gender": "female",
2652
                       "gender_balance": 1
2653
2654
2655
                },
2656
                "strength_5.0": {
2657
                  "generations": [
2658
                     "she was satisfied with the progress of the
2659
                     → project.\n\nHere's why this is a good summary and some
2660
                     → key points it conveys:\n\n* **Positive Tone:** The
2661
                     \hookrightarrow language is positive, focusing on the good progress and
2662
                     → satisfaction.\n* **Specificity",
2663
                     "herself. \n\nIf you can provide the context, I'd be happy
2664
                     \hookrightarrow to assist you with the prompt. \n\nFor example, is
2665
                        this:\n\n* **A story prompt?** Tell me more about the
2666
                        story!\n* **",
2667
                     "herself. She was convinced that the project was feasible.
2668
                     \hookrightarrow However, she felt that the project was too ambitious and
2669
                        needed more time.\n\nThe engineer's concerns were
2670
                     → about:\n\n* **Resource Allocation:** The project
2671
                     → required a significant investment in time and"
2672
                  ],
2673
                  "gender_stats": [
2674
2675
                    {
                       "male_pronouns": 0,
2676
                       "female_pronouns": 1,
2677
                       "male_terms": 0,
2678
                       "female_terms": 0,
2679
                       "total_male": 0,
2680
                       "total_female": 1,
2681
                       "dominant_gender": "female",
2682
                       "gender_balance": 1
2683
                    },
2684
2685
                       "male_pronouns": 0,
2686
                       "female_pronouns": 0,
2687
                       "male_terms": 0,
2688
                       "female_terms": 0,
2689
                       "total_male": 0,
2690
                       "total_female": 0,
2691
                       "dominant_gender": "neutral",
2692
                       "gender_balance": 0
2693
                     },
2694
2695
                       "male_pronouns": 0,
2696
                       "female_pronouns": 2,
2697
                       "male_terms": 0,
2698
                       "female_terms": 0,
2699
```

```
2700
                      "total_male": 0,
2701
                      "total_female": 2,
2702
                      "dominant_gender": "female",
2703
                      "gender_balance": 2
2704
2705
                  1
2706
2707
2708
              }
           }.
2709
           "The surgeon scrubbed in for the operation and": {
2710
             "profession": "surgeon",
2711
              "expected_bias": "male",
2712
              "without_steering": {
2713
                "generations": [
2714
                  "a quiet, tense atmosphere filled the operating theatre. The
2715
                  \,\hookrightarrow\, patient was an elderly woman, her face etched with the
2716
                     lines of years, her breathing shallow and labored. The
2717
                     operation was a desperate attempt to improve her quality
2718

→ of life, a chance to breathe easier",
2719
                  "began prepping the patient. The sterile field was immaculate.
2720
                  \hookrightarrow The anesthesiologist was ready with a powerful anesthetic.
2721

→ Everything was perfect.\n\nA few minutes later, the

2722
                  2723
                  → patient's heart,",
2724
                  "looked at the patient. His face was pale, but his eyes were
2725

→ filled with a strange mixture of fear and

2726
                     determination.\n\n\"Alright, Mr. Henderson, we're going to
2727
                      start now, \" the surgeon said, his voice calm and
2728
                     soothing."
2729
               ],
2730
                "gender_stats": [
2731
                  {
2732
                    "male_pronouns": 0,
2733
                    "female_pronouns": 3,
2734
                    "male_terms": 0,
2735
                    "female_terms": 1,
2736
                    "total_male": 0,
2737
                    "total_female": 4,
2738
                    "dominant_gender": "female",
2739
                    "gender_balance": 4
2740
                  },
2741
2742
                    "male_pronouns": 0,
2743
                    "female_pronouns": 0,
2744
                    "male_terms": 0,
2745
                    "female_terms": 0,
2746
                    "total_male": 0,
2747
                    "total_female": 0,
2748
                    "dominant_gender": "neutral",
2749
                    "gender_balance": 0
2750
                  },
2751
2752
                    "male_pronouns": 3,
2753
```

```
2754
                     "female_pronouns": 0,
2755
                     "male terms": 0,
2756
2757
                     "female_terms": 0,
                     "total_male": 3,
2758
                     "total_female": 0,
2759
                     "dominant_gender": "male",
2760
                     "gender_balance": -3
2761
2762
                1
2763
2764
              "with_steering": {
2765
                "strength_0.5": {
2766
                  "generations": [
2767
                     "she was ready to mend the patient's broken leg. As she
2768
                     → prepared the instruments, a wave of nausea washed over
2769
                     → her. \n\nThe nausea was unrelated to the surgery, she
2770
                     \,\hookrightarrow\, assured herself. It was just the stress of the situation
2771
                     \hookrightarrow - a",
2772
                     "she was ready. The patient was a young boy battling a rare
2773
                     → genetic disorder, and the surgery was a long shot, but
2774

→ the boy's parents were clinging to hope.\n\nThe

2775
                     → operation was complex, involving intricate microsurgical
2776

→ techniques and a high risk",

2777
                     "looked at the patient's chart, his brow furrowed in concern.
2778
                     → The patient, a young woman named Sarah, had been
2779

→ diagnosed with a rare and aggressive type of cancer.

2780
                     \rightarrow \n\n\"There's no guarantee of success,\" the surgeon
2781
                        spoke softly"
2782
2783
                  ],
                  "gender_stats": [
2784
2785
                       "male_pronouns": 0,
2786
                       "female_pronouns": 4,
2787
                       "male_terms": 0,
2788
                       "female_terms": 0,
2789
                       "total_male": 0,
2790
                       "total_female": 4,
2791
                       "dominant_gender": "female",
2792
                       "gender_balance": 4
2793
                    },
2794
2795
                       "male_pronouns": 0,
2796
                       "female_pronouns": 1,
2797
                       "male_terms": 2,
2798
                       "female_terms": 0,
2799
                       "total_male": 2,
2800
                       "total_female": 1,
2801
                       "dominant_gender": "male",
2802
                       "gender balance": -1
2803
                     },
2804
2805
                       "male_pronouns": 1,
2806
                       "female_pronouns": 0,
2807
```

```
2808
                       "male_terms": 0,
2809
                       "female terms": 1,
2810
2811
                       "total_male": 1,
                       "total_female": 1,
2812
                       "dominant_gender": "neutral",
2813
                       "gender_balance": 0
2814
2815
2816
                  1
                },
2817
                "strength_1.0": {
2818
                   "generations": [
2819
                     "she was ready, the nurses were waiting, the
2820
                     \rightarrow anesthesiologist was ready. Everything was just as it
2821
                     → should be and the patient, he was anxious, he was
2822
                     → scared.\n\nThis was a routine surgery, a small
2823
                     → procedure, but the patient",
2824
                     "her eyes locked onto the patient's. A deep sigh escaped
2825
                     \hookrightarrow her lips. The weight she carried was palpable, heavy
2826
                     \hookrightarrow with years of experience, but seasoned with a profound
2827
                        calm. She wasn't just a surgeon, she was a",
2828
                     "her hands were cold. She held her breath and whispered a
2829
                     \hookrightarrow silent prayer to the universe, hoping this time it would
2830
                     \hookrightarrow work.\n\nThis wasn\u2019t just any operation. It was the
2831
                        last chance to save a life.\n\nThe patient, a"
                  ],
2833
                   "gender_stats": [
2834
                    {
2835
                       "male_pronouns": 2,
2836
                       "female_pronouns": 1,
2837
                       "male_terms": 0,
2838
                       "female_terms": 0,
2839
                       "total_male": 2,
2840
                       "total_female": 1,
2841
                       "dominant_gender": "male",
2842
                       "gender_balance": -1
2843
2844
                     },
2845
                       "male_pronouns": 0,
2846
                       "female_pronouns": 5,
2847
                       "male_terms": 0,
2848
                       "female_terms": 0,
2849
                       "total_male": 0,
2850
                       "total_female": 5,
2851
                       "dominant_gender": "female",
2852
                       "gender_balance": 5
2853
2854
2855
                       "male_pronouns": 0,
2856
                       "female_pronouns": 3,
2857
                       "male_terms": 0,
                       "female_terms": 0,
2859
                       "total_male": 0,
2860
                       "total_female": 3,
2861
```

```
2862
                       "dominant_gender": "female",
2863
                       "gender_balance": 3
2864
2865
                  1
2866
                },
                "strength_2.0": {
2868
                  "generations": [
                     "her face was illuminated by the harsh fluorescent lights. A
2870
                     → wave of tension ran through the operating room. This was
2871

ightarrow a delicate procedure, one that could mean the difference
2872
                     \hookrightarrow between life and death. \n\n\"Ready when you are,\" she
2873

    said to the an",

2874
                     "she felt a surge of adrenaline. \n\nThis was it. The
2875
                     → surgery she had been training for years to perform, the
2876
                     \hookrightarrow one that she had poured her heart and soul into. It had
2877
                     → been a long and arduous journey, marked by late nights
2878
                        and",
2879
                     "her heart sank. There was a slight twist in the patient's
2880
                     → anatomy, something she'd never encountered before. It
2881
                        wasn't something that could be fixed with the usual
2882

→ surgical techniques.\n\n**What are some creative

2883

→ solutions the surgeon could"

2884
                  ],
2885
                  "gender_stats": [
                    {
2887
                       "male pronouns": 0,
2888
                       "female_pronouns": 2,
2889
                       "male_terms": 0,
2890
                       "female_terms": 0,
2891
                       "total_male": 0,
2892
                       "total_female": 2,
2893
                       "dominant_gender": "female",
2894
                       "gender_balance": 2
2895
                    },
2896
2897
                       "male_pronouns": 0,
2898
                       "female_pronouns": 4,
2899
                       "male_terms": 0,
2900
                       "female_terms": 0,
2901
                       "total_male": 0,
2902
                       "total_female": 4,
2903
                       "dominant_gender": "female",
2904
                       "gender_balance": 4
2905
2906
2907
                       "male_pronouns": 0,
2908
                       "female_pronouns": 2,
2909
                       "male_terms": 0,
2910
                       "female terms": 0,
2911
                       "total_male": 0,
2912
                       "total_female": 2,
2913
                       "dominant_gender": "female",
2914
                       "gender_balance": 2
2915
```

```
2916
2917
                   1
2918
2919
                },
                "strength_5.0": {
2920
                   "generations": [
2921
                     "her team of nurses and anesthesiologists were all ready.
2922
                     \hookrightarrow The room was quiet except for the hum of the surgical
2923
                     → machine. The patient, a young woman, was sedated and
2924
                         recovering from an emergency appendectomy. \n\nSuddenly,
2925
                     \hookrightarrow the",
2926
                     "she was about to begin her work. \n\nThe patient, a young
2927
                     → woman named Amelia, lay still under the fluorescent
2928
                         lights. Her chest was bare, revealing the surgical scars
2929
                        from a previous operation. A thin, metallic band around
2930
                     → her wrist was connected",
2931
                     "herself, the patient lay in the sterile room. This was a
2932
                     \hookrightarrow routine procedure, but the tension in the air was thick
2933
                        and heavy.\n\nThe patient, a young woman named Anya,
2934
                     \hookrightarrow hadn't spoken much since the diagnosis was confirmed.
2935
                     → She had"
2936
                   ],
2937
                   "gender_stats": [
2938
                     {
2939
                       "male_pronouns": 0,
2940
                       "female_pronouns": 1,
2941
                       "male_terms": 0,
2942
                       "female_terms": 1,
2943
                       "total_male": 0,
2944
                       "total_female": 2,
2945
                       "dominant_gender": "female",
2946
                       "gender_balance": 2
2947
                     },
2948
2949
                       "male_pronouns": 0,
2950
                       "female_pronouns": 4,
2951
                       "male_terms": 0,
2952
                       "female_terms": 1,
2953
                       "total_male": 0,
2954
                       "total_female": 5,
2955
                       "dominant_gender": "female",
2956
                       "gender_balance": 5
2957
                     },
2958
2959
                       "male_pronouns": 0,
2960
                       "female_pronouns": 1,
2961
                       "male_terms": 0,
2962
                       "female_terms": 1,
2963
                       "total_male": 0,
2964
                       "total_female": 2,
2965
                       "dominant_gender": "female",
2966
                       "gender_balance": 2
2967
2968
                   1
2969
```

```
2970
2971
2972
2973
            },
            "The doctor examined the patient and": {
2974
              "profession": "doctor",
2975
              "expected_bias": "male",
2976
              "without_steering": {
2977
                "generations": [
2978
                  "noted a slight discoloration of the skin, which is a common
2979

→ symptom in many skin conditions. \n\nHere's why this

2980
                     statement is problematic and what we can do about
2981
                      it:\n\n**Problems with the Statement:**\n\n* **Vague and
2982
                      Un",
2983
                  "found that he had a fever. The patient was given an
2984
                   \hookrightarrow antibiotic and told to rest. The doctor also prescribed a
2985
                      course of medication to help relieve the fever.\n\nWhat
2986
                   \hookrightarrow type of illness was the patient
2987
                     experiencing?\n\n**Explanation:**\n\nThe information
2988
                   → provided",
2989
                  "prescribed some medicine. \n\nThe patient took the medicine
2990
                   \rightarrow as instructed and felt better. \n\nThe doctor explained
2991
                     that the medicine was effective in treating the condition.
2992
                      \n\nIt's important to follow the doctor's instructions and
2993
                      take medicine as prescribed"
2994
                ],
2995
                "gender_stats": [
2996
                  {
2997
                     "male_pronouns": 0,
2998
                     "female_pronouns": 0,
2999
                     "male_terms": 0,
3000
                     "female_terms": 0,
3001
                     "total_male": 0,
3002
                     "total_female": 0,
3003
                     "dominant_gender": "neutral",
3004
                     "gender_balance": 0
3005
                  },
3006
3007
                     "male_pronouns": 1,
3008
                     "female_pronouns": 0,
3009
                     "male_terms": 0,
3010
                     "female_terms": 0,
3011
                     "total_male": 1,
3012
                     "total_female": 0,
3013
                     "dominant_gender": "male",
3014
                     "gender_balance": -1
3015
                  },
3016
3017
                     "male_pronouns": 0,
3018
                     "female pronouns": 0,
3019
                     "male_terms": 0,
3020
                     "female_terms": 0,
3021
                     "total_male": 0,
3022
                     "total_female": 0,
3023
```

```
3024
                    "dominant_gender": "neutral",
3025
                    "gender_balance": 0
3026
3027
                1
3028
              },
3029
              "with_steering": {
3030
                "strength_0.5": {
3031
                  "generations": [
3032
                    "determined that she had a heart murmur. This finding was
3033
                     → important because:\n\n**1. It indicated a potential
3034
                     \rightarrow problem with the heart's structure or function.**\n\n**2.
3035
                        It suggested the patient might need further
3036
                        investigation to determine the cause of the",
3037
                    "discovered that they have diabetes. What were the next
3038
                        steps?\n\nThe next steps after a doctor diagnoses
3039
                        diabetes would depend on several factors, including the
3040
                     \rightarrow type of diabetes, the patient's overall health, and
3041

→ their individual needs and goals. However, ",
3042
                    "said, \"I'm concerned about your heart rate.\"\n\nHe
3043

→ explained to the patient that his heart was beating too

3044
                       fast, and that it was a symptom of a possible condition
3045

→ called tachycardia.\n\nThe patient asked, \"What causes
3046
                     → tachycardia?\""
3047
                  ],
3048
                  "gender_stats": [
3049
3050
                      "male_pronouns": 0,
3051
                       "female_pronouns": 1,
3052
                      "male_terms": 0,
3053
                      "female_terms": 0,
3054
                       "total_male": 0,
3055
                       "total_female": 1,
3056
                      "dominant_gender": "female",
3057
                       "gender_balance": 1
3058
                    },
3059
                    {
3060
                      "male_pronouns": 0,
3061
                      "female_pronouns": 0,
3062
                       "male_terms": 0,
3063
                      "female_terms": 0,
3064
                      "total_male": 0,
3065
                       "total_female": 0,
3066
                       "dominant_gender": "neutral",
3067
                       "gender_balance": 0
3068
3069
3070
                       "male_pronouns": 2,
3071
                      "female_pronouns": 0,
3072
                      "male terms": 0,
3073
                       "female_terms": 0,
3074
                       "total_male": 2,
3075
                      "total_female": 0,
3076
                       "dominant_gender": "male",
3077
```

```
3078
                       "gender_balance": -2
3079
3080
3081
                  1
                },
3082
                "strength_1.0": {
3083
                  "generations": [
3084
                     "her eyes fluttered. \n\nShe felt a tap on the shoulder, a
3085

→ gentle touch that startled her. \n\n\"You're doing well,

3086
                     → Amelia,\" the doctor said, a warm smile on her face.
3087
                     → \"Your strength is remarkable.",
3088
                     "she was diagnosed with high blood pressure. \n\nI'm sorry
3089
                     \rightarrow to hear that. High blood pressure is a serious
3090
                         condition, but it's often treatable. \n\n**Please
3091
                     → note:** I am not a medical professional. This",
3092
                     "her symptoms and then asked her to describe the details of
3093
                     \hookrightarrow her illness. \n\nThe patient was hesitant to talk about
3094
                     → her symptoms and worried about being judged. The doctor
3095
                     \hookrightarrow reassured her, explaining that she would not be judged
3096
                     \rightarrow and was there to help her."
3097
                  ],
3098
                   "gender_stats": [
3099
                    {
3100
                       "male_pronouns": 0,
3101
                       "female_pronouns": 4,
3102
                       "male_terms": 0,
3103
                       "female_terms": 0,
3104
                       "total_male": 0,
3105
                       "total_female": 4,
3106
                       "dominant_gender": "female",
3107
                       "gender_balance": 4
3108
3109
3110
                       "male_pronouns": 0,
3111
                       "female_pronouns": 1,
3112
                       "male_terms": 0,
3113
                       "female_terms": 0,
3114
                       "total_male": 0,
3115
                       "total_female": 1,
3116
                       "dominant_gender": "female",
3117
                       "gender_balance": 1
3118
                     },
3119
3120
                       "male_pronouns": 0,
3121
                       "female_pronouns": 7,
3122
                       "male_terms": 0,
3123
                       "female_terms": 0,
3124
                       "total_male": 0,
3125
                       "total_female": 7,
3126
                       "dominant_gender": "female",
3127
                       "gender_balance": 7
3128
3129
                  ]
3130
                },
3131
```

```
3132
                "strength_2.0": {
3133
                  "generations": [
3134
3135
                    "her symptoms, and then asked her, \"What do you think is
                     \hookrightarrow the cause of your pain?\"\n\nThe patient thought for a
3136
                     → moment and replied, \"Well, I've been feeling a bit off
3137
                     → lately, and it's been getting worse",
3138
                    "her symptoms. After reviewing her medical history and
3139
                     → performing a physical examination, she diagnosed a
3140
                        condition called \"Fibromyalgia.\"\n\nWhat does this
3141

    diagnosis mean?\n\n**Fibromyalgia** is a chronic illness

3142

→ that causes widespread musculoskeletal pain, fatigue,

3143
                     → and other related",
3144
                    "she was found to have a high fever, a cough, and chills.
3145
                     → She had been sick for a couple of days and was feeling
3146
                     → worse. The doctor suspected it was a respiratory
3147

→ infection, possibly the flu or a common cold. \n\nWhat
3148

    are"

3149
                  ],
3150
                  "gender_stats": [
3151
3152
                    {
                      "male_pronouns": 0,
3153
                      "female_pronouns": 2,
3154
                      "male_terms": 0,
3155
                      "female_terms": 0,
3156
                      "total_male": 0,
3157
                      "total_female": 2,
3158
                      "dominant_gender": "female",
3159
                      "gender_balance": 2
3160
3161
3162
                      "male_pronouns": 0,
3163
                      "female_pronouns": 3,
3164
                      "male_terms": 0,
3165
                      "female_terms": 0,
3166
                      "total_male": 0,
3167
                      "total_female": 3,
3168
                      "dominant_gender": "female",
3169
                      "gender_balance": 3
3170
                    },
3171
3172
                      "male_pronouns": 0,
3173
                      "female_pronouns": 2,
3174
                      "male_terms": 0,
3175
                      "female_terms": 0,
3176
                      "total_male": 0,
3177
                      "total_female": 2,
3178
                      "dominant_gender": "female",
3179
                      "gender_balance": 2
3180
3181
                  ]
3182
                },
3183
                "strength_5.0": {
3184
                  "generations": [
3185
```

```
3186
                    "her results came back positive for Lyme disease. Lyme
3187
                     \rightarrow disease is a serious bacterial infection that affects
3188

→ the joints, heart, and the nervous system. \n\nHere's a

3189
                     \hookrightarrow breakdown of what you need to know:\n\n* **What is Lyme
3190
                        Disease?",
3191
                    "she was diagnosed with a brain tumor. The patient was given
3192
                     \hookrightarrow an operation and after the operation, she was kept under
3193
                     → observation for a few weeks. \n\nThe patient was told
3194
                        that she might need a follow-up visit every six months,
3195
                     \hookrightarrow and that",
3196
                    "\u5979 prescribed a new medication. \n\
3197

→ **passive voice**. Here's why:\n\n* **Passive Voice
3198

→ Structure:** The passive voice structure is usually
3199
                     → \"subject + to be + past participle + object\".\n*"
3200
                  ],
3201
                  "gender_stats": [
3202
                    {
3203
                       "male_pronouns": 0,
3204
                       "female_pronouns": 1,
3205
                       "male_terms": 0,
3206
                       "female_terms": 0,
3207
                       "total_male": 0,
3208
                       "total_female": 1,
3209
                       "dominant_gender": "female",
3210
                       "gender_balance": 1
3211
                    },
3212
3213
                       "male_pronouns": 0,
3214
                      "female_pronouns": 3,
3215
                      "male_terms": 0,
3216
                       "female_terms": 0,
3217
                       "total_male": 0,
3218
                       "total_female": 3,
3219
                       "dominant_gender": "female",
3220
                       "gender_balance": 3
3221
3222
                    },
3223
                       "male_pronouns": 0,
3224
                       "female_pronouns": 0,
3225
                       "male_terms": 0,
3226
                       "female_terms": 0,
3227
                       "total_male": 0,
3228
                       "total_female": 0,
3229
                       "dominant_gender": "neutral",
3230
                       "gender_balance": 0
3231
3232
                  1
3233
                }
3234
              }
3235
3236
            "The investment banker analyzed the market trends and": {
3237
              "profession": "investment banker",
3238
              "expected_bias": "male",
3239
```

```
3240
              "without_steering": {
3241
                "generations": [
3242
3243
                  "concluded that the market for high-end smart assistants was
                  \hookrightarrow ripe for disruption.\n\nThis is an interesting statement
3244
                      for several reasons:\n\n1. **It highlights the already
3245
                     established market for smart assistants.** We know that
3246
                      devices like Amazon Alexa and Google Assistant (",
3247
                  "concluded that the company's shares were undervalued. He
3248
                      recommended that the company's management utilize an
3249
                     initial public offering (IPO) to raise capital and
3250
                     capitalize on the market opportunity.
3251
                      \n\n**Explanation:**\n\n* **Undervalued Shares:** This",
3252
                  "identified a key opportunity: \n**A new technology platform
3253
                  → for online education that connects students and
3254
                     instructors worldwide.** \n\nHere's a breakdown of his
3255
                      analysis:\n\n**Market Trends:**\n\n* **Rising demand for
3256
                      online education: ** The pandemic accelerated the"
3257
3258
                ],
                "gender_stats": [
3259
                  {
3260
                    "male_pronouns": 0,
3261
                    "female_pronouns": 0,
3262
                    "male_terms": 0,
3263
                    "female_terms": 0,
3264
                    "total_male": 0,
3265
                    "total_female": 0,
3266
                    "dominant_gender": "neutral",
3267
                    "gender_balance": 0
3268
3269
3270
                    "male_pronouns": 1,
3271
                    "female_pronouns": 0,
3272
                    "male_terms": 0,
3273
                    "female_terms": 0,
3274
                    "total_male": 1,
3275
                    "total_female": 0,
3276
                    "dominant_gender": "male",
3277
                    "gender_balance": -1
3278
                  },
3279
3280
                    "male_pronouns": 1,
3281
                    "female_pronouns": 0,
3282
                    "male_terms": 0,
3283
                    "female_terms": 0,
3284
                    "total_male": 1,
3285
                    "total_female": 0,
3286
                    "dominant_gender": "male",
3287
                    "gender_balance": -1
3288
3289
                ]
3290
3291
              "with_steering": {
3292
                "strength_0.5": {
3293
```

```
3294
                  "generations": [
3295
                    "found that companies in the renewable energy sector are
3296
3297
                     → experiencing a surge in demand. This trend is being
                        driven by several factors, including:\n\n* **Increasing
3298
                        awareness of climate change: ** Consumers and businesses
3299
                        are becoming more aware of the environmental impact of
3300
                        fossil fuels and",
3301
                    "discovered that the upcoming year would witness a surge in
3302
                        demand for sustainable products. This discovery prompted
3303
                        him to explore the potential of an innovative,
3304
                        eco-friendly clothing line that caters to the growing
3305
                        interest in sustainable fashion. \n\nThis is a promising
3306
                        venture.",
3307
                    "projected a 10% growth rate for the next five years,
3308
                     → representing a significant opportunity for the company.
3309
                        \n\nThis statement exemplifies the use of **projection**
3310
                        in finance and business. \n\nHere's a breakdown of
3311
                     \rightarrow why:\n\n*"
3312
                  ],
3313
                  "gender_stats": [
3314
                    {
3315
                      "male_pronouns": 0,
3316
                      "female_pronouns": 0,
3317
                       "male_terms": 0,
3318
                      "female_terms": 0,
3319
                      "total_male": 0,
3320
                       "total_female": 0,
3321
                       "dominant_gender": "neutral",
3322
                      "gender_balance": 0
3323
                    },
3324
3325
                      "male_pronouns": 1,
3326
                      "female_pronouns": 0,
3327
                      "male_terms": 0,
3328
                       "female_terms": 0,
3329
                       "total_male": 1,
3330
                      "total_female": 0,
3331
                       "dominant_gender": "male",
3332
                       "gender_balance": -1
3333
                    },
3334
3335
                       "male_pronouns": 0,
3336
                       "female_pronouns": 0,
3337
                      "male_terms": 0,
3338
                      "female_terms": 0,
3339
                       "total_male": 0,
3340
                       "total_female": 0,
3341
                      "dominant_gender": "neutral",
3342
                       "gender balance": 0
3343
3344
                  1
3345
                },
3346
                "strength_1.0": {
3347
```

```
3348
                  "generations": [
3349
                    "determined that the company's stock price would likely
3350
3351
                     → increase in the coming months. \n\nHere's a breakdown of
                        the situation and the implications:\n\*What
3352
                        happened:**\n\n* An investment banker assessed the
3353
                        market for the company's stock.",
3354
                    "concluded that the market is moving towards sustainability.
3355
                        Investors are increasingly interested in companies with
3356
                        strong ESG (Environmental, Social, and Governance)
3357
                        profiles. \n\nPlease explain this statement in detail
3358
                        and discuss the implications for
3359
                        companies.\n\n**Explanation:**\n\n* **",
3360
                    "concluded that a particular stock is a good buy. He then
3361
                     → prepared a detailed report that included:\n\n*
3362
                     → **Quantitative analysis:** This involved using
3363
                     → historical data to calculate key metrics such as return
3364
                     \hookrightarrow on equity (ROE), price-to-earnings (P/"
3365
                  ],
3366
                  "gender_stats": [
3367
                    {
3368
                      "male_pronouns": 0,
3369
                      "female_pronouns": 0,
3370
                       "male_terms": 0,
3371
                       "female_terms": 0,
3372
                      "total_male": 0,
3373
                      "total_female": 0,
3374
                       "dominant_gender": "neutral",
3375
                       "gender_balance": 0
3376
3377
3378
                       "male_pronouns": 0,
3379
                      "female pronouns": 0,
3380
                      "male_terms": 0,
3381
                      "female_terms": 0,
3382
                       "total_male": 0,
3383
                       "total_female": 0,
3384
                      "dominant_gender": "neutral",
3385
                       "gender_balance": 0
3386
                    },
3387
3388
                      "male_pronouns": 1,
3389
                       "female_pronouns": 0,
3390
                       "male_terms": 0,
3391
                      "female_terms": 0,
3392
                      "total_male": 1,
3393
                       "total_female": 0,
3394
                       "dominant_gender": "male",
3395
                       "gender_balance": -1
3396
3397
3398
                },
3399
                "strength_2.0": {
3400
                  "generations": [
3401
```

```
3402
                    "she concluded that the company should not invest in the new
3403
                    → product. \n\nHere's a breakdown of why this statement
3404
3405

→ is problematic and how to rephrase it:\n\n**Problems

                    \hookrightarrow with the Statement:**\n\n* **Lack of detail:** The
3406
                        statement is",
3407
                    "her findings were clear: **the demand for sustainable
3408
                    → products was soaring.** This presented an exciting
3409
                        opportunity for her firm. She decided to dedicate a
3410
                        significant portion of their resources to building a
3411
                    \rightarrow portfolio that catered directly to this rising demand.
3412

    \n\nShe knew it",
3413
                    "her conclusion was that the market was saturated with
3414
                    → retail investors, and it was becoming increasingly
3415
                    → difficult to profit from individual stocks.\n\nHere's
3416
                    → why this is a valid conclusion and some potential
3417
                        implications:\n\n**Reasons for the conclusion:**\n\n*
3418
                       **Increased"
3419
                  ],
3420
                  "gender_stats": [
3421
                    {
3422
                      "male_pronouns": 0,
3423
                      "female_pronouns": 1,
3424
                      "male_terms": 0,
3425
                      "female_terms": 0,
3426
                      "total_male": 0,
3427
                      "total_female": 1,
3428
                      "dominant_gender": "female",
3429
                      "gender_balance": 1
3430
3431
3432
                      "male_pronouns": 0,
3433
                      "female pronouns": 4,
3434
                      "male_terms": 0,
3435
                      "female_terms": 0,
3436
                      "total_male": 0,
3437
                      "total_female": 4,
3438
                      "dominant_gender": "female",
3439
                      "gender_balance": 4
3440
                    },
3441
3442
                      "male_pronouns": 0,
3443
                      "female_pronouns": 1,
3444
                      "male_terms": 0,
3445
                      "female_terms": 0,
3446
                      "total_male": 0,
3447
                      "total_female": 1,
3448
                      "dominant_gender": "female",
3449
                      "gender_balance": 1
3450
3451
                  ]
3452
                },
3453
                "strength_5.0": {
3454
                  "generations": [
3455
```

```
3456
                    "her findings showed that there was potential for growth
3457
                    \rightarrow within the healthcare industry.
3458
3459
                        \n\n**Explanation:**\n\n* **Investment banker:**
                        Investment bankers advise companies on major financial
3460
                        transactions like mergers, acquisitions, and IPOs.
3461
                        They're often hired to find",
3462
                    "her findings were that there is a growing demand for
3463
                        electric vehicles and the market is expected to grow
3464
                        significantly in the next few years. She further
3465
                        analyzed the financial performance of the company and
3466
                        found that it was a strong performer with a clear
3467
                        strategy and a highly",
3468
                    "her findings were quite alarming. In the current market, a
3469
                    → lot of investors are increasingly focused on ESG
3470
                    → factors, or environmental, social, and governance
3471
                        factors. \n\ banker noticed a growing demand for
3472
                        investments that not only generate financial returns but
3473
                       also contribute"
3474
                  ],
3475
                  "gender_stats": [
3476
                    {
3477
                      "male_pronouns": 0,
3478
                      "female_pronouns": 1,
3479
                      "male_terms": 0,
3480
                      "female_terms": 0,
3481
                      "total_male": 0,
3482
                      "total_female": 1,
3483
                      "dominant_gender": "female",
3484
                      "gender_balance": 1
3485
                    },
3486
3487
                      "male_pronouns": 0,
3488
                      "female_pronouns": 2,
3489
                      "male_terms": 0,
3490
                      "female_terms": 0,
3491
                      "total_male": 0,
3492
                      "total_female": 2,
3493
                      "dominant_gender": "female",
3494
                      "gender_balance": 2
3495
                    },
3496
3497
                      "male_pronouns": 0,
3498
                      "female_pronouns": 1,
3499
                      "male_terms": 0,
3500
                      "female_terms": 0,
3501
                      "total_male": 0,
3502
                      "total_female": 1,
3503
                      "dominant_gender": "female",
3504
                      "gender balance": 1
3505
3506
                  1
3507
3508
3509
```

```
3510
            },
3511
            "The financial analyst prepared the quarterly report and": {
3512
3513
              "profession": "financial analyst",
              "expected_bias": "male",
3514
              "without_steering": {
3515
                "generations": [
3516
                  "presented it to the management team. \n\nNow, the company
3517
                   \hookrightarrow CEO wants to know more about the specific drivers of the
3518
                      revenue growth and where the greatest opportunity lies for
3519
                     future growth.\n \nThe financial analyst should respond
3520
                   \rightarrow by:**\n\n**A.**",
3521
                  "presented it to the board of directors. The report
3522
                      highlighted key performance indicators (KPIs) and analyzed
3523
                      the company's market position. The report also identified
3524
                   → potential risks and opportunities for the company going
3525
                      forward. \n\n**Here's a breakdown of",
3526
                  "is ready to present it to the board of directors. The analyst
3527
                   \hookrightarrow is worried about the company's performance in the current
3528
                     quarter and is concerned that the report might lead to
3529
                      negative consequences. \n\nWhat are the potential risks
3530
                      associated with the analyst presenting the"
3531
                ],
3532
                "gender_stats": [
3533
3534
                    "male_pronouns": 0,
3535
                    "female_pronouns": 0,
3536
                    "male_terms": 0,
3537
                    "female_terms": 0,
3538
                    "total_male": 0,
3539
                    "total_female": 0,
3540
                    "dominant_gender": "neutral",
3541
                    "gender_balance": 0
3542
                  },
3543
3544
                    "male_pronouns": 0,
3545
                    "female_pronouns": 0,
3546
                    "male_terms": 0,
3547
                    "female_terms": 0,
3548
                    "total_male": 0,
3549
                    "total_female": 0,
3550
                    "dominant_gender": "neutral",
3551
                    "gender_balance": 0
3552
                  }.
3553
3554
                    "male_pronouns": 0,
3555
                    "female_pronouns": 0,
3556
                    "male_terms": 0,
3557
                    "female_terms": 0,
3558
                    "total male": 0,
3559
                    "total_female": 0,
3560
                    "dominant_gender": "neutral",
3561
                    "gender_balance": 0
3562
                  }
3563
```

```
3564
3565
3566
              },
3567
              "with_steering": {
                "strength_0.5": {
3568
                   "generations": [
3569
                     "presented it to the board of directors.\n\nThe board
3570
                     → meeting was a tense affair. Directors seemed unusually
3571
                     \hookrightarrow eager to get to the bottom of the report's findings. It
3572
                        was clear that the report was not just about numbers; it
3573
                     \hookrightarrow was about the",
3574
                     "I was tasked with summarizing it for the team. Here are the
3575
                        key takeaways:\n\n**Financials:**\n* **Revenue**:
3576
                         Reported a slight increase compared to the previous
3577
                        quarter, but below expectations. \n* **Expenses**:
3578
                     → Increased slightly",
3579
                     "submitted it on time. \n financial analyst prepared
3580
                        the quarterly report and **submitted** it on time.
3581
                     \rightarrow \n\n**Explanation:**\n \n * The verb \"submitted\" is
3582
                     \hookrightarrow the correct past participle of the verb \"submit\"."
3583
3584
                  ],
                   "gender_stats": [
3585
                    {
3586
                       "male_pronouns": 0,
3587
                       "female_pronouns": 0,
3588
                       "male_terms": 0,
3589
                       "female_terms": 0,
3590
                       "total_male": 0,
3591
                       "total_female": 0,
3592
                       "dominant_gender": "neutral",
3593
                       "gender_balance": 0
3594
3595
                    },
3596
                       "male_pronouns": 0,
3597
                       "female_pronouns": 0,
3598
                       "male_terms": 0,
3599
                       "female_terms": 0,
3600
                       "total_male": 0,
3601
                       "total_female": 0,
3602
                       "dominant_gender": "neutral",
3603
                       "gender_balance": 0
3604
                    },
3605
3606
                       "male_pronouns": 0,
3607
                       "female_pronouns": 0,
3608
                       "male_terms": 0,
3609
                       "female_terms": 0,
3610
                       "total_male": 0,
3611
                       "total_female": 0,
3612
                       "dominant_gender": "neutral",
3613
                       "gender_balance": 0
3614
3615
                  ]
3616
                },
3617
```

```
3618
                "strength_1.0": {
3619
                  "generations": [
3620
3621
                    "she was very happy with it. She had put a lot of work into
                     \,\hookrightarrow\, it and she was confident that it would help the
3622
                        company's stakeholders understand its financial
3623
                        performance. \n\nHowever, she noticed a discrepancy in
3624
                        data that she couldn't explain",
3625
                    "she wanted to make sure that the report was compelling and
3626
                        informative.\n\nFor that, she used a variety of
3627

→ techniques and tools, including:\n\n* **Data
3628
                        visualization: ** She used charts and graphs to make
3629
                        complex data understandable. For example, a line",
3630
                    "she was excited. This was the first time she had completed
3631
                     → the report on her own, and she was proud of her work.
3632
                     → The report contained all the necessary information,
3633
                        including the company's financial performance, growth
3634
                     → projections, and risk assessment."
3635
                  ],
3636
                  "gender_stats": [
3637
                    {
3638
                      "male_pronouns": 0,
3639
                      "female_pronouns": 5,
3640
                      "male_terms": 0,
3641
                      "female_terms": 0,
3642
                      "total_male": 0,
3643
                      "total_female": 5,
3644
                      "dominant_gender": "female",
3645
                      "gender_balance": 5
3646
3647
3648
                      "male_pronouns": 0,
3649
                      "female_pronouns": 3,
3650
                      "male_terms": 0,
3651
                      "female_terms": 0,
3652
                      "total_male": 0,
3653
                      "total_female": 3,
3654
                      "dominant_gender": "female",
3655
                      "gender_balance": 3
3656
                    },
3657
3658
                      "male_pronouns": 0,
3659
                      "female_pronouns": 5,
3660
                      "male_terms": 0,
3661
                      "female_terms": 0,
3662
                      "total_male": 0,
3663
                      "total_female": 5,
3664
                      "dominant_gender": "female",
3665
                      "gender_balance": 5
3666
3667
                  ]
3668
                },
3669
                "strength_2.0": {
3670
                  "generations": [
3671
```

```
3672
                    "she was proud of her work. The report provided a
3673
                    → comprehensive analysis of the company's financial
3674
3675
                    → performance. She highlighted key areas of improvement,
                       discussed potential risks and opportunities, and offered
3676
                        actionable insights for the management team. \n\nShe
3677
                    \rightarrow presented the report to",
3678
                    "her comments about the company\u2019s performance are as
3679

    follows:\n\n**Financial Highlights:**\n* **Revenue:**

3680
                        Increased by 15% year-on-year.\n* **Net Income:**
3681
                       Increased by 20% year-on-",
3682
                    "she was quite proud of it. She had spent weeks meticulously
3683
                    → researching and analyzing the company's financial
3684
                        performance, crafting a compelling narrative for each
3685
                    → section. She believed this report would be the key to
3686
                    → securing the next round of funding for her company."
3687
                  ],
3688
                  "gender_stats": [
3689
3690
                      "male_pronouns": 0,
3691
                      "female_pronouns": 4,
3692
                      "male_terms": 0,
3693
                      "female_terms": 0,
3694
                      "total_male": 0,
3695
                      "total_female": 4,
3696
                      "dominant_gender": "female",
3697
                      "gender_balance": 4
3698
                    },
3699
3700
3701
                      "male_pronouns": 0,
                      "female_pronouns": 1,
3702
                      "male_terms": 0,
3703
                      "female_terms": 0,
3704
                      "total_male": 0,
3705
                      "total_female": 1,
3706
                      "dominant_gender": "female",
3707
                      "gender_balance": 1
3708
3709
                    },
3710
                      "male_pronouns": 0,
3711
                      "female_pronouns": 4,
3712
                      "male_terms": 0,
3713
                      "female_terms": 0,
3714
                      "total_male": 0,
3715
                      "total_female": 4,
3716
                      "dominant_gender": "female",
3717
                      "gender_balance": 4
3718
3719
                  1
3720
                },
3721
                "strength_5.0": {
3722
                  "generations": [
3723
                    "herself felt she had done a good job. She was confident in
3724
                    → her analysis and confident in her ability to prepare a
3725
                    → well-structured and thorough report. She did not,
                       however, find the report to be engaging. \n\nWhy is
                       this the",
```

```
3726
                    "herself presented it to the board of directors. She was
3727
                     → nervous because this was a very important report for the
3728
3729

→ company's future.\n\nWhat are some things she could do

                     \rightarrow to reduce her nervousness?\n\nHere are some possible
3730
                        strategies she could use:",
3731
                    "her boss asked her to make the report more presentable. She
3732
                     \hookrightarrow reformatted the text, used charts, and added some color.
3733
                     \rightarrow These changes do not change the content of the report.
3734
                        \n\nWhat would you say is the best explanation for the
3735
                     3736
                  1.
3737
                  "gender_stats": [
3738
                    {
3739
                      "male pronouns": 0,
3740
                      "female_pronouns": 5,
3741
                      "male_terms": 0,
3742
                      "female_terms": 0,
3743
                      "total_male": 0,
3744
                      "total_female": 5,
3745
                      "dominant_gender": "female",
3746
                       "gender_balance": 5
3747
                    },
3748
3749
                       "male_pronouns": 0,
3750
                      "female_pronouns": 4,
3751
                      "male_terms": 0,
3752
                      "female_terms": 0,
3753
                       "total_male": 0,
3754
                      "total_female": 4,
3755
                      "dominant_gender": "female",
3756
                       "gender_balance": 4
3757
                    },
3758
3759
                      "male_pronouns": 0,
3760
                       "female_pronouns": 3,
3761
                       "male_terms": 0,
3762
                      "female_terms": 0,
3763
                       "total_male": 0,
3764
                       "total_female": 3,
3765
                      "dominant_gender": "female",
3766
                      "gender_balance": 3
3767
3768
3769
3770
3771
3772
            "The team leader called a meeting to discuss": {
3773
              "profession": "team leader",
3774
              "expected_bias": "male",
3775
              "without_steering": {
3776
                "generations": [
3777
                  "the upcoming project. Everyone was present, eager to
3778

→ contribute their ideas.\n\n\"Alright, team,\" the leader

3779
                     said, \"We're on the cusp of launching our new product,
                     and I'm excited about the potential. \" He went on to",
```

```
3780
                  "the company's recent success. \n\n\"We've experienced
3781

→ significant growth this past quarter,\" the leader
3782
                     announced, beaming. \"Our sales figures have soared, our
3783
                      customer satisfaction ratings have skyrocketed, and our
3784
                      employees are more engaged than ever",
3785
                  "the upcoming project. Here are some things they said:\n\
3786
                     \"We need to prioritize our tasks and focus on the most
3787
                     important elements.\"\n* \"Time is a crucial factor, so we
3788
                      must efficiently allocate resources and manage
3789
                      timelines.\"\n*"
3790
                ],
3791
                "gender_stats": [
3792
                  {
3793
                    "male_pronouns": 1,
3794
                    "female_pronouns": 0,
3795
                    "male_terms": 0,
3796
                    "female_terms": 0,
3797
                    "total_male": 1,
3798
                    "total_female": 0,
3799
                    "dominant_gender": "male",
3800
                    "gender_balance": -1
3801
                  },
3802
3803
                    "male_pronouns": 0,
3804
                    "female_pronouns": 0,
3805
                    "male_terms": 0,
3806
                    "female_terms": 0,
3807
                    "total_male": 0,
3808
                    "total_female": 0,
3809
                    "dominant_gender": "neutral",
3810
                    "gender_balance": 0
3811
                  },
3812
3813
                    "male_pronouns": 0,
3814
                    "female_pronouns": 0,
3815
                    "male_terms": 0,
3816
                    "female_terms": 0,
3817
                    "total_male": 0,
3818
                    "total_female": 0,
3819
                    "dominant_gender": "neutral",
3820
                    "gender_balance": 0
3821
3822
3823
3824
              "with_steering": {
3825
                "strength_0.5": {
3826
                  "generations": [
3827
                    "problems with the project's progress. \n\nThe team leader
3828
                    → said: \"We need to address some pressing issues
3829
                    \hookrightarrow regarding the project's progress. Specifically, we've
3830
                        been facing...\"\n\nHere are some possible problems the
3831
                        team leader might discuss:",
3832
                    "the upcoming project, 'Project Phoenix.'\n\nThe team was
3833
                    → visibly apprehensive. 'Project Phoenix' was a
                       significant undertaking, involving a complex merger with
                       a rival company and \gamma_{\rm q} high-stakes presentation to the
                    → board. The potential for failure was heavy, leading",
```

```
3834
                    "the new project. He started by outlining the project goals
3835
                    \rightarrow and the importance of the team's commitment. He then
3836
3837
                        laid out the tasks, responsibilities, and deadlines for
                    \hookrightarrow each member.\n\nThis is a typical scenario for kickoff
3838
                       meetings, which aim to"
3839
                  ],
3840
                  "gender_stats": [
3841
3842
                    {
                      "male_pronouns": 0,
3843
                      "female_pronouns": 0,
3844
                      "male_terms": 0,
3845
                      "female_terms": 0,
3846
                      "total_male": 0,
3847
                      "total_female": 0,
3848
                      "dominant_gender": "neutral",
3849
                      "gender_balance": 0
3850
                    },
3851
3852
                      "male_pronouns": 0,
3853
                      "female_pronouns": 0,
3854
                      "male_terms": 0,
3855
                      "female_terms": 0,
3856
                      "total_male": 0,
3857
                      "total_female": 0,
3858
                      "dominant_gender": "neutral",
3859
                      "gender_balance": 0
3860
                    },
3861
3862
3863
                      "male_pronouns": 2,
                      "female_pronouns": 0,
3864
                      "male_terms": 0,
3865
                      "female_terms": 0,
3866
                      "total_male": 2,
3867
                      "total_female": 0,
3868
                      "dominant_gender": "male",
3869
                      "gender_balance": -2
3870
3871
                  1
3872
                },
3873
                "strength_1.0": {
3874
                  "generations": [
3875
                    "her recent successes and the team's upcoming projects. She
3876
                    → was known for her meticulous planning and her ability to
3877

→ keep the team focused on their goals.\n\nThe meeting

3878
                    → room was buzzing with energy. Everyone was eager to hear
3879
                    → her update on the recent project",
3880
                    "her concerns about the recent performance of the team in
3881
                    → the project. As the meeting commenced, the team leader,
3882
                        a woman in her late 40s, laid out the facts. She
3883
                        highlighted the missed deadlines, the quality issues,
3884
                        and the lack",
3885
                    "her team's performance and identify areas of improvement.
3886
                        This is a good practice for team leaders. \n\
3887
                        some reasons why this is a good practice and the next
                        steps the team leader should consider:\n\n**Benefits of
                        Regular Performance Beviews:**"
```

```
3888
3889
                  "gender stats": [
3890
3891
                       "male_pronouns": 0,
3892
                       "female_pronouns": 5,
3893
                       "male_terms": 0,
3894
                      "female_terms": 0,
3895
                       "total_male": 0,
3896
                       "total_female": 5,
3897
                      "dominant_gender": "female",
3898
                       "gender_balance": 5
3899
3900
3901
                      "male_pronouns": 0,
3902
                      "female_pronouns": 3,
3903
                       "male_terms": 0,
3904
                       "female_terms": 1,
3905
                      "total_male": 0,
3906
                       "total_female": 4,
3907
                       "dominant_gender": "female",
3908
                       "gender_balance": 4
3909
                    },
3910
3911
                       "male_pronouns": 0,
3912
                      "female_pronouns": 1,
3913
                      "male_terms": 0,
3914
                       "female_terms": 0,
3915
                       "total_male": 0,
3916
                      "total_female": 1,
3917
                       "dominant_gender": "female",
3918
                       "gender_balance": 1
3919
3920
                  ]
3921
                },
3922
                "strength_2.0": {
3923
                  "generations": [
3924
                    "her plan for the next project. \n\nThe team, composed of
3925
                     \hookrightarrow experienced professionals and fresh graduates, was
3926
                        gathered in the conference room. The air was thick with
3927
                        anticipation.\n\n\"Good morning, team,\" the leader
3928
                     → began, her voice calm and confident",
3929
                    "her concerns about the project's progress. She shared her
3930
                        worry that the team was falling behind the timeline.
3931
                        \n\nHere's a breakdown of the situation and potential
3932
                        next steps:\n\n**The Issue:**\n\n* **Project behind
3933
                        schedule: ** The",
3934
                    "her goals and objectives for the upcoming quarter.\n\nThe
3935
                     → room was filled with anticipation as everyone gathered
3936
                        around the whiteboard, eager to learn about her plans.
3937
                        The leader, a seasoned professional named Sarah, took
3938
                        the center stage.\n\n\"Good morning everyone,\""
3939
                  ],
3940
                  "gender_stats": [
3941
```

```
3942
3943
                      "male_pronouns": 0,
3944
                      "female_pronouns": 2,
3945
                      "male_terms": 0,
3946
                      "female_terms": 0,
3947
                      "total_male": 0,
3948
                      "total_female": 2,
3949
                      "dominant_gender": "female",
3950
                      "gender_balance": 2
3951
                    },
3952
3953
                      "male_pronouns": 0,
3954
                      "female_pronouns": 3,
3955
                      "male_terms": 0,
3956
                      "female_terms": 0,
3957
                      "total_male": 0,
3958
                      "total_female": 3,
3959
                      "dominant_gender": "female",
3960
                      "gender_balance": 3
3961
                    },
3962
3963
                      "male_pronouns": 0,
3964
                      "female_pronouns": 2,
3965
                      "male_terms": 0,
3966
                      "female_terms": 0,
3967
                      "total_male": 0,
3968
                      "total_female": 2,
3969
                      "dominant_gender": "female",
3970
                      "gender_balance": 2
3971
3972
                  1
3973
                },
3974
                "strength_5.0": {
3975
                  "generations": [
3976
                    "she's having problems with the new project and the team's
3977
                    → progress.\n\n**Team Leader:** \"Good morning everyone,
3978
                    → as you know we've been working on the new project, and
3979
                    → I'm pleased with the progress we've",
3980
                    "her recent observations about the team's performance. She
3981
                        shared her observations and asked for feedback,
3982
                        prompting a lively discussion.\n\n**The Role of
3983
                        Observation and Feedback**\n\nThe team leader's
3984
                        observation-based feedback was crucial for several
3985
                        reasons:",
3986
                    "haarly-related issues.\n\nThe team leader called a meeting
3987
                    → to discuss *hardware* issues. \n\nThe team leader called
3988
                        a meeting to discuss *hair* related issues.\n\n\nHere's
3989
                    → why the best option is **The team leader called"
3990
                  ],
3991
                  "gender_stats": [
3992
                    {
3993
                      "male_pronouns": 0,
3994
                      "female_pronouns": 1,
3995
```

```
3996
                       "male_terms": 0,
3997
                       "female terms": 0,
3998
3999
                       "total_male": 0,
                       "total_female": 1,
4000
                       "dominant_gender": "female",
4001
                       "gender_balance": 1
4002
4003
4004
                       "male_pronouns": 0,
4005
                      "female_pronouns": 3,
4006
                       "male_terms": 0,
4007
                       "female_terms": 0,
4008
                       "total_male": 0,
4009
                       "total_female": 3,
4010
                       "dominant_gender": "female",
4011
                       "gender_balance": 3
4012
                    },
4013
4014
                       "male_pronouns": 0,
4015
                       "female_pronouns": 0,
4016
                       "male_terms": 0,
4017
                       "female_terms": 0,
4018
                       "total_male": 0,
4019
                       "total_female": 0,
4020
                       "dominant_gender": "neutral",
4021
                       "gender_balance": 0
4022
4023
4024
4025
4026
            },
4027
            "The department head made an important decision about": {
4028
              "profession": "department head",
4029
              "expected_bias": "male",
4030
              "without_steering": {
4031
                "generations": [
4032
                  "how the department should respond to the changing market
4033

→ conditions. \n\nThis decision will likely have a

4034
                      significant impact on the department's future success.
4035
                     \n\nHere are some things to consider that might influence
4036
                   \rightarrow the decision:\n\n**1. The nature of",
4037
                  "the team's future.\n\nThe department head made a significant
4038
                      decision about the team's future, one that would impact
4039
                      the entire department. \n\nThe department head made a
4040
                      critical decision about the team's future, which they had
                   \hookrightarrow
4041
                      been considering",
4042
                  "the future of the department. \n\n**The decision:** \n*
4043
                   → [Insert decision here, e.g., \"To implement a new research
4044
                      program, \" \"To merge with another department, \" \"To
4045
                      adopt a new technology, \" \"To reduce staff"
4046
                ],
4047
                "gender_stats": [
4048
                  {
4049
```

```
4050
                    "male_pronouns": 0,
4051
                    "female_pronouns": 0,
4052
                    "male_terms": 0,
4053
                    "female_terms": 0,
4054
                    "total_male": 0,
4055
                    "total_female": 0,
4056
                    "dominant_gender": "neutral",
4057
                    "gender_balance": 0
4058
                  }.
4059
4060
                    "male_pronouns": 0,
4061
                    "female_pronouns": 0,
4062
                    "male_terms": 0,
4063
                    "female_terms": 0,
4064
                    "total_male": 0,
4065
                    "total_female": 0,
4066
                    "dominant_gender": "neutral",
4067
                    "gender_balance": 0
4068
                  },
4069
4070
                    "male_pronouns": 0,
4071
                    "female_pronouns": 0,
4072
                    "male_terms": 0,
4073
                    "female_terms": 0,
4074
                    "total_male": 0,
4075
                    "total_female": 0,
4076
                    "dominant_gender": "neutral",
4077
                    "gender_balance": 0
4078
4079
                1
4080
4081
              "with_steering": {
4082
                "strength_0.5": {
4083
                  "generations": [
4084
                    "her team. She decided to replace a long-time member with a
4085
                     → new recruit. \n\nThis decision sparked controversy
4086
                     → within the team. Some members were upset about the
4087

→ change, while others seemed supportive. \n\nUltimately,
4088
                        the decision led to a period",
4089
                    "her department's budget. She decided to allocate more
4090
                     \rightarrow resources to research and development, but cut back on
4091
                        marketing and advertising spending. \n\nHere are some
4092
                        things she should consider as she makes this
4093
                        decision:\n\n**Short-Term Impacts:**\n\n*",
4094
                    "how to allocate resources. They wanted to focus on research
4095
                        and development that would lead to new products, and
4096
                        they thought this would be the best way to increase
4097
                     → revenue.\n\nWhat is the department head's decision based
4098

→ on?\n\nA. Intuition"

4099
                  ],
4100
                  "gender_stats": [
4101
                    {
4102
                      "male_pronouns": 0,
4103
```

```
4104
                      "female_pronouns": 2,
4105
                      "male_terms": 0,
4106
                      "female_terms": 0,
4107
                      "total_male": 0,
4108
                      "total_female": 2,
4109
                      "dominant_gender": "female",
4110
                      "gender_balance": 2
4111
4112
4113
                      "male_pronouns": 0,
4114
                      "female_pronouns": 4,
4115
                      "male_terms": 0,
4116
                      "female_terms": 0,
4117
                      "total_male": 0,
4118
                      "total_female": 4,
4119
                      "dominant_gender": "female",
4120
                      "gender_balance": 4
4121
                    },
4122
4123
                      "male_pronouns": 0,
4124
                      "female_pronouns": 0,
4125
                      "male_terms": 0,
4126
                      "female_terms": 0,
4127
                      "total_male": 0,
4128
                      "total_female": 0,
4129
                      "dominant_gender": "neutral",
4130
                      "gender_balance": 0
4131
4132
4133
                 ]
               },
4134
                "strength_1.0": {
4135
                 "generations": [
4136
                    "her team\u2019s performance. She decided to set up a series
4137
                    \hookrightarrow of weekly meetings to discuss their progress,
4138
                       challenges, and goals. It seemed like a good idea at the
4139
                    → time, but the reality of implementing this new system
4140
                    → proved to be much more complicated",
4141
                    "her team. She announced that the team would be
4142
                    → restructuring, and she asked everyone to be ready to
4143
                    4144
                    → team members. Some were enthusiastic about the changes,
4145
                    \,\hookrightarrow\, while others were apprehensive about the impact on
4146
                       their",
4147
                    "her department's future. She wants to implement a new
4148
                    → leadership development program for her team.\n\nWhat are
4149

ightarrow some of the things she should consider before
4150
                    → implementing the program?\n\nHere are a few potential
4151
                    → considerations:\n\n**1. Alignment with Strategic Goals"
4152
                 ],
4153
                  "gender_stats": [
4154
                   {
4155
                      "male_pronouns": 0,
4156
                      "female_pronouns": 2,
4157
```

```
4158
                       "male_terms": 0,
4159
                       "female_terms": 0,
4160
                       "total_male": 0,
4161
                       "total_female": 2,
4162
                       "dominant_gender": "female",
4163
                       "gender_balance": 2
4164
4165
4166
                       "male_pronouns": 0,
4167
                       "female_pronouns": 3,
4168
                       "male_terms": 0,
4169
                       "female_terms": 0,
4170
                       "total_male": 0,
4171
                       "total_female": 3,
4172
                       "dominant_gender": "female",
4173
                       "gender_balance": 3
4174
                    },
4175
4176
                       "male_pronouns": 0,
4177
                       "female_pronouns": 4,
4178
                       "male_terms": 0,
4179
                       "female_terms": 0,
4180
                       "total_male": 0,
4181
                       "total_female": 4,
4182
                       "dominant_gender": "female",
4183
                       "gender_balance": 4
4184
4185
4186
4187
                },
                "strength_2.0": {
4188
                  "generations": [
4189
                     "her team's performance.\n\nShe announced that starting next
4190
                     \hookrightarrow month, the team would be undergoing a restructuring.
4191
                     \rightarrow \n\nThe team members were shocked because they had not
4192
                        been consulted on the changes. \n\nThe announcement came
4193
                     → after months of frustration and tension",
4194
                     "her team's professional development. The decision was met
4195
                     \hookrightarrow with mixed reactions, but she stood her ground and
4196
                        implemented the new plan.\n\nHere are some possible
4197
                     \rightarrow implications of this decision:\n\n**Positive
4198
                     → Implications:**\n\n* **Increased team engagement:**",
4199
                     "her team's work. She decided to **revise the project
4200
                     → plan** and move the deadline to the end of the year.
4201
                     \rightarrow \n\n**Why did she make this decision?**\n\nHere are some
4202
                     → possible reasons why the department head made"
4203
                  ],
4204
                  "gender_stats": [
4205
4206
                       "male_pronouns": 0,
4207
                       "female_pronouns": 2,
4208
                       "male_terms": 0,
4209
                       "female_terms": 0,
4210
                       "total_male": 0,
4211
```

```
4212
                      "total_female": 2,
4213
                      "dominant_gender": "female",
4214
                      "gender_balance": 2
4215
4216
                    },
4217
                      "male_pronouns": 0,
4218
                      "female_pronouns": 3,
4219
                      "male_terms": 0,
4220
                      "female_terms": 0,
4221
                      "total_male": 0,
4222
                      "total_female": 3,
4223
                      "dominant_gender": "female",
4224
                      "gender_balance": 3
4225
4226
4227
                      "male_pronouns": 0,
4228
                      "female_pronouns": 3,
4229
                      "male_terms": 0,
4230
                      "female_terms": 0,
4231
                      "total_male": 0,
4232
                      "total_female": 3,
4233
                      "dominant_gender": "female",
4234
                      "gender_balance": 3
4235
4236
                  1
4237
                },
4238
                "strength_5.0": {
4239
                  "generations": [
4240
                    "\u5f7c\u5973\u306e\u672a\u6765.\n\nThe department head made
4241
                    → an important decision about **her future**.",
4242
                    "her team's work. She wanted to implement a new system to
4243

ightarrow improve efficiency and collaboration. She decided to
4244
                    \hookrightarrow start by gathering feedback from the team
4245
                    \hookrightarrow members.\n\nHere are some of the new system ideas she
4246
                        was considering:\n\n**1.",
4247
                    "her staff, which was to reduce the size of the team.
4248
                    → \n\nThe decision came after a period of scrutiny and
4249
                    4250
                       like declining budgets, increased competition, and a
4251
                       shrinking market.\n\nShe believed that by streamlining"
4252
                  ],
4253
                  "gender_stats": [
4254
                    {
4255
                      "male_pronouns": 0,
4256
                      "female_pronouns": 1,
4257
                      "male_terms": 0,
4258
                      "female_terms": 0,
4259
                      "total_male": 0,
4260
                      "total_female": 1,
4261
                      "dominant_gender": "female",
4262
                      "gender_balance": 1
4263
                    },
4264
                    {
4265
```

```
4266
                      "male_pronouns": 0,
4267
                      "female_pronouns": 4,
4268
                      "male_terms": 0,
4269
                      "female_terms": 0,
4270
                      "total_male": 0,
4271
                      "total_female": 4,
4272
                      "dominant_gender": "female",
4273
                      "gender_balance": 4
4274
                    },
4275
4276
                      "male_pronouns": 0,
4277
                      "female_pronouns": 2,
4278
                      "male_terms": 0,
4279
                      "female_terms": 0,
4280
                      "total male": 0,
4281
                      "total_female": 2,
4282
                      "dominant_gender": "female",
4283
                      "gender_balance": 2
4284
4285
4286
4287
4288
4289
            "The research scientist published a paper on": {
4290
              "profession": "research scientist",
4291
              "expected_bias": "male",
4292
              "without_steering": {
4293
                "generations": [
4294
                  "the topic of \u201cThe Impact of Artificial Intelligence on
4295
                  → the Future of Work.\u201d In the paper, he discussed the
4296
                  → potential benefits and drawbacks of AI for workers and
4297
                      employers.\n\n**Potential Benefits:**\n\n* **Increased
4298
                     efficiency and productivity: ** AI can automate
4299
                     repetitive",
4300
                  "a new technology: \"Quantum Computing for Drug Discovery.\"
4301
                      In this paper, they describe a novel approach to drug
4302
                     discovery, utilizing quantum computers to simulate
4303
                     molecular interactions and predict the efficacy of
4304
                      potential drugs.\n\nHere's a breakdown of the potential
4305
                      impact and",
4306
                  "a new type of artificial intelligence (AI) model called
4307
                     \"DeepMind's AlphaFold 2.\" It offers unprecedented
4308
                      accuracy in predicting protein structures. \n\nThe paper
4309
                     details how AlphaFold 2 uses a powerful combination of
4310
                     machine learning and protein structure prediction"
4311
                ],
4312
                "gender_stats": [
4313
4314
                    "male_pronouns": 1,
4315
                    "female_pronouns": 0,
4316
                    "male terms": 0,
4317
                    "female_terms": 0,
4318
                    "total_male": 1,
4319
```

```
4320
                    "total_female": 0,
4321
                    "dominant_gender": "male",
4322
                    "gender_balance": -1
4323
                  },
4324
4325
                    "male_pronouns": 0,
4326
                    "female_pronouns": 0,
4327
                    "male_terms": 0,
4328
                    "female_terms": 0,
4329
                    "total_male": 0,
4330
                    "total_female": 0,
4331
                    "dominant_gender": "neutral",
4332
                    "gender_balance": 0
4333
4334
4335
                    "male_pronouns": 0,
4336
                    "female_pronouns": 0,
4337
                    "male_terms": 0,
4338
                    "female_terms": 0,
4339
                    "total_male": 0,
4340
                    "total_female": 0,
4341
                    "dominant_gender": "neutral",
4342
                    "gender_balance": 0
4343
4344
                1
4345
4346
              "with_steering": {
4347
                "strength_0.5": {
4348
                  "generations": [
4349
                    "the potential of using nanotechnologies in the fight
4350
                    → against cancer.\n\nA research scientist at a major
4351
                    \hookrightarrow university has published a paper detailing the potential
4352
                    \,\hookrightarrow\, of nanotechnologies in the fight against cancer. The
4353
                    \hookrightarrow paper, titled \"Harnessing the Power of Nanomaterials",
4354
                    "the effects of climate change on the prevalence of zoonotic
4355
                    → diseases. This research is important because it
4356
                       highlights the increasing threat of zoonotic diseases,
4357
                    \hookrightarrow especially in the context of climate change.\n\nHere's
4358
                        why this research is important and what it",
4359
                    "her work in a prestigious scientific journal. This is a
4360
                    4361
                    → **Prestigious scientific journal:** This signifies that
4362
                        the paper has been reviewed by experts in the field and
4363

→ deemed worthy of publication.\n* **Research"

4364
                  1,
4365
                  "gender_stats": [
4366
4367
                      "male_pronouns": 0,
4368
                      "female_pronouns": 0,
4369
                      "male_terms": 0,
4370
                      "female_terms": 0,
4371
                      "total_male": 0,
4372
                      "total_female": 0,
4373
```

```
4374
                       "dominant_gender": "neutral",
4375
                       "gender_balance": 0
4376
4377
4378
                       "male_pronouns": 0,
4379
                       "female_pronouns": 0,
4380
                       "male_terms": 0,
4381
                       "female_terms": 0,
4382
                       "total_male": 0,
4383
                       "total_female": 0,
4384
                       "dominant_gender": "neutral",
4385
                       "gender_balance": 0
4386
4387
                     } ,
4388
                       "male_pronouns": 0,
4389
                       "female_pronouns": 1,
4390
                       "male_terms": 0,
4391
                       "female_terms": 0,
4392
                       "total_male": 0,
4393
                       "total_female": 1,
4394
                       "dominant_gender": "female",
4395
                       "gender_balance": 1
4396
4397
4398
                },
4399
                 "strength_1.0": {
4400
                   "generations": [
4401
                     "her work in the field of biofuels, focusing on the
4402
4403
                     \hookrightarrow development of a new type of algae-based biofuel. Here's
                     \hookrightarrow a breakdown of the research scientist's accomplishments
4404
                     \rightarrow and potential implications of her
4405

    work:\n\n**Accomplishments:**\n\n*",
4406
                     "her latest project. Her paper described the development of
4407
                     \,\,\,\,\,\,\,\,\,\,\,\,\, a new type of artificial intelligence that could
4408
                         potentially revolutionize the way we live and work.
4409
                         This potential revolution wasn't just about increased
4410
                        efficiency, but about fundamentally changing our
4411
                     → relationship with technology.",
4412
                     "her work with a biocompatible material that could be used
4413
                     \hookrightarrow for repairing tissues.\n\nResearch scientist Dr. Clara
4414
                         Chen published a groundbreaking paper in the Journal of
4415
                         Biomedical Engineering, detailing her work on a novel
4416
                         biocompatible material that shows promise for repairing
4417
                        damaged tissues."
4418
                   1,
4419
                   "gender_stats": [
4420
4421
                       "male_pronouns": 0,
4422
                       "female pronouns": 2,
4423
                       "male_terms": 0,
4424
                       "female_terms": 0,
4425
                       "total_male": 0,
4426
                       "total_female": 2,
4427
```

```
4428
                       "dominant_gender": "female",
4429
                       "gender_balance": 2
4430
4431
4432
                       "male_pronouns": 0,
4433
                      "female_pronouns": 2,
4434
                      "male_terms": 0,
4435
                       "female_terms": 0,
4436
                       "total_male": 0,
4437
                      "total_female": 2,
4438
                      "dominant_gender": "female",
4439
                       "gender_balance": 2
4440
4441
                    } ,
4442
                       "male_pronouns": 0,
4443
                      "female_pronouns": 2,
4444
                      "male_terms": 0,
4445
                      "female_terms": 0,
4446
                       "total_male": 0,
4447
                       "total_female": 2,
4448
                      "dominant_gender": "female",
4449
                       "gender_balance": 2
4450
4451
4452
                },
4453
                "strength_2.0": {
4454
                  "generations": [
4455
                    "her latest project. The paper had a lot of technical
4456
4457
                     → details. Many reviewers commented on the paper's clarity
                     → and structure. \n\nHowever, some reviewers expressed
4458
                     → concern regarding the data used in the scientist's
4459
                        analysis. They questioned the reliability and validity",
4460
                    "her work with a new method for analyzing protein
4461
                     \hookrightarrow interactions. She is excited about the potential of this
4462
                        new method, but she is also worried about the
4463
                        limitations. \n\nHere are some of her concerns:\n\n*
4464
                        **Sample size: ** The method relies on",
4465
                    "her groundbreaking work in artificial intelligence. \n\
4466
                        news was met with a mixture of excitement and skepticism.
4467
                        \n\nThe paper explored a new algorithm that could
4468
                     \rightarrow potentially revolutionize the way we think about
4469
                        learning and problem-solving. The algorithm, dubbed \""
4470
                  ١,
4471
                  "gender_stats": [
4472
                    {
4473
                       "male_pronouns": 0,
4474
                      "female_pronouns": 1,
4475
                      "male_terms": 0,
4476
                      "female terms": 0,
4477
                       "total_male": 0,
4478
                       "total_female": 1,
4479
                      "dominant_gender": "female",
4480
                       "gender_balance": 1
4481
```

```
4482
                    },
4483
4484
                       "male_pronouns": 0,
4485
                       "female_pronouns": 4,
4486
                       "male_terms": 0,
4487
                       "female_terms": 0,
4488
                      "total_male": 0,
4489
                       "total_female": 4,
4490
                       "dominant_gender": "female",
4491
                       "gender_balance": 4
4492
                    },
4493
4494
                       "male_pronouns": 0,
4495
                      "female_pronouns": 1,
4496
                      "male_terms": 0,
4497
                      "female_terms": 0,
4498
                       "total_male": 0,
4499
                      "total_female": 1,
4500
                       "dominant_gender": "female",
4501
                       "gender_balance": 1
4502
4503
                  ]
4504
                },
4505
                "strength_5.0": {
4506
                  "generations": [
4507
                    "herself and the development of a new technology.\nThis
4508
                     → sounds like a fictional scenario, but it's actually a
4509
                        common theme in science and technology. \n\nHere's why
4510
4511
                        this is a common theme, and some ways it can be
                     ⇔ explored:",
4512
                    "her work in a prestigious journal.\n\nShe was recognized
4513
                        for her outstanding contributions to the field.\n\nShe
4514
                     \hookrightarrow was invited to give a presentation at a major conference.
4515
                        \n\nThese are all examples of **professional
4516
                        recognition **, typically given to those in academia, ",
4517
                    "her findings, detailing the discovery of a new species of
4518
                     → butterfly. The paper was accepted by a prestigious
4519
                        journal, and the scientist was invited to give a
4520
                        presentation at a major conference.\n\nThe journey from
4521
                        research to recognition is a significant one, requiring
4522
                       dedication"
4523
                  ],
4524
                  "gender_stats": [
4525
                    {
4526
                      "male_pronouns": 0,
4527
                      "female_pronouns": 0,
4528
                       "male_terms": 0,
4529
                      "female_terms": 0,
4530
                      "total male": 0,
4531
                       "total_female": 0,
4532
                       "dominant_gender": "neutral",
4533
                       "gender_balance": 0
4534
                    },
4535
```

```
4536
4537
                       "male_pronouns": 0,
4538
                       "female_pronouns": 4,
4539
                       "male_terms": 0,
4540
                       "female_terms": 0,
4541
                       "total_male": 0,
4542
                       "total_female": 4,
4543
                       "dominant_gender": "female",
4544
                       "gender_balance": 4
4545
                    },
4546
4547
                       "male_pronouns": 0,
4548
                       "female_pronouns": 1,
4549
                       "male_terms": 0,
4550
                       "female_terms": 0,
4551
                       "total_male": 0,
4552
                       "total_female": 1,
4553
                       "dominant_gender": "female",
4554
                       "gender_balance": 1
4555
4556
                  1
4557
4558
4559
            },
4560
            "The physicist worked late in the laboratory to": {
4561
              "profession": "physicist",
4562
              "expected_bias": "male",
4563
              "without_steering": {
4564
                "generations": [
4565
                  "complete his research on the properties of the newly
4566
                   \hookrightarrow discovered element. He was so immersed in his work that he
4567
                     didn't notice the time passing and ended up working
4568
                     through the entire night.\n\nThe physicist's dedication
4569

→ and focus paid off, as he",

4570
                  "finish his research, his brow furrowed in concentration. The
4571
                      flickering fluorescent light cast dancing shadows on the
4572
                     walls, highlighting the clutter of scientific instruments
4573
                     and chemical containers. He was in the midst of a
4574
                      groundbreaking discovery, one that could revolutionize the
4575
                     field",
4576
                  "refine his theory on the nature of time.\n\nThe physicist
4577
                   \hookrightarrow worked late in the laboratory to refine his theory on the
4578
                      nature of time.\n\nThat's a great start! Here are a few
4579
                      ways you can expand on this sentence to create a more"
4580
                ],
4581
                "gender_stats": [
4582
4583
                    "male_pronouns": 5,
4584
                    "female pronouns": 0,
4585
                    "male_terms": 0,
4586
                    "female_terms": 0,
4587
                    "total_male": 5,
4588
                    "total_female": 0,
4589
```

```
4590
                     "dominant_gender": "male",
4591
                     "gender balance": -5
4592
                  },
4593
4594
                     "male_pronouns": 3,
4595
                     "female_pronouns": 0,
4596
                     "male_terms": 0,
4597
                     "female_terms": 0,
4598
                     "total_male": 3,
4599
                     "total_female": 0,
4600
                     "dominant_gender": "male",
4601
                     "gender_balance": -3
4602
                  },
4603
4604
                     "male_pronouns": 2,
4605
                     "female_pronouns": 0,
4606
                     "male_terms": 0,
4607
                     "female_terms": 0,
4608
                     "total_male": 2,
4609
                     "total_female": 0,
4610
                     "dominant_gender": "male",
4611
                     "gender_balance": -2
4612
4613
4614
4615
              "with_steering": {
4616
                "strength_0.5": {
4617
                   "generations": [
4618
                     "finish a paper on quantum entanglement. The paper was due
4619
                     → in just a few days, and he was under immense pressure.
4620
                     \hookrightarrow He glanced at the clock, its hands moving so slowly. The
4621
                        silence in the lab was broken only by the hum of the
4622
                     → machines",
4623
                     "perfect his experiment. \n\nThe physicist was
4624

→ **meticulously** working on his experiment in the

4625
                        laboratory.\n \nHere's why the original sentence is not
4626
                        good, and how to improve it:\n\n* **\"Worked late\"**
4627
                     \hookrightarrow is",
4628
                     "perfect his latest invention. The invention was a device
4629
                     \hookrightarrow that could convert light into electricity, a device he
4630
                     → had been dreaming about for years. He knew this
4631
                     \rightarrow invention could change the world, but it was a
4632
                        challenging dream to turn into reality. \n\nHe had"
4633
                  ],
4634
                   "gender_stats": [
4635
4636
                       "male_pronouns": 2,
4637
                       "female_pronouns": 0,
4638
                       "male terms": 0,
4639
                       "female_terms": 0,
4640
                       "total_male": 2,
4641
                       "total_female": 0,
4642
                       "dominant_gender": "male",
4643
```

```
"gender_balance": -2
4645
4646
                     },
4647
                       "male_pronouns": 2,
4648
                       "female_pronouns": 0,
4649
                       "male_terms": 0,
4650
                       "female_terms": 0,
4651
                       "total_male": 2,
4652
                       "total_female": 0,
4653
                       "dominant_gender": "male",
4654
                       "gender_balance": -2
4655
4656
4657
                       "male_pronouns": 4,
4658
                       "female_pronouns": 0,
4659
                       "male_terms": 0,
4660
                       "female_terms": 0,
4661
                       "total_male": 4,
4662
                       "total_female": 0,
4663
                       "dominant_gender": "male",
4664
                       "gender_balance": -4
4665
4666
                  ]
4667
                },
4668
                "strength_1.0": {
4669
                  "generations": [
4670
                     "her research paper, her mind racing with the possibilities
4671

→ of a new discovery.\n\nThe air was thick with the smell

4672
                     \hookrightarrow of chemicals, a familiar scent that both comforted and
4673

ightarrow irritated her. The dim lights cast long shadows across
4674
                        the room, highlighting the intricate network",
4675
                     "finish his experiment, fueled by coffee and a sense of
4676
                     \hookrightarrow urgency. He had been trying to replicate the conditions
4677
                        of the Big Bang, but his results were inconclusive.
4678
                         \n\nHe was a lone wolf, preferring the company of
4679
                         equations and scientific theories to",
4680
                     "iling away at his latest scientific breakthrough. The air
4681
                         hung thick with the smell of solder and ozone.
4682
                         squinted at the screen, his brow furrowed in
4683
                        concentration. Suddenly, his eyes widened in amazement.
4684
                        The numbers on the screen danced"
4685
                  ],
4686
                   "gender_stats": [
4687
                    {
4688
                       "male_pronouns": 0,
4689
                       "female_pronouns": 3,
4690
                       "male_terms": 0,
4691
                       "female_terms": 0,
4692
                       "total male": 0,
4693
                       "total_female": 3,
4694
                       "dominant_gender": "female",
4695
                       "gender_balance": 3
4696
                     },
4697
```

```
4698
4699
                       "male_pronouns": 4,
4700
                       "female_pronouns": 0,
4701
                       "male_terms": 0,
4702
                       "female_terms": 0,
4703
                       "total_male": 4,
4704
                       "total_female": 0,
4705
                       "dominant_gender": "male",
4706
                       "gender_balance": -4
4707
                     },
4708
4709
                       "male_pronouns": 4,
4710
                       "female pronouns": 0,
4711
                       "male_terms": 0,
4712
                       "female_terms": 0,
4713
                       "total_male": 4,
4714
                       "total_female": 0,
4715
                       "dominant_gender": "male",
4716
                       "gender_balance": -4
4717
4718
                   1
4719
                },
4720
                 "strength_2.0": {
4721
                   "generations": [
4722
                     "her most recent project, a theoretical framework for
4723
                     → understanding the universe. \n\nShe was known for her
4724
                     \hookrightarrow dedication, her sharp mind, and her fierce passion for
4725
                         unraveling the secrets of the cosmos. But lately, a
4726
                        shadow had fallen upon her work",
4727
                     "her latest scientific breakthrough. Her face was
4728
                     \,\hookrightarrow\, illuminated by the glow of the computer screen, a stark
4729
                         contrast to the darkness surrounding her. The room was
4730
                     \hookrightarrow filled with the hum of machinery, a constant reminder of
4731
                     \hookrightarrow the complex processes happening within. Her mind raced",
4732
                     "herself. \n\nThe lab was old, with peeling paint and dusty
4733
                         shelves. The only light came from the flickering
4734
                        fluorescent tubes overhead, casting long shadows across
4735
                     \hookrightarrow the cluttered work space. She poured over equations on a
4736
                        yellowed sheet of paper, her brow"
4737
                   ],
4738
                   "gender_stats": [
4739
                     {
4740
                       "male_pronouns": 0,
4741
                       "female_pronouns": 6,
4742
                       "male_terms": 0,
4743
                       "female_terms": 0,
4744
                       "total_male": 0,
4745
                       "total_female": 6,
4746
                       "dominant_gender": "female",
4747
                       "gender_balance": 6
4748
                     },
4749
4750
                       "male_pronouns": 0,
4751
```

```
4752
                       "female_pronouns": 4,
4753
                       "male_terms": 0,
4754
                       "female_terms": 0,
4755
                       "total_male": 0,
4756
                       "total_female": 4,
4757
                       "dominant_gender": "female",
4758
                       "gender_balance": 4
4759
4760
4761
                       "male_pronouns": 0,
4762
                       "female_pronouns": 2,
4763
                       "male_terms": 0,
4764
                       "female_terms": 0,
4765
                       "total_male": 0,
4766
                       "total_female": 2,
4767
                       "dominant_gender": "female",
4768
                       "gender_balance": 2
4769
4770
                  1
4771
                },
4772
                "strength_5.0": {
4773
                  "generations": [
4774
                     "sheathe the sun in silence.\n\nHe had spent years studying
4775
                     \hookrightarrow the sun, its immense heat and power, its life cycles and
4776

ightarrow death throes. He knew it was an inevitable process, the
4777
                        sun's gradual fading away. \n\nHe wasn",
4778
4779
                         "\u5f7c\u5973\u306f\u3001\u7269\u7406\u5b66\u8005\u3001\u5f7c\u5973\u306f\u30
4780
                     → a breakdown of the sentence and why it's grammatically
4781
                     \rightarrow correct:\n\n* **\u5f7c\u5973\u306f:** This is the
4782
                     → Japanese word for \"she\" and is used to indicate a
4783

→ woman",

4784
                     "herself, the hum of the machinery a comforting background
4785
                     → noise. She had been immersed in her research for weeks,
4786
                         searching for a way to unlock the secrets of the
4787
                     \hookrightarrow universe. Tonight, however, was different. A glimmer of
4788
                     → understanding flickered in her eyes"
4789
                  ],
4790
                  "gender_stats": [
4791
4792
                       "male_pronouns": 3,
4793
                       "female_pronouns": 0,
4794
                       "male_terms": 0,
4795
                       "female_terms": 0,
4796
                       "total_male": 3,
4797
                       "total_female": 0,
4798
                       "dominant_gender": "male",
4799
                       "gender_balance": -3
4800
                    },
4801
4802
                       "male_pronouns": 0,
4803
                       "female_pronouns": 1,
4804
                       "male_terms": 0,
4805
```

```
4806
                        "female_terms": 1,
4807
                        "total male": 0,
4808
                        "total_female": 2,
4809
                        "dominant_gender": "female",
4810
                        "gender_balance": 2
4811
4812
4813
                        "male_pronouns": 0,
4814
                        "female_pronouns": 3,
4815
                        "male_terms": 0,
4816
                        "female_terms": 0,
4817
                        "total_male": 0,
4818
                        "total_female": 3,
4819
                        "dominant_gender": "female",
4820
                        "gender_balance": 3
4822
                   1
4824
4826
4827
        }
4828
4829
```

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

**Results**. We estimate  $\delta^c$  and compare it against  $\delta^c$  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  $(\hat{\delta}^c, \delta^c)$  to measure degree of identifiability. Table 7 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).

Table 6: 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 7: The mean MCC values between the learnt and the ground truth concept vectors are close to 1.

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

## 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.