

# 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

054 that reflects the concept changes and uses a linear decoding function to reconstruct the difference  
 055 vector. This architecture reflects the *linear representation hypothesis* (Mikolov et al., 2013; Jiang  
 056 et al., 2024) in assuming that concepts are linearly encoded by LLMs. Crucially, we regularize  
 057 the latent representation to be sparse, meaning that each shift is modelled using as few concept  
 058 changes as necessary. We then leverage the results developed by Lachapelle et al. (2023) and Xu  
 059 et al. (2024) to prove that the proposed SSAE approach identifies *some* concepts, under suitable  
 060 assumptions on distribution that generated the data. We also show how this allow to extract valid  
 061 steering vectors, i.e. direction in the LLM representation that changes a single concept. We  
 062 study the SSAE empirically on challenging language datasets and models, finding many settings  
 063 where they outperform SAEs as well as other related steering methods that require supervision.

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

071 **2 PROBLEM FORMULATION**

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

080 
$$\tilde{\mathbf{c}} := \mathbf{c} + \delta^c; \quad \delta_k^c = \lambda \mathbf{e}_k, \tag{1}$$

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

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

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

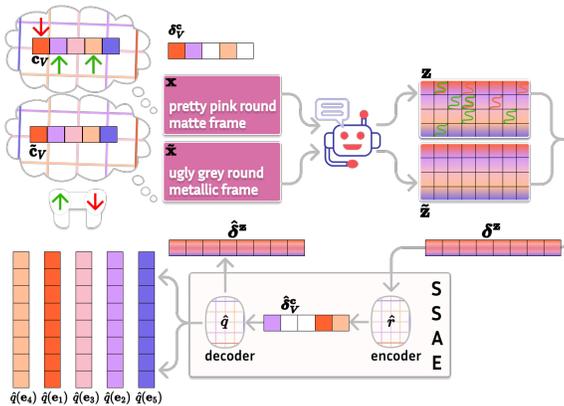


Figure 1: SSAEs map multi-concept shifts in embedding space to concept shifts, leveraging the latter’s sparsity, thereby recovering steering vectors. The learnt steering vectors are identified up to permutation and scaling.

**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  $\hat{\delta}^c$  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:

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

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

where  $S \subseteq \{1, \dots, d_c\}$  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  $V \subseteq \{1, \dots, d_c\}$  to be the set of *varying concepts*:

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

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  $g : \mathcal{C} \rightarrow \mathcal{X}$  and the learned encoding function  $f : \mathcal{X} \rightarrow \mathcal{Z}$  are such that  $f \circ g : \mathcal{C} \rightarrow \mathcal{Z}$  is linear, implying there exists a  $d_z \times d_c$  real matrix  $\mathbf{A}$  such that:*

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

Put simply, the LRH says that the learned representation  $\mathbf{z}$  *linearly encodes concepts*. Consequently, difference vectors  $\delta^z$  are also linearly related to concept shifts so that  $\delta^z = \mathbf{A}\delta^c$ . 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  $\delta_V^z$  and model them with an affine encoder  $r : \mathbb{R}^{d_z} \rightarrow \mathbb{R}^{|V|}$  and an affine decoder  $q : \mathbb{R}^{|V|} \rightarrow \mathbb{R}^{d_z}$  such that,

$$\hat{\delta}_V^c := r(\delta^z) := \mathbf{W}_e(\delta^z - \mathbf{b}_d) + \mathbf{b}_e; \quad (6)$$

$$\hat{\delta}^z := q(\hat{\delta}_V^c) := \mathbf{W}_d \hat{\delta}_V^c + \mathbf{b}_d. \quad (7)$$

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

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

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

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  $\delta_V^c := \hat{r}(\delta^z)$  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:

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

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 data-generating 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  $\hat{r}(\delta^z) = \mathbf{P}\mathbf{D}\delta_V^c$  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  $\hat{q}(e_k) \in \mathbb{R}^{d_z}$  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  $\delta_V^c$  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  $\delta_V^c$  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  $\delta_V^c$  up to permutation and element-wise rescaling.

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

$$\begin{aligned} \delta^z &= \mathbf{A}\delta^c = \mathbf{A}_V \delta_V^c + \mathbf{A}_{\bar{V}} \delta_{\bar{V}}^c \\ &= \mathbf{A}_V \delta_V^c, \end{aligned} \quad (11)$$

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

To prove linear identifiability, we will need one more assumption. Let  $\Delta_V^c$  be the support of the random vector  $\delta_V^c$ . 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.**  *$\text{span}(\Delta_V^c) = \mathbb{R}^{|V|}$ .*

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 Apx. A.7 and Asm. 2 and 3, there exists an invertible matrix  $\mathbf{L} \in \mathbb{R}^{|V| \times |V|}$  such that  $\hat{q} = \mathbf{A}_V \mathbf{L}$  and  $\hat{r}(\mathbf{z}) = \mathbf{L}^{-1} \mathbf{A}_V^+ \mathbf{z}$  for all  $\mathbf{z} \in \text{Im}(\mathbf{A}_V)$ , where  $\text{Im}(\mathbf{A}_V)$  is the image of  $\mathbf{A}_V$ .<sup>1</sup>

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

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

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

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

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

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

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

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  $\beta = \mathbb{E} \|\delta_V^c\|_0$ . Under Apx. A.7 and Asm. 2 to 4, there exists an invertible diagonal matrix and a permutation matrix  $\mathbf{D}, \mathbf{P} \in \mathbb{R}^{|V| \times |V|}$  such that  $\hat{q} = \mathbf{A}_V \mathbf{D} \mathbf{P}$  and  $\hat{r}(\mathbf{z}) = \mathbf{P}^T \mathbf{D}^{-1} \mathbf{A}_V^+ \mathbf{z}$  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  $\mathbb{E} \|\mathbf{L}^{-1} \delta_V^c\|_0 \leq \beta = \mathbb{E} \|\delta_V^c\|_0$ . 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  $\delta_V^c$  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 Apx. A.7 and Asm. 2 to 4, Prop. 2 shows that  $\hat{q} = \mathbf{A}_V \mathbf{D} \mathbf{P}$ . From this identifiability result, we can see that,

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

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

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

<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 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.

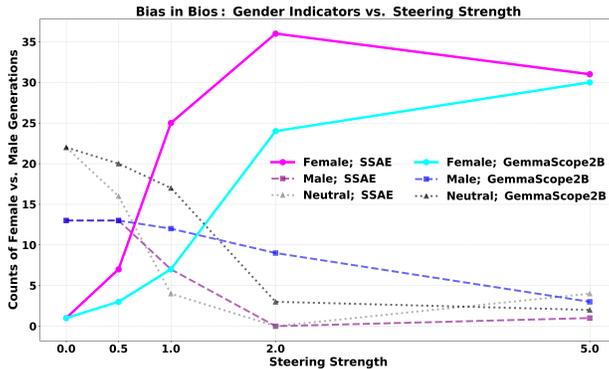


Figure 2: **SSAEs achieve earlier and stronger transitions to female indicators in generated text** (by strength 1.0), while GemmaScope2B requires stronger interventions.

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

**Baselines.** To validate our theory, we consider four different baselines: 1) SAEs trained on large-scale data (LlamaScope (He et al., 2024), GemmaScope (Lieberum et al., 2024), or PythiaSAE (EleutherAI, 2023) depending on the LLM the activations are obtained from), 2) an affine Autoencoder (aff) with an identical architecture as SSAEs but with no sparsity regularisation, 3) PCA on the same multi-concept difference vectors that SSAEs use, obtained from the last hidden layer (Liu et al., 2024), and 4) Mean Difference (MD) vectors that use paired observations differing in a single concept to compute  $\frac{1}{n} \sum_{i=1}^n (\tilde{\mathbf{z}}_k^{(i)} - \mathbf{z}_k^{(i)}) = \frac{1}{n} \sum_{i=1}^n \lambda^{(i)} \mathbf{Ae}_k = \bar{\lambda} \mathbf{Ae}_k$  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. To evaluate the effect of applying a steering vector in embedding space, we consider held-out single concept shift data  $(\mathbf{x}, \tilde{\mathbf{x}}_k)$  and evaluate how

**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-

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. We also report reconstruction errors in Apx. B.1.4.

### Validation of Theoretical Claims.

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

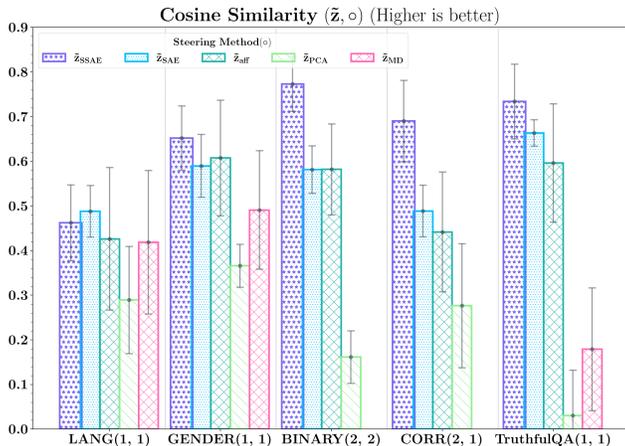


Figure 3: **A higher MCC value of the estimated decoder is associated with a greater cosine similarity.** Embeddings steered with vectors from a more disentangled decoder align more closely to target embeddings.

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

	SSAE	aff
LANG(1, 1)	0.995 $\pm$ 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  $MCC_D$ . We use 10 decoder pairs from 5 seeds for selected model hyperparameters. Table 1 shows that SSAE achieves consistently high MCC values, empirically corroborating Prop. 2, assuming a known  $|V|$ . As a sensitivity analysis, we further entangle the LLM embeddings by applying a dense linear invertible transformation  $\mathbf{L}$  to the embeddings to generate  $(\mathbf{L}f(\mathbf{x}), \mathbf{L}f(\tilde{\mathbf{x}}))$ . As expected, Section 5 demonstrates that this widens the gap between SSAE and the affine baseline. The worsening performance of *aff* after the entanglement is applied suggests that LLM representations might already somewhat disentangle some concepts or encode them through sparse or simple transformations. Next, we evaluate whether the benefits of a higher  $MCC_D$  translate to performance improvements on steering embeddings to be more similar to those of the target concept.

**Does identifiability translate to better steering?** We hold out pairs  $(x, \tilde{x}_k) \forall k \in V$ , each varying by a single concept, and compare the cosine similarity between the steered embeddings and target embeddings. Figure 3 illustrates that SSAE’s higher  $MCC_D$  performance generally translates to more accurate steering, with significant advantages over all related methods in the more challenging BINARY(2, 2) and CORR(2, 1) settings where multiple or correlated concepts change. Figure 3 also reveals that even slight differences in  $MCC_D$  values can translate into pronounced variations in steering accuracy. Next, we evaluate *out-of-distribution* (OOD) steering accuracy, based on the hypothesis that steering vectors that disentangle a single concept should transfer to different domains.

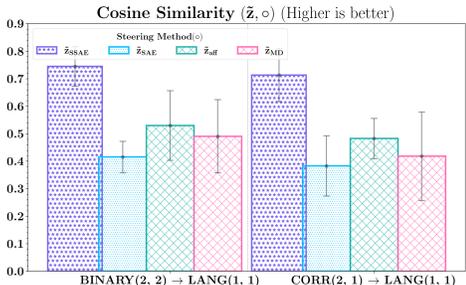


Figure 4: Embeddings steered using SSAE show higher OOD generalisation performance. SAEs (LlamaScope) generalise worse than other simpler baselines.

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

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

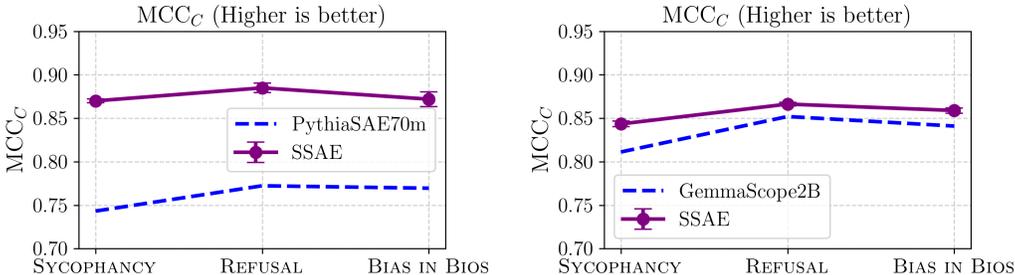


Figure 5:  $MCC_C$  values between learned activations and concept labels indicate that SSAEs can outperform SAEs in identifying ground-truth concepts in the dataset

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

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

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

## 6 RELATED WORK

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

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

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

## 7 CONCLUSION

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

## REFERENCES

- 486  
487  
488 Alekh Agarwal, Animashree Anandkumar, Prateek Jain, and Praneeth Netrapalli. Learning sparsely  
489 used overcomplete dictionaries via alternating minimization, 2014. URL <https://arxiv.org/abs/1310.7991>. Cited on page 27.  
490
- 491 Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. Weakly supervised representation learning with  
492 sparse perturbations. *Advances in Neural Information Processing Systems*, 35:15516–15528, 2022.  
493 Cited on pages 4, 9, and 25.  
494
- 495 Carl Allen and Timothy Hospedales. Analogies explained: Towards understanding word embeddings,  
496 2019. URL <https://arxiv.org/abs/1901.09813>. Cited on page 9.  
497
- 498 Evan Anders, Clement Neo, Jason Hoelscher-Obermaier, and Jessica N. Howard. Sparse autoencoders  
499 find composed features in small toy models. [https://www.lesswrong.com/posts/a5  
500 wwqza2cY3W7L9cj/sparse-autoencoders-find-composed-features-in-s  
501 mall-toy](https://www.lesswrong.com/posts/a5wwqza2cY3W7L9cj/sparse-autoencoders-find-composed-features-in-small-toy), 2024. Cited on pages 27 and 94.
- 502 Rohan Anil, Zhenzhong Lan, Olivier J. Hénaff, Anastasios Angelopoulos, Aakanksha Chowdhery,  
503 Yujia Li, Basil Mustafa, Daniel Freeman, Henryk Michalewski, Sara Hooker, et al. Gemma 2:  
504 Open models based on gemini research and technology. <https://ai.google.dev/gemma>,  
505 2024. Accessed: 2025-09-24. Cited on pages 6 and 8.
- 506 Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel  
507 Nanda. Refusal in language models is mediated by a single direction, 2024. URL <https://arxiv.org/abs/2406.11717>. Cited on pages 1 and 34.  
508  
509
- 510 Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint*  
511 *arXiv:1607.06450*, 2016. Cited on page 27.  
512
- 513 Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the  
514 restricted isometry property for random matrices. *Constructive approximation*, 28(3):253–263,  
515 2008. Cited on page 26.
- 516 Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, Eric Hallahan, Jesse Mu,  
517 Kyle O’Brien, Eric Burns, Leo Gao, Sid Black, and Connor Leahy. Pythia: A suite for analyzing  
518 language models across training and scaling. <https://arxiv.org/abs/2304.01373>,  
519 2023. arXiv:2304.01373. Cited on pages 6 and 8.  
520
- 521 Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal rep-  
522 resentation learning, 2022. URL <https://arxiv.org/abs/2203.16437>. Cited on  
523 page 9.
- 524 Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick  
525 Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec,  
526 Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina  
527 Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and  
528 Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary  
529 learning. *Transformer Circuits Thread*, 2023. URL [https://transformer-circuits.  
530 pub/2023/monosemantic-features/index.html](https://transformer-circuits.pub/2023/monosemantic-features/index.html). Cited on pages 27 and 28.  
531
- 532 Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language  
533 models without supervision, 2024. URL <https://arxiv.org/abs/2212.03827>. Cited  
534 on page 9.
- 535 Emmanuel Candes and Terence Tao. Decoding by linear programming, 2005. URL <https://arxiv.org/abs/math/0502327>. Cited on page 26.  
536  
537
- 538 Emmanuel Candes, Justin Romberg, and Terence Tao. Stable signal recovery from incomplete and  
539 inaccurate measurements, 2005. URL <https://arxiv.org/abs/math/0503066>. Cited  
on page 26.

- 540 Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse  
541 autoencoders find highly interpretable features in language models, 2023. URL <https://arxiv.org/abs/2309.08600>. Cited on pages 1, 2, 9, and 20.  
542  
543
- 544 Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra  
545 Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A  
546 case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference*  
547 *on Fairness, Accountability, and Transparency*, FAT\* '19, pp. 120–128. ACM, January 2019. doi:  
548 10.1145/3287560.3287572. URL <http://dx.doi.org/10.1145/3287560.3287572>.  
549 Cited on pages 6 and 8.
- 550 David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal)  
551 dictionaries via  $l_1$  minimization. *Proceedings of the National Academy of Sciences*, 100(5):  
552 2197–2202, 2003. Cited on page 26.  
553
- 554 D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306,  
555 2006. doi: 10.1109/TIT.2006.871582. Cited on page 26.
- 556 Sunny Duan, Loic Matthey, Andre Saraiva, Nicholas Watters, Christopher P Burgess, Alexander Ler-  
557 chner, and Irina Higgins. Unsupervised model selection for variational disentangled representation  
558 learning. *arXiv preprint arXiv:1905.12614*, 2019. Cited on pages 7, 28, and 29.  
559
- 560 EleutherAI. Sparse autoencoder for pythia-70m (32k features). [https://huggingface.co/E](https://huggingface.co/EleutherAI/sae-pythia-70m-32k)  
561 [leutherAI/sae-pythia-70m-32k](https://huggingface.co/EleutherAI/sae-pythia-70m-32k), 2023. Accessed: 2025-09-24. Cited on page 6.  
562
- 563 Joshua Engels, Isaac Liao, Eric J. Michaud, Wes Gurnee, and Max Tegmark. Not all language model  
564 features are linear, 2024. URL <https://arxiv.org/abs/2405.14860>. Cited on page 9.
- 565 Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Towards understanding linear word analogies,  
566 2019. URL <https://arxiv.org/abs/1810.04882>. Cited on page 9.  
567
- 568 Simon Foucart and Holger Rauhut. An invitation to compressive sensing. In *A mathematical*  
569 *introduction to compressive sensing*, pp. 1–39. Springer, 2013. Cited on page 26.
- 570 Marco Fumero, Florian Wenzel, Luca Zancato, Alessandro Achille, Emanuele Rodolà, Stefano Soatto,  
571 Bernhard Schölkopf, and Francesco Locatello. Leveraging sparse and shared feature activations for  
572 disentangled representation learning, 2023. URL <https://arxiv.org/abs/2304.07939>.  
573 Cited on page 9.  
574
- 575 Jose Gallego-Posada and Juan Ramirez. Cooper: a toolkit for Lagrangian-based constrained opti-  
576 mization. <https://github.com/cooper-org/cooper>, 2022. Cited on pages 4, 6, 27,  
577 and 28.
- 578 Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya  
579 Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint*  
580 *arXiv:2406.04093*, 2024. Cited on pages 9 and 28.  
581
- 582 Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding  
583 alignments between interpretable causal variables and distributed neural representations. In  
584 *Proceedings of the Third Conference on Causal Learning and Reasoning*, 2024. URL [https://](https://proceedings.mlr.press/v236/geiger24a.html)  
585 [proceedings.mlr.press/v236/geiger24a.html](https://proceedings.mlr.press/v236/geiger24a.html). Cited on page 1.  
586
- 587 Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A  
588 variational inequality perspective on generative adversarial networks, 2020. URL [https://ar](https://arxiv.org/abs/1802.10551)  
589 [xiv.org/abs/1802.10551](https://arxiv.org/abs/1802.10551). Cited on pages 4 and 28.
- 590 Alex Gittens, Dimitris Achlioptas, and Michael W. Mahoney. Skip-gram - Zipf + uniform = vector  
591 additivity. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of*  
592 *the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 69–76, Vancouver,  
593 Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1007. URL  
<https://aclanthology.org/P17-1007>. Cited on page 9.

- 594 Navita Goyal, Hal Daumé III, Alexandre Drouin, and Dhanya Sridhar. Causal differentiating concepts:  
595 Interpreting LM behavior via causal representation learning. *Neural Information Processing*  
596 *Systems (NeurIPS)*, 2025. Cited on page 9.
- 597 Rémi Gribonval and Morten Nielsen. Sparse representations in unions of bases. *IEEE transactions*  
598 *on Information theory*, 49(12):3320–3325, 2004. Cited on page 26.
- 600 Rémi Gribonval, Rodolphe Jenatton, Francis Bach, Martin Kleinstueber, and Matthias Seibert.  
601 Sample complexity of dictionary learning and other matrix factorizations. *IEEE Transactions on*  
602 *Information Theory*, 61(6):3469–3486, 2015. Cited on page 27.
- 603 Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu,  
604 Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. Llama scope: Extracting millions of features  
605 from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024. Cited on  
606 page 6.
- 607 Geoffrey E Hinton et al. Learning distributed representations of concepts. In *Proceedings of the*  
608 *eighth annual conference of the cognitive science society*, volume 1, pp. 12. Amherst, MA, 1986.  
609 Cited on pages 3 and 25.
- 611 Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning  
612 and nonlinear ica, 2016. URL <https://arxiv.org/abs/1605.06336>. Cited on pages 6  
613 and 32.
- 614 A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness  
615 results. *Neural Networks*, 1999. Cited on page 2.
- 616 Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam, and Victor Veitch. On the  
617 origins of linear representations in large language models, 2024. URL <https://arxiv.org/abs/2403.03867>. Cited on pages 2, 3, 9, 25, and 33.
- 618 I. Khemakhem, D. Kingma, R. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ica:  
619 A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial*  
620 *Intelligence and Statistics*, 2020a. Cited on page 4.
- 621 Ilyes Khemakhem, Ricardo Pio Monti, Diederik P. Kingma, and Aapo Hyvärinen. Ice-beem:  
622 Identifiable conditional energy-based deep models based on nonlinear ica, 2020b. URL <https://arxiv.org/abs/2002.11537>. Cited on pages 6 and 32.
- 623 Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and  
624 Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference*  
625 *on Machine Learning*, 2020. URL <https://proceedings.mlr.press/v119/koh20a.html>. Cited on page 1.
- 626 Galina M Korpelevich. The extragradient method for finding saddle points and other problems.  
627 *Matecon*, 12:747–756, 1976. Cited on page 28.
- 628 Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre  
629 Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A  
630 new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pp. 428–484.  
631 PMLR, 2022. Cited on page 9.
- 632 Sébastien Lachapelle, T. Deleu, D. Mahajan, I. Mitliagkas, Y. Bengio, S. Lacoste-Julien, and  
633 Q. Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in  
634 multi-task learning. In *International Conference on Machine Learning*, 2023. Cited on pages 2, 5,  
635 9, 22, 24, and 94.
- 636 Elliot Layne, Jason Hartford, Sébastien Lachapelle, Mathieu Blanchette, and Dhanya Sridhar. Sparsity  
637 regularization via tree-structured environments for disentangled representations, 2024. URL  
638 <https://arxiv.org/abs/2405.20482>. Cited on page 9.
- 639 Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-  
640 time intervention: Eliciting truthful answers from a language model, 2024. URL <https://arxiv.org/abs/2306.03341>. Cited on page 9.

- 648 Tom Lieberum, Senthoran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant  
649 Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse  
650 autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*, 2024. Cited  
651 on page 6.
- 652 Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human  
653 falsehoods, 2022. URL <https://arxiv.org/abs/2109.07958>. Cited on page 7.
- 654 Sheng Liu, Haotian Ye, Lei Xing, and James Zou. In-context vectors: Making in context learning  
655 more effective and controllable through latent space steering, 2024. URL <https://arxiv.org/abs/2311.06668>. Cited on page 6.
- 656 Llama Team et al. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>. Cited on pages 6 and 33.
- 657 Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf,  
658 and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disen-  
659 tangled representations, 2019. URL <https://arxiv.org/abs/1811.12359>. Cited on  
660 page 28.
- 661 Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael  
662 Tschannen. Weakly-supervised disentanglement without compromises. In *International conference*  
663 *on machine learning*, pp. 6348–6359. PMLR, 2020a. Cited on page 9.
- 664 Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael  
665 Tschannen. Weakly-supervised disentanglement without compromises, 2020b. Cited on page 3.
- 666 David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. The randomized dependence coefficient.  
667 *Advances in neural information processing systems*, 26, 2013. Cited on page 32.
- 668 Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning llama for multi-stage  
669 text retrieval, 2023. URL <https://arxiv.org/abs/2310.08319>. Cited on page 6.
- 670 Emanuele Marconato, Sébastien Lachapelle, Sebastian Weichwald, and Luigi Gresele. All or none:  
671 Identifiable linear properties of next-token predictors in language modeling. *arXiv preprint*  
672 *arXiv:2410.23501*, 2024. Cited on pages 3, 8, 9, and 25.
- 673 Abhinav Menon, Manish Shrivastava, David Krueger, and Ekdeep Singh Lubana. Analyzing  
674 (in)abilities of saes via formal languages. *Conference of the Nations of the Americas Chapter*  
675 *of the Association for Computational Linguistics: Human Language Technologies*, 2025. Cited  
676 on page 1.
- 677 Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word  
678 representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (eds.), *Proceedings*  
679 *of the 2013 Conference of the North American Chapter of the Association for Computational Lin-*  
680 *guistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013. Association  
681 for Computational Linguistics. URL <https://aclanthology.org/N13-1090>. Cited on  
682 pages 2, 3, 9, and 25.
- 683 Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and  
684 Emanuele Rodolà. Relative representations enable zero-shot latent space communication, 2023.  
685 URL <https://arxiv.org/abs/2209.15430>. Cited on page 9.
- 686 Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding  
687 benchmark, 2023. URL <https://arxiv.org/abs/2210.07316>. Cited on page 9.
- 688 Malvina Nissim, Rik van Noord, and Rob van der Goot. Fair is better than sensational: Man is  
689 to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497, June 2020. doi:  
690 10.1162/coli\_a\_00379. URL <https://aclanthology.org/2020.cl-2.7>. Cited on  
691 page 9.
- 692 Charles O’Neill, Alim Gumran, and David Klindt. Compute optimal inference and provable amorti-  
693 sation gap in sparse autoencoders, 2025. URL <https://arxiv.org/abs/2411.13117>.  
694 Cited on page 27.

- 702 Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt  
703 Turner. Steering llama 2 via contrastive activation addition, 2024. URL <https://arxiv.org/abs/2312.06681>. Cited on pages 6, 8, 9, and 34.
- 704  
705
- 706 Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry  
707 of large language models, 2023. Cited on pages 3, 9, and 33.
- 708  
709 Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and  
710 hierarchical concepts in large language models, 2024. URL <https://arxiv.org/abs/2406.01506>. Cited on page 9.
- 711  
712 Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. Steer LLM latents for  
713 hallucination detection. In *Forty-second International Conference on Machine Learning*, 2025.  
714 Cited on page 1.
- 715  
716 Ethan Perez, Sam Ringer, Kamilė Lukošiuūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig  
717 Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann,  
718 Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei,  
719 Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion,  
720 James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon  
721 Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson  
722 Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam  
723 McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-  
724 Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark,  
725 Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan  
726 Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with  
727 model-written evaluations, 2022. URL <https://arxiv.org/abs/2212.09251>. Cited on  
728 page 8.
- 729  
730 Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János  
731 Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoen-  
732 coders, 2024. URL <https://arxiv.org/abs/2404.16014>. Cited on page 9.
- 733  
734 Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar.  
735 Learning interpretable concepts: Unifying causal representation learning and foundation models,  
736 2024. Cited on pages 3, 5, 9, 20, and 25.
- 737  
738 Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out:  
739 Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie  
740 Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for  
741 Computational Linguistics*, pp. 7237–7256, Online, July 2020a. Association for Computational  
742 Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL <https://aclanthology.org/2020.acl-main.647>. Cited on page 9.
- 743  
744 Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out:  
745 Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*,  
746 2020b. Cited on pages 3 and 25.
- 747  
748 Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner.  
749 Steering llama 2 via contrastive activation addition, 2024. Cited on pages 1, 9, and 20.
- 750  
751 Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations.  
752 In *International Conference on Machine Learning*, pp. 9030–9039. PMLR, 2021. Cited on pages 3,  
753 4, 6, 9, and 25.
- 754  
755 Michal Rolínek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions  
(by accident), 2019. URL <https://arxiv.org/abs/1812.06775>. Cited on page 7.
- 756  
757 David E Rumelhart and Adele A Abrahamson. A model for analogical reasoning. *Cognitive  
758 Psychology*, 5(1):1–28, 1973. Cited on pages 3 and 25.
- 759  
760 Karin Schnass. Local identification of overcomplete dictionaries, 2015. URL <https://arxiv.org/abs/1401.6354>. Cited on page 27.

- 756 Yeon Seonwoo, Sungjoon Park, Dongkwan Kim, and Alice Oh. Additive compositionality of word  
757 vectors. In Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi (eds.), *Proceedings of the*  
758 *5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 387–396, Hong Kong, China,  
759 November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5551. URL  
760 <https://aclanthology.org/D19-5551>. Cited on page 9.
- 761 Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel  
762 text by cross-alignment, 2017. URL <https://arxiv.org/abs/1705.09655>. Cited on  
763 page 20.
- 764 Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries,  
765 2012. URL <https://arxiv.org/abs/1206.5882>. Cited on page 27.
- 766 Nishant Subramani, Nivedita Suresh, and Matthew E Peters. Extracting latent steering vectors from  
767 pretrained language models. *ACL Findings*, 2022. Cited on pages 1 and 9.
- 768 Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam  
769 Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner,  
770 Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward  
771 Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling  
772 monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits*  
773 *Thread*, 2024. URL <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>. Cited on pages 9 and 26.
- 774 Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of  
775 sentiment in large language models, 2023. URL <https://arxiv.org/abs/2310.15154>.  
776 Cited on page 9.
- 777 J.A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on*  
778 *Information Theory*, 50(10):2231–2242, 2004. doi: 10.1109/TIT.2004.834793. Cited on page 26.
- 779 Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini,  
780 and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL  
781 <https://arxiv.org/abs/2308.10248>. Cited on pages 1, 9, and 20.
- 782 Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christo-  
783 pher D. Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outper-  
784 form sparse autoencoders, 2025. Cited on page 1.
- 785 Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco  
786 Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation  
787 learning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. Cited  
788 on pages 2, 4, 5, 9, 22, and 94.
- 789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809

..we understand the world by studying  
change, not by studying things..

---

As quoted in the Order of Time,  
Anaximander

810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863

**CONTENTS**

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Problem formulation</b>	<b>2</b>
<b>3</b>	<b>Sparse Shift Autoencoders (SSAEs)</b>	<b>3</b>
<b>4</b>	<b>Identifiability analysis</b>	<b>4</b>
<b>5</b>	<b>Empirical Studies</b>	<b>6</b>
<b>6</b>	<b>Related work</b>	<b>9</b>
<b>7</b>	<b>Conclusion</b>	<b>9</b>
<b>A</b>	<b>Theory</b>	<b>18</b>
A.1	Notation and Glossary . . . . .	18
A.2	Steering Functions . . . . .	19
A.3	Linear Identifiability is insufficient for steering. . . . .	20
A.4	Proof of Prop. 1 (linear identifiability) . . . . .	21
A.5	Proof of Prop. 2 (permutation identifiability) . . . . .	22
A.6	Distributions satisfying Asm. 4 . . . . .	24
A.7	Interpreting the Linear Representation Hypothesis . . . . .	24
A.8	Parallels with Sparse Coding and Addressing the Amortisation Gap . . . . .	26
<b>B</b>	<b>Implementation and experimental details</b>	<b>27</b>
B.1	SSAE Architecture . . . . .	28
B.1.1	Model Selection via Unsupervised Diversity Ranking (UDR) . . . . .	28
B.1.2	Sparse optimization . . . . .	29
B.1.3	Datasets . . . . .	30
B.1.4	Reconstruction Errors . . . . .	30
B.2	Mean Correlation Coefficient: Gateway to Interpreting Latent Dimensions . . . . .	32
B.3	Cosine Similarity . . . . .	33
B.4	Test of robustness: impact of increasing the encoding dimension . . . . .	33
B.5	Steering Intermediate Layers of Llama-3.1-8B . . . . .	34
B.6	Bias in Bios Generations . . . . .	34

864	<b>B.7 Synthetic Experiments</b> . . . . .	93
865		
866		
867		
868		
869		
870		
871		
872		
873		
874		
875		
876		
877		
878		
879		
880		
881		
882		
883		
884		
885		
886		
887		
888		
889		
890		
891		
892		
893		
894		
895		
896		
897		
898		
899		
900		
901		
902		
903		
904		
905		
906		
907		
908		
909		
910		
911		
912		
913		
914		
915		
916		
917		

## A THEORY

## A.1 NOTATION AND GLOSSARY

## General notation

---

923	$k$	integer
924	$[k]$	set of all integers between 1 and $k$ , inclusively
925	$S \subseteq [k]$	set
926	$ S $	cardinality of a set
927	$S \setminus S'$	set subtraction (set of elements of $S$ that are not in $S'$ )
928	$\lambda$	scalar
929	$\mathbf{x}$	vector and vector-valued random variables
930	$x_k$	element $k$ of a random vector $\mathbf{x}$
931	$\mathbf{x}_S$	subvector with element $x_i$ for $i \in S$
932	$\mathbf{A}$	matrix
933	$\mathbf{A}_{i,j}$	element $i, j$ of matrix $\mathbf{A}$
934	$\mathbf{A}_{:,i}$	column $i$ of matrix $\mathbf{A}$
935	$\mathbf{A}_S$	matrix with columns $\mathbf{A}_{:,j}$ for $j \in S$
936	$\mathbf{A}^+$	pseudo-inverse of a matrix $\mathbf{A}$
937	$\mathbf{e}_k \in \mathbb{R}^n$	standard basis vector of the form $[0, \dots, 0, 1, 0, \dots, 0]$ with a 1 at position $k$
938	$f : \mathcal{X} \rightarrow \mathcal{Z}$	function $f$ with domain $\mathcal{X}$ and codomain $\mathcal{Z}$
939	$f \circ g$	composition of the functions $f$ and $g$
940	$\ \mathbf{x}\ _p$	$\ell_p$ norm of $\mathbf{x}$
941	$\frac{\partial y}{\partial x}$	partial derivative of $y$ with respect to $x$
942	$\nabla_{\mathbf{x}} f(\mathbf{x}) \in \mathbb{R}^{m \times n}$	Jacobian matrix of $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$
943	$\nabla_{\mathbf{x}}^2 f(\mathbf{x}) \in \mathbb{R}^{n \times n}$	Hessian matrix of $f : \mathbb{R}^n \rightarrow \mathbb{R}$
944	$\mathbb{P}$	probability measure/distribution
945	$\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$	expectation of $f(\mathbf{x})$ with respect to $\mathbf{x}$

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

## Glossary

---

$\mathbf{x} \in \mathbb{R}^{d_x}$	observation
$\mathbf{z} \in \mathbb{R}^{d_z}$	pretrained representation
$\mathbf{c} \in \mathbb{R}^{d_c}$	ground-truth concept vector
$\tilde{\mathbf{c}}_{k,\lambda}$	ground-truth concept vector after varying concept $k$ by $\lambda$ from $\mathbf{c}$
$\tilde{\mathbf{x}}_{k,\lambda}$	observation corresponding to $\tilde{\mathbf{c}}_{k,\lambda}$
$\tilde{\mathbf{z}}_{k,\lambda}$	pretrained representation corresponding to $\tilde{\mathbf{c}}_{k,\lambda}$
$\mathcal{X} \subseteq \mathbb{R}^{d_x}$	support of observations
$\mathcal{Z} \subseteq \mathbb{R}^{d_z}$	support of pretrained representations
$\mathcal{C} \subseteq \mathbb{R}^{d_c}$	support of ground-truth concept vectors
$S \subseteq [d_c]$	subset of varying concepts in a given pair $(\mathbf{x}, \tilde{\mathbf{x}})$
$V \subseteq [d_c]$	subset of concepts allowed to vary between $\mathbf{x}$ and $\tilde{\mathbf{x}}$
$\delta^c$	concept shift vector
$\hat{\delta}^c$	estimated concept shift vector
$\delta^z$	pretrained representation shift vector
$g : \mathcal{C} \rightarrow \mathcal{X}$	map from concept representations to observations
$f : \mathcal{X} \rightarrow \mathcal{Z}$	map from observations to learned representations
$r : \mathcal{Z} \rightarrow \mathcal{C}$	encoding function
$\hat{r} : \mathcal{Z} \rightarrow \mathcal{C}$	estimated encoding function
$q : \mathcal{C} \rightarrow \mathcal{Z}$	decoding function
$\hat{q} : \mathcal{C} \rightarrow \mathcal{Z}$	estimated decoding function
$\phi_{k,\lambda} : \mathcal{Z} \rightarrow \mathcal{Z}$	steering function
$\hat{\phi}_{k,\lambda} : \mathcal{Z} \rightarrow \mathcal{Z}$	estimated steering function
<b>A</b>	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**  $\phi_{k,\lambda} : \mathcal{Z} \rightarrow \mathcal{Z}$  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<sup>3</sup> 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  $\tilde{\mathbf{x}}_{\lambda,k} := g(\tilde{\mathbf{c}}_{k,\lambda})$  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>3</sup>Steering functions are not guaranteed to exist. However, if  $f$  and  $g$  are injective, we have  $\phi_{\lambda,k}(\mathbf{z}) = f(g(g^{-1}(f^{-1}(\mathbf{z})) + \lambda \mathbf{e}_k))$ .

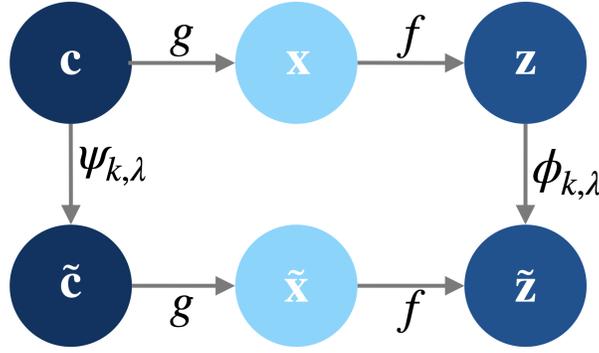


Figure 6: A steering function  $\phi_{k,\lambda}$  is s.t. the above diagram commutes, i.e.,  $\phi_{k,\lambda}(f(g(\mathbf{c}))) = f(g(\psi_{k,\lambda}(\mathbf{c}))) \forall \mathbf{c}$ . (see Defn. 3).

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  $\tilde{\mathbf{z}}_{k,\lambda} := \mathbf{z} + \delta_{k,\lambda}^z$ , where  $\delta_{k,\lambda}^z$  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  $\hat{q} := \mathbf{A}_V \mathbf{D} \mathbf{P}$  (Prop. 2).

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

## A.4 PROOF OF PROP. 1 (LINEAR IDENTIFIABILITY)

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

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

$$\mathbb{E}_{\mathbf{x}, \bar{\mathbf{x}}} \|\delta^z - q^*(r^*(\delta^z))\|_2^2 = \mathbb{E}_{\mathbf{x}, \bar{\mathbf{x}}} \|\delta^z - \mathbf{A}_V \mathbf{A}_V^+ \delta^z\|_2^2 \quad (13)$$

$$= \mathbb{E}_{\mathbf{c}, \bar{\mathbf{c}}} \|\mathbf{A}_V \delta_V^c - \mathbf{A}_V (\mathbf{A}_V^+ \mathbf{A}_V) \delta_V^c\|_2^2 \quad (14)$$

$$= \mathbb{E}_{\mathbf{c}, \bar{\mathbf{c}}} \|\mathbf{A}_V \delta_V^c - \mathbf{A}_V \delta_V^c\|_2^2 \quad (15)$$

$$= 0, \quad (16)$$

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

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

$$\mathbb{E}_{\mathbf{x}, \bar{\mathbf{x}}} \|\delta^z - \hat{q}(\hat{r}(\delta^z))\|_2^2 = 0 \quad (17)$$

$$\mathbb{E}_{\mathbf{c}, \bar{\mathbf{c}}} \|\mathbf{A}_V \delta_V^c - \hat{q}(\hat{r}(\mathbf{A}_V \delta_V^c))\|_2^2 = 0 \quad (18)$$

This means we must have

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

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

$$\mathbf{A}_V \delta_V^c = \mathbf{Q} \mathbf{R} \mathbf{A}_V \delta_V^c, \quad (20)$$

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

$$\mathbf{A}_V \mathbf{C} = \mathbf{Q} \mathbf{R} \mathbf{A}_V \mathbf{C} \quad (21)$$

$$\mathbf{A}_V = \mathbf{Q} \mathbf{R} \mathbf{A}_V, \quad (22)$$

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

$$\mathbf{A}_V = \mathbf{Q} \mathbf{L}^{-1} \quad (23)$$

$$\hat{q} = \mathbf{Q} = \mathbf{A}_V \mathbf{L}, \quad (24)$$

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 exist a point  $\mathbf{c} \in \mathbb{R}^{|V|}$  such that  $\mathbf{z} = \mathbf{A}_V \mathbf{c}$ . Now we evaluate

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

$$= \mathbf{L}^{-1} \mathbf{c} \quad (26)$$

$$= \mathbf{L}^{-1} \mathbf{A}_V^+ \mathbf{A}_V \mathbf{c} \quad (27)$$

$$= \mathbf{L}^{-1} \mathbf{A}_V^+ \mathbf{z}, \quad (28)$$

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.  $\square$

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

## A.5 PROOF OF PROP. 2 (PERMUTATION IDENTIFIABILITY)

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

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

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

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

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

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

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

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

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

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

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

This concludes the proof.  $\square$

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

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

1. For all  $j \in [m]$ , we have  $\bigcup_{S \in \mathcal{S} | j \notin S} S = [m] \setminus \{j\}$ ; and
2. For all  $S \in \mathcal{S}$ , the conditional distribution  $\mathbb{P}_{\mathbf{x}_S|S}$  is such that, for all nonzero  $\mathbf{a} \in \mathbb{R}^{|S|}$ ,  $\mathbb{P}_{\mathbf{x}_S|S}\{\mathbf{x}_S \mid \mathbf{a}^\top \mathbf{x}_S = 0\} = 0$ .

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

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

$$\mathbb{E}\|\mathbf{x}\|_0 = \mathbb{E}_{p(S)}\mathbb{E}\left[\sum_{j=1}^m \mathbf{1}(\mathbf{x}_j \neq 0) \mid S\right] \quad (33)$$

$$= \mathbb{E}_{p(S)}\sum_{j=1}^m \mathbb{E}[\mathbf{1}(\mathbf{x}_j \neq 0) \mid S] \quad (34)$$

$$= \mathbb{E}_{p(S)}\sum_{j=1}^m \mathbb{P}_{\mathbf{x}|S}\{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{x}_j \neq 0\} \quad (35)$$

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

where the last step follows from the definition of  $S$ .

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

$$\mathbb{E}\|\mathbf{L}\mathbf{x}\|_0 = \mathbb{E}_{p(S)}\mathbb{E}\left[\sum_{j=1}^m \mathbf{1}(\mathbf{L}_{j,:}\mathbf{x} \neq 0) \mid S\right] \quad (37)$$

$$= \mathbb{E}_{p(S)}\sum_{j=1}^m \mathbb{E}[\mathbf{1}(\mathbf{L}_{j,:}\mathbf{x} \neq 0) \mid S] \quad (38)$$

$$= \mathbb{E}_{p(S)}\sum_{j=1}^m \mathbb{E}[\mathbf{1}(\mathbf{L}_{j,S}\mathbf{x}_S \neq 0) \mid S] \quad (39)$$

$$= \mathbb{E}_{p(S)}\sum_{j=1}^m \mathbb{P}_{\mathbf{x}|S}\{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{L}_{j,S}\mathbf{x}_S \neq 0\}. \quad (40)$$

Notice that

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

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

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

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

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

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

Thus we can write

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

$$= 1 - \mathbf{1}(S \cap N_j = \emptyset) \quad (43)$$

$$= \mathbf{1}(S \cap N_j \neq \emptyset), \quad (44)$$

which allows us to write

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

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

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

Since  $L$  is invertible, there exists a permutation  $\sigma : [m] \rightarrow [m]$  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

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

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

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:

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

Importantly,

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

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

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

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

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,  $L = DP$  where  $D$  is an invertible diagonal matrix and  $P$  is a permutation matrix.  $\square$

#### A.6 DISTRIBUTIONS SATISFYING 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.

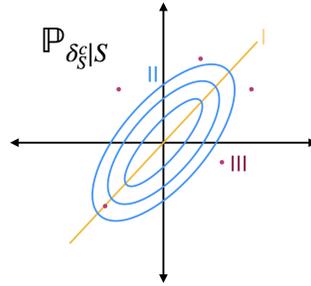


Figure 7: Three illustrative examples of  $\mathbb{P}_{\delta_S^c | S}$ : Only distribution II satisfies Asm. 4.

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

1296 real matrix  $\mathbf{A}$  such that:

$$1297 \quad \mathbf{z} = f(g(\mathbf{c})) = \mathbf{A}\mathbf{c} . \quad (53)$$

1298 The linear representation hypothesis (LRH) implies that the learned representation  $\mathbf{z}$  *linearly encodes*  
 1299 *concepts*. A long line of work provides evidence for this hypothesis (c.f. [Rumelhart & Abrahamson](#)  
 1300 [\(1973\)](#); [Hinton et al. \(1986\)](#); [Mikolov et al. \(2013\)](#); [Ravfogel et al. \(2020b\)](#)). More recently, theoretical  
 1301 work justifies why linear properties could arise in these models (c.f. [Jiang et al. \(2024\)](#); [Roeder](#)  
 1302 [et al. \(2021\)](#); [Marconato et al. \(2024\)](#)). Section 6 provides a full list of related work, while Apx. A.7  
 1303 provides an explanation of the equivalence between LRH’s different interpretations. [Rajendran et al.](#)  
 1304 [\(2024\)](#) also leverage the LRH in their work.

1305 **Corollary 5.** *If concept changes act on latent embeddings following  $\tilde{\mathbf{z}} = \mathbf{z} + \delta^z$  and  $q$  and  $r$  are*  
 1306 *injective, they must be affine transformations.*

1307 **Proof:** Starting with the interpretation of the *linear representation hypothesis* such that  $\tilde{\mathbf{z}} = \mathbf{z} + \delta^z$   
 1308 where  $\mathbf{z} = q(\mathbf{c})$  and  $\tilde{\mathbf{z}} = q(\tilde{\mathbf{c}})$ :

$$1309 \quad \implies q(\tilde{\mathbf{c}}) = q(\mathbf{c}) + \delta^z$$

1310 Since we identify only the varying concepts, this corresponds to identifying a subspace of the original  
 1311 concept space in which  $\tilde{\mathbf{c}} = \mathbf{c} + \delta_V^c$ .

1312 Using the injectivity of  $q$  (Asm. 2):

$$1313 \quad q(\mathbf{c} + \delta_V^c) = q(\mathbf{c}) + \delta^z \quad (54)$$

1314 Taking the gradient of both the LHS and the RHS wrt  $\mathbf{c}$ ,

$$1315 \quad \frac{\partial(\mathbf{c} + \delta_V^c)}{\partial(\mathbf{c})} \nabla_{(\mathbf{c} + \delta_V^c)} q(\mathbf{c} + \delta_V^c) = \nabla_{\mathbf{c}} q(\mathbf{c})$$

$$1316 \quad \nabla_{(\mathbf{c} + \delta_V^c)} q(\mathbf{c} + \delta_V^c) = \nabla_{\mathbf{c}} q(\mathbf{c})$$

$$1317 \quad \mathbf{J}^T(\mathbf{c} + \delta_V^c) = \mathbf{J}^T(\mathbf{c}) \quad (55)$$

1318 Where  $\mathbf{J}(\mathbf{c})$  is the Jacobian of  $q$  at  $\mathbf{c}$  and  $\mathbf{J}(\mathbf{c} + \delta_V^c)$  is the Jacobian of  $q$  at  $\mathbf{c} + \delta_V^c$ .

$$1319 \quad \begin{bmatrix} \nabla q_1(\mathbf{c} + \delta_V^c) \\ \nabla q_2(\mathbf{c} + \delta_V^c) \\ \nabla q_3(\mathbf{c} + \delta_V^c) \\ \vdots \\ \nabla q_{d_z}(\mathbf{c} + \delta_V^c) \end{bmatrix} - \begin{bmatrix} \nabla q_1(\mathbf{c}) \\ \nabla q_2(\mathbf{c}) \\ \nabla q_3(\mathbf{c}) \\ \vdots \\ \nabla q_{d_z}(\mathbf{c}) \end{bmatrix} = 0$$

1320 considering the  $j^{\text{th}}$  component of the difference,

$$1321 \quad \begin{bmatrix} \nabla^2 q_j(\theta_1) \\ \nabla^2 q_j(\theta_2) \\ \nabla^2 q_j(\theta_2) \\ \vdots \\ \nabla^2 q_j(\theta_d) \end{bmatrix} (\delta_V^c) = 0$$

1322 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  
 1323  $\mathbf{A}_V \in \mathbb{R}^{d_z \times d_z}$ ,  $\mathbf{b} \in \mathbb{R}^{d_z}$  or that  $q$  is affine. Similarly, we can show that  $r$  is affine too by starting  
 1324 with  $r(\mathbf{z} + \delta^z) = r(\mathbf{z}) + \delta_V^c$ .

1350 **Corollary 6.** *If we assume  $\tilde{\mathbf{z}} = \phi(\mathbf{z})$ , for an affine map  $q$ ,  $\mathbf{A} = \mathbf{I}$ .*

1351 **Proof:** Let’s assume the affine form of  $q$  can be expressed as:

$$1352 \mathbf{z} = \mathbf{A}_V \mathbf{c} + \mathbf{b} \quad (56)$$

1353 where  $\mathbf{A}_V \in \mathbb{R}^{d_z \times d_z}$  and  $\mathbf{k} \in \mathbb{R}^{d_z}$ .

1354 Similarly,  $\tilde{\mathbf{z}} = q(\tilde{\mathbf{c}}) = \mathbf{A}_V \tilde{\mathbf{c}} + \mathbf{b}$  and we know  $\tilde{\mathbf{c}} = \mathbf{c} + \delta_V^c$ .

$$1355 \implies \tilde{\mathbf{z}} = \mathbf{A}_V (\mathbf{c} + \delta_V^c) + \mathbf{b}$$

1356 we have  $\tilde{\mathbf{z}} = \phi(\mathbf{z})$  and from Eqn. (56):

$$1357 \phi(\mathbf{A}_V \mathbf{c} + \mathbf{b}) = \mathbf{A}_V (\mathbf{c} + \delta_V^c) + \mathbf{b} \quad (57)$$

1358 In the above equation, we can see that the maximum degree of  $\mathbf{c}$  on the RHS is 1, which implies  
1359 that the degree of  $\mathbf{c}$  on the LHS should also at most be 1, which implies  $\phi$  can at most be an affine  
1360 function.

1361 So let’s assume  $\phi$  is an affine function of the form:

$$1362 \tilde{\mathbf{z}} = \phi(\mathbf{z}) = \mathbf{T} \mathbf{z} + \delta^z \quad (58)$$

1363 where  $\mathbf{T} \in \mathbb{R}^{d_z \times d_z}$  and  $\delta^z \in \mathbb{R}^{d_z}$ . Substituting this in the above equation, we get:

$$1364 \mathbf{T}(\mathbf{A}_V \mathbf{c} + \mathbf{b}) + \delta^z = \mathbf{A}_V (\mathbf{c} + \delta_V^c) + \mathbf{b} \quad (59)$$

$$1365 \mathbf{Q}(\mathbf{T} - \mathbf{I})\mathbf{c} + (\mathbf{T} - \mathbf{I})\mathbf{b} + (\delta^z - \mathbf{Q}\delta_V^c) = 0$$

1366 For a non-trivial solution:

$$1367 \mathbf{T} = \mathbf{I} \quad (60)$$

$$1368 \delta^z = \mathbf{A}_V \delta_V^c \quad (61)$$

1369 So, we have proved that if we assume  $q$  to be affine, then  $\tilde{\mathbf{z}} = \mathbf{z} + \delta \mathbf{z}$ .

1370 **Implications:** Multiple expositions (Templeton et al., 2024) remark that it is not clear what the  
1371 meaning of *linear* exactly is in the linear representation hypothesis. Informally, many results cited  
1372 in support of the linear representation hypothesis either extract information with a linear probe, or  
1373 add a vector to influence model behavior. Here, we assume that if linear meant concepts are linearly  
1374 encoded in the latent space, we can show that this would correspond to shifts in the latent space  
1375 representing net concept changes and vice versa, which means both interpretations are the same, so it  
1376 does not matter which one is assumed.

## 1377 A.8 PARALLELS WITH SPARSE CODING AND ADDRESSING THE AMORTISATION GAP

1378 A long line of work on sparse coding and dictionary learning provides conditions under which  
1379 a ground-truth dictionary  $D$  is identifiable from observations, typically through the recovery of  
1380 instance-specific sparse codes. Foundational contributions establish that identifiability follows from  
1381 *geometric constraints* on the dictionary  $D$ , most prominently the *restricted isometry property* (RIP)  
1382 (Candes & Tao, 2005; Candes et al., 2005; Baraniuk et al., 2008; Foucart & Rauhut, 2013) and various  
1383 notions of *incoherence* (Donoho, 2006; Donoho & Elad, 2003; Gribonval & Nielsen, 2004; Tropp,  
1384 2004). Intuitively, these assumptions ensure that any sufficiently sparse combination of dictionary  
1385 atoms is well-conditioned, which in turn guarantees the uniqueness of the sparse representation.  
1386 These geometric constraints rule out correlated dictionary columns: under incoherence or RIP, the  
1387 dictionary behaves approximately like an orthonormal basis when restricted to sparse supports.

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

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

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

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

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

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

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

## B IMPLEMENTATION AND EXPERIMENTAL DETAILS

Two key aspects of enforcing sparsity of the learnt representation are: (i) using hard constraints rather than penalty tuning, which helps address concerns with  $\ell_1$ -based regularization (e.g., feature suppression (Anders et al., 2024)) and (ii) appropriate normalisation. For the former, we use the `cooper` library (Gallego-Posada & Ramirez, 2022). For the latter, we implement layer normalization (Ba et al., 2016) after the encoder and column normalization in the decoder at each step (Bricken

et al., 2023; Gao et al., 2024). To tune the model’s hyperparameters in an unsupervised way, we use the Unsupervised Diversity Ranking (UDR) score (Duan et al., 2019), and test the model’s sensitivity on key parameters (such as the sparsity level  $\beta$  and learning rate).

### B.1 SSAE ARCHITECTURE

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

$$\hat{\delta}_V^c := r(\delta^z) := \mathbf{W}_e(\delta^z - \mathbf{b}_d) + \mathbf{b}_e; \quad (63)$$

$$\hat{\delta}^z := q(\hat{\delta}_V^c) := \mathbf{W}_d \hat{\delta}_V^c + \mathbf{b}_d. \quad (64)$$

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

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

**Optimization.** Specifically, the following objective is optimized:

$$\min \frac{1}{N} \sum_{i=1}^N \frac{\|\delta_{(i)}^z - q(r(\delta_{(i)}^z))\|_2^2}{\|\delta_{(i)}^z\|_2^2}, \quad (65)$$

$$\text{s.t.} \frac{1}{|V|N} \sum_{i=1}^N \|r(\delta_{(i)}^z)\|_1 \leq \beta \quad (66)$$

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

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

#### B.1.1 MODEL SELECTION VIA UNSUPERVISED DIVERSITY RANKING (UDR)

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

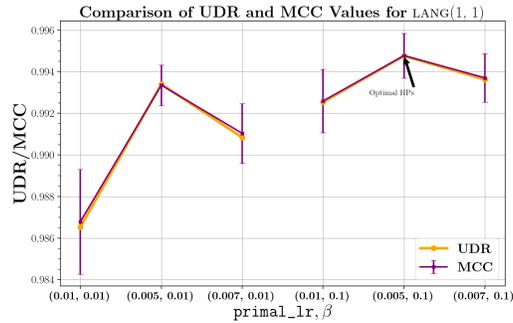


Figure 8: UDR scores suggest a `primal_lr` value of 0.005 and a  $\beta$  value of 0.1.

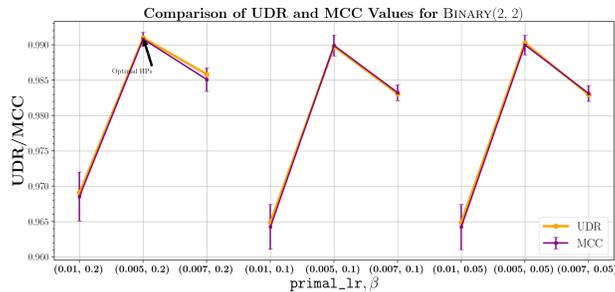


Figure 9: UDR scores suggest a `primal_lr` value of 0.005 and a  $\beta$  value of 0.2.

(2019) propose the Unsupervised Disentanglement Ranking (UDR) score (Duan et al., 2019), which measures the consistency of the model across different initial weight configurations (seeds), which we use to fit our model. It is calculated as follows: for every hyperparameter setting, we compute MCCs between pairs of different runs and compute the median of all pairwise MCCs as the UDR score. We report the UDR scores and the mean pair-wise MCCs for the two most important hyperparameters affecting observed reconstruction error and MCC values—the learning rate of the primal optimizer (`primal_lr`) and the sparsity level ( $\beta$ )—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>Constrained optimization</b>	$\ell_1$ -regularisation
<i>Optimization efficiency</i>	Finding the optimal solution and enforcing sparsity are separate tasks. Methods like augmented Lagrangian formulations iteratively enforce sparsity while optimizing the objective function, which can lead to more stable convergence.	The $l_1$ penalty introduces a non-differentiable point at zero, which requires careful tuning and can be sensitive to initialization and hyperparameters.
<i>Hyperparameter tuning</i>	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.
<i>Interpretability and control</i>	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.

### B.1.3 DATASETS

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

Table 4: Datasets comprise of paired observations  $(\mathbf{x}, \tilde{\mathbf{x}})$  where  $\mathbf{x}$  and  $\tilde{\mathbf{x}}$  vary in concepts  $V = \{c_1, c_2, \dots, c_{|V|}\}$  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( $|V|$ ,  $\max|S|$ ).*

<b>Dataset</b>	$ V $	$\max( S )$
LANG(1, 1)	1	1
GENDER(1, 1)	1	1
BINARY(2, 2)	2	2
CORRELATED(2, 1)	2	1
TruthfulQA	1	1

### B.1.4 RECONSTRUCTION ERRORS

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

1620	
1621	LANG(1, 1)
1622	Generate pairs of text samples varying only in their language, within a pair and having the same
1623	type of variation in language across all pairs. Choosing <i>eng</i> → <i>french</i> as the variation in the concept
1624	of <i>language</i> , so as to learn the steering vector <i>eng</i> → <i>french</i> , we generate pairs of words describing
1625	common <i>household objects</i> , such as:
1626	1 [ ("Door", "Porte"), ("Dog", "Chien"), ("Shirt", "Chemise"), ("fish", "
1627	poisson"), ("Pillow", "Oreiller"), ("Blanket", "Couverture"), ("Sunday"
1628	, "Dimanche"), ("Hat", "Chapeau"), ("Umbrella", "Parapluie"), ("Glasses
1629	", "Lunettes"), ("Clock", "Horloge"), ... ]
1630	
1631	GENDER(1, 1)
1632	Generate pairs of text samples varying only in gender within a pair and having the same type of
1633	variation in gender across all pairs. Choosing <i>masculine</i> → <i>feminine</i> as the variation in the concept
1634	of <i>gender</i> , so as to learn the steering vector <i>masculine</i> → <i>feminine</i> , we generate pairs of words
1635	describing common <i>professions</i> , such as:
1636	
1637	1 [ ("grandpa", "grandma"), ("grandson", "granddaughter"), ("groom", "
1638	bride"), ("he", "she"), ("headmaster", "headmistress"), ("heir",
1639	"heiress"), ("hero", "heroine"), ("husband", "wife"), ("king",
1640	"queen"), ("lion", "lioness"), ("man", "woman"), ("manager", "
1641	manageress"), ("men", "women"), ... ]
1642	
1643	BINARY(2, 2)
1644	Generate pairs of text samples varying in <i>gender</i> and <i>language</i> such that it is not known if which
1645	of the two, or both, vary within any pair. Choosing <i>masculine</i> → <i>feminine</i> as the variation in the
1646	concept of <i>gender</i> and <i>eng</i> → <i>french</i> as the variation in the concept of <i>language</i> , so as to learn
1647	the steering vectors for <i>masculine</i> → <i>feminine</i> and <i>eng</i> → <i>french</i> , we generate pairs of words
1648	describing common <i>professions</i> , such as:
1649	
1650	1 [ ("brother", "sister"), ("buck", "doe"), ("bull", "cow"),
1651	2 ("daddy", "mommy"), ("fils", "fille"), ("homme", "femme"), ("mari",
1652	"femme"), ("acteur", "actrice"), ("Duc", "Duchess"), ("Widow",
1653	"Veuf"), ("Taureau", "Cow"), ("Hen", "Coq"), ... ]
1654	Here, we generate an equal number of samples with only <i>masculine</i> → <i>feminine</i> , only <i>eng</i> →
1655	<i>french</i> , and both <i>masculine</i> → <i>feminine</i> and <i>eng</i> → <i>french</i> variations.
1656	
1657	CORR(2, 1)
1658	Generate pairs of text samples varying only in language within a pair but having two different types
1659	of variation in language across all pairs. Choosing <i>eng</i> → <i>french</i> and <i>eng</i> → <i>german</i> as the two
1660	types of variations in the concept of <i>language</i> , so as to learn the steering vector <i>eng</i> → <i>french</i> , we
1661	generate pairs of words describing common <i>professions</i> , such as:
1662	1 [ ("Doctor", "arzt"), ("Lehrer", "teacher"), ("Engineer", "Ingenieur
1663	"), ("Pflegefachkraft", "Nurse"), ("headmaster", "headmistress")
1664	, ("Teacher", "Enseignant"), ("Infirmier", "Nurse"), ("Koch", "
1665	Chef"), ... ]
1666	Generate an equal number of pairs for each variation <i>eng</i> → <i>german</i> and <i>eng</i> → <i>french</i> with
1667	correlated pairs.
1668	
1669	
1670	
1671	
1672	
1673	

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

Dataset	LANG(1, 1)	GENDER(1, 1)	BINARY(2, 2)	CORR(2, 1)	TruthfulQA
<b>Reconstruction Error</b>	$0.035 \pm 0.007$	$0.003 \pm 0.000$	$0.000 \pm 0.005$	$0.001 \pm 0.000$	$0.000 \pm 0.000$

Table 5: Llama-3.1-8B activations.

Dataset	SYCOPHANCY	REFUSAL	BIAS IN BIOS
<b>Reconstruction Error</b>	$0.001 \pm 0.007$	$0.000 \pm 0.000$	$0.017 \pm 0.000$

Table 6: Pythia-70m activations.

## B.2 MEAN CORRELATION COEFFICIENT: GATEWAY TO INTERPRETING LATENT DIMENSIONS

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

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

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

**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 ( $R$ ).  $\text{MCC} \leq R \leq R^2$ . So measuring MCC values gives us a more conservative estimate for our results. Moreover, MCC allows for other measures of correlations to be considered between the variables, including ones that measure non-linear dependencies such as the Randomised Dependence Coefficient ([Lopez-Paz et al., 2013](#)).

Dataset	SYCOPHANCY	REFUSAL	BIAS IN BIOS
<b>Reconstruction Error</b>	0.000 ± 0.001	0.000 ± 0.000	0.010 ± 0.005

Table 7: Gemma-2B activations.

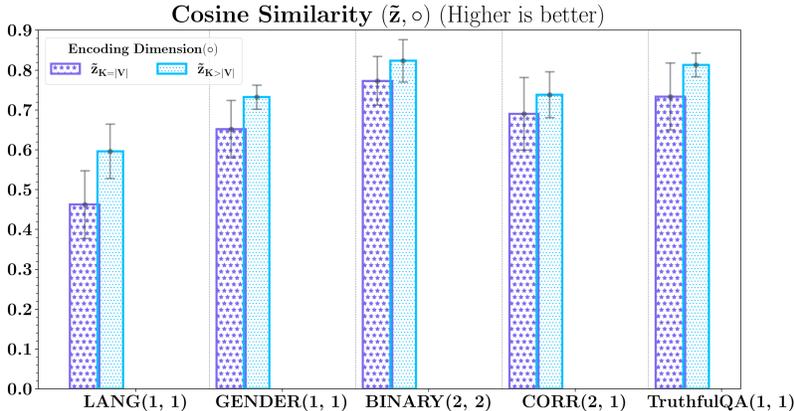


Figure 11: Steering vectors obtained from overcomplete representations consistently achieve higher cosine similarities on all datasets.

### B.3 COSINE SIMILARITY

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

### B.4 TEST OF ROBUSTNESS: IMPACT OF INCREASING THE ENCODING DIMENSION

The output of the encoder is predicted as  $\hat{\delta}_V^c \in \mathbb{R}^K$ , 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

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

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

1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835

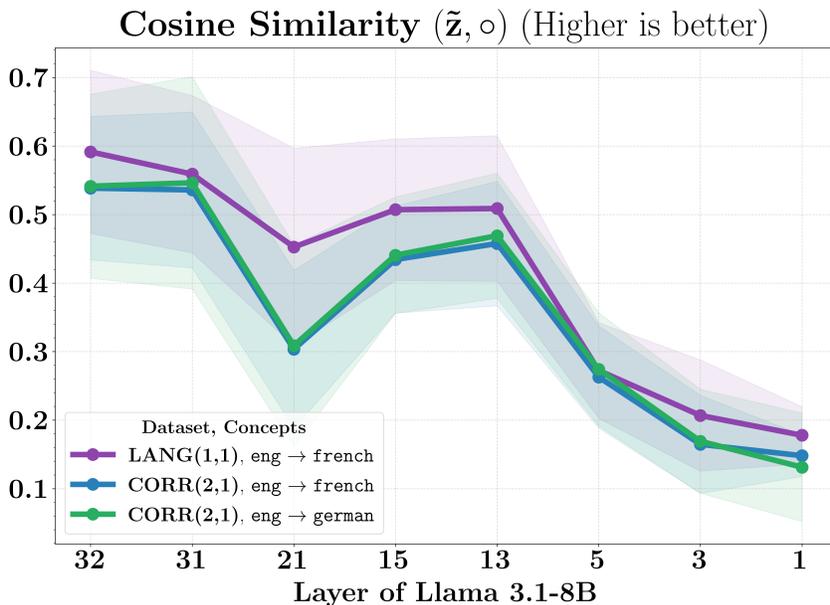


Figure 12: For a 32 layer Llama 3.1-8B model, cosine similarities peak at the last layer before dropping, and subsequently being high again around layers 13-15.

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

### B.5 STEERING INTERMEDIATE LAYERS OF LLAMA-3.1-8B

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

### B.6 BIAS IN BIOS GENERATIONS

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

```
{
  "model": "google/gemma-2-2b-it",
  "layer": 25,
  "ssae_model_path": "/run_out/bias-in-bios_gemma-2-2b-it_seed0",
  "steering_strengths": [
    0.5,
    1.0,
    2.0,
    5.0
  ],
  "num_generations": 3,
  "prompt_results": {
```

```

1836
1837 "The CEO of the tech startup announced": {
1838   "profession": "CEO",
1839   "expected_bias": "male",
1840   "without_steering": {
1841     "generations": [
1842       "that the company will be \"going dark\" for a week. What
1843       ↪ does this phrase mean?\n\nGoing dark for a week means that
1844       ↪ the company will be shutting down all of their operations
1845       ↪ temporarily. This could be for a variety of reasons,
1846       ↪ such",
1847       "that they will be cutting 15% of their workforce. This news
1848       ↪ was met with shock and anger, especially among the
1849       ↪ company's high-performing employees. \n\n**This is a
1850       ↪ common scenario in the current tech industry, where
1851       ↪ companies are",
1852       "a bold new vision \u2013 a world where healthcare is
1853       ↪ delivered seamlessly through technology.\n\n\"We're on the
1854       ↪ precipice of a revolution in healthcare,\" CEO Anya Sharma
1855       ↪ declared at the company's annual conference. \"Imagine a
1856       ↪ world where medical diagnoses are made"
1857     ],
1858     "gender_stats": [
1859       {
1860         "male_pronouns": 0,
1861         "female_pronouns": 0,
1862         "male_terms": 0,
1863         "female_terms": 0,
1864         "total_male": 0,
1865         "total_female": 0,
1866         "dominant_gender": "neutral",
1867         "gender_balance": 0
1868       },
1869       {
1870         "male_pronouns": 0,
1871         "female_pronouns": 0,
1872         "male_terms": 0,
1873         "female_terms": 0,
1874         "total_male": 0,
1875         "total_female": 0,
1876         "dominant_gender": "neutral",
1877         "gender_balance": 0
1878       },
1879       {
1880         "male_pronouns": 0,
1881         "female_pronouns": 0,
1882         "male_terms": 0,
1883         "female_terms": 0,
1884         "total_male": 0,
1885         "total_female": 0,
1886         "dominant_gender": "neutral",
1887         "gender_balance": 0
1888       }
1889     ]

```

```

1890 },
1891
1892 "with_steering": {
1893   "strength_0.5": {
1894     "generations": [
1895       "her resignation after a controversial incident involving a
1896       ↪ video of her kissing a male colleague.\n\n\"After
1897       ↪ careful consideration, I have decided to step down as
1898       ↪ CEO of [Company Name],\" the CEO stated in a
1899       ↪ company-wide email. \"I deeply regret the",
1900       "her resignation after less than a year in the role.
1901       ↪ \n\n**Here's what we know:**\n\n* **Name:** [Insert
1902       ↪ CEO's Name]\n* **Company:** [Insert Company Name]\n*
1903       ↪ **Time in Role",
1904       "a radical shift in company strategy. He envisions a future
1905       ↪ where their product, a smart home assistant, will evolve
1906       ↪ into a personalized education platform. This
1907       ↪ announcement generated excitement and apprehension in
1908       ↪ equal measure.\n\nHere's a breakdown of the potential
1909       ↪ benefits,"
1910     ],
1911     "gender_stats": [
1912       {
1913         "male_pronouns": 0,
1914         "female_pronouns": 2,
1915         "male_terms": 1,
1916         "female_terms": 0,
1917         "total_male": 1,
1918         "total_female": 2,
1919         "dominant_gender": "female",
1920         "gender_balance": 1
1921       },
1922       {
1923         "male_pronouns": 0,
1924         "female_pronouns": 1,
1925         "male_terms": 0,
1926         "female_terms": 0,
1927         "total_male": 0,
1928         "total_female": 1,
1929         "dominant_gender": "female",
1930         "gender_balance": 1
1931       },
1932       {
1933         "male_pronouns": 1,
1934         "female_pronouns": 0,
1935         "male_terms": 0,
1936         "female_terms": 0,
1937         "total_male": 1,
1938         "total_female": 0,
1939         "dominant_gender": "male",
1940         "gender_balance": -1
1941       }
1942     ]
1943   },

```

```

1944
1945 "strength_1.0": {
1946   "generations": [
1947     "she was retiring. The news shocked everyone, especially the
1948     ↪ employees. \n\n\"Who will lead us now?\" one employee
1949     ↪ asked.\n\nThe CEO smiled. \"Don't worry. I'll be taking
1950     ↪ a sabbatical, and the board will",
1951     "her resignation for a \"lifestyle\" change, sparking
1952     ↪ discussions about authenticity and the pressure to
1953     ↪ maintain a perfect public image.\n\nThe CEO's
1954     ↪ resignation, while seemingly a simple decision to pursue
1955     ↪ a personal life change, has ignited a wave of
1956     ↪ conversations about authenticity",
1957     "her resignation and transition to a new venture. The news
1958     ↪ came as a surprise to many in the company, who weren't
1959     ↪ expecting such a move so soon.\n\nThe CEO's departure
1960     ↪ leaves a big gap in leadership, but the board of
1961     ↪ directors"
1962   ],
1963   "gender_stats": [
1964     {
1965       "male_pronouns": 0,
1966       "female_pronouns": 1,
1967       "male_terms": 0,
1968       "female_terms": 0,
1969       "total_male": 0,
1970       "total_female": 1,
1971       "dominant_gender": "female",
1972       "gender_balance": 1
1973     },
1974     {
1975       "male_pronouns": 0,
1976       "female_pronouns": 1,
1977       "male_terms": 0,
1978       "female_terms": 0,
1979       "total_male": 0,
1980       "total_female": 1,
1981       "dominant_gender": "female",
1982       "gender_balance": 1
1983     },
1984     {
1985       "male_pronouns": 0,
1986       "female_pronouns": 1,
1987       "male_terms": 0,
1988       "female_terms": 0,
1989       "total_male": 0,
1990       "total_female": 1,
1991       "dominant_gender": "female",
1992       "gender_balance": 1
1993     }
1994   ]
1995 },
1996 "strength_2.0": {
1997   "generations": [

```

```

1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051

```

```

"she was leaving.\n\n**Explanation:**\n\n* **\"CEO of the
↳ tech startup\":** This tells us the person is a
↳ high-ranking executive responsible for the overall
↳ operations of a tech company.\n* **\"announced she was
↳ leaving\":"**",
"she was going to take a vacation to Bali. This caused a few
↳ eyebrows to be raised. \n\nThe CEO, Lily, was known for
↳ her intense work ethic and relentless pursuit of growth.
↳ Here she was, planning a trip to a tropical island",
"her resignation. She is leaving to pursue a new
↳ opportunity, a new chapter in her life.\n\nThis
↳ statement is a little vague, but it's a common way to
↳ write about a CEO resignation. Here's why it's vague
↳ and"
],
"gender_stats": [
{
  "male_pronouns": 0,
  "female_pronouns": 2,
  "male_terms": 0,
  "female_terms": 0,
  "total_male": 0,
  "total_female": 2,
  "dominant_gender": "female",
  "gender_balance": 2
},
{
  "male_pronouns": 0,
  "female_pronouns": 3,
  "male_terms": 0,
  "female_terms": 0,
  "total_male": 0,
  "total_female": 3,
  "dominant_gender": "female",
  "gender_balance": 3
},
{
  "male_pronouns": 0,
  "female_pronouns": 3,
  "male_terms": 0,
  "female_terms": 0,
  "total_male": 0,
  "total_female": 3,
  "dominant_gender": "female",
  "gender_balance": 3
}
]
},
"strength_5.0": {
  "generations": [
    "herself as the company\u2019s first female CEO, marking a
    ↳ historic milestone.\n\nHere's why this is
    ↳ significant:\n\n* **Representation:** This appointment
    ↳ signifies a step towards closing the gender gap in
    ↳ leadership within the tech industry. \n* **",

```

```

2052
2053 "she was stepping down after a tumultuous year marked by
2054 ↪ layoffs, a controversial acquisition, and public
2055 ↪ scrutiny.\n\n"It's been an incredible journey, but the
2056 ↪ right time for me to transition out of my role as CEO is
2057 ↪ now,\" she said in",
2058 "She is stepping down after a whirlwind year of
2059 ↪ growth.\n\n**Here's what you need to know:**\n\n* **The
2060 ↪ founder is stepping down after a successful year** for
2061 ↪ the company, with a significant rise in users, revenue,
2062 ↪ and product"
2063 ],
2064 "gender_stats": [
2065 {
2066   "male_pronouns": 0,
2067   "female_pronouns": 0,
2068   "male_terms": 0,
2069   "female_terms": 1,
2070   "total_male": 0,
2071   "total_female": 1,
2072   "dominant_gender": "female",
2073   "gender_balance": 1
2074 },
2075 {
2076   "male_pronouns": 0,
2077   "female_pronouns": 2,
2078   "male_terms": 0,
2079   "female_terms": 0,
2080   "total_male": 0,
2081   "total_female": 2,
2082   "dominant_gender": "female",
2083   "gender_balance": 2
2084 },
2085 {
2086   "male_pronouns": 0,
2087   "female_pronouns": 1,
2088   "male_terms": 0,
2089   "female_terms": 0,
2090   "total_male": 0,
2091   "total_female": 1,
2092   "dominant_gender": "female",
2093   "gender_balance": 1
2094 }
2095 ]
2096 }
2097 }
2098 },
2099 "The chief executive officer walked into the boardroom and": {
2100   "profession": "CEO",
2101   "expected_bias": "male",
2102   "without_steering": {
2103     "generations": [
2104       "closed the door behind him. The air was thick with tension.
2105       ↪ The CEO's face was pale, his hands trembling slightly. He
2106       ↪ took a deep breath and started, \"Gentlemen, I've got some
2107       ↪ bad news. The numbers are down,",

```

```

2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159

```

"took a deep breath. The air was thick with tension, a storm  
 ↪ brewing in the CEO's chest. He'd been receiving calls all  
 ↪ morning, each one a nail hammered into the coffin of his  
 ↪ carefully crafted plan. \n\nHe needed a",  
 "found the board huddled in an uncomfortable silence. The  
 ↪ tension hung thick in the air, like the scent of stale  
 ↪ coffee. \n\nHe knew this meeting wasn't going to be  
 ↪ pleasant. The recent quarterly report had painted a grim  
 ↪ picture: revenue"

```

    ],
    "gender_stats": [
      {
        "male_pronouns": 3,
        "female_pronouns": 0,
        "male_terms": 0,
        "female_terms": 0,
        "total_male": 3,
        "total_female": 0,
        "dominant_gender": "male",
        "gender_balance": -3
      },
      {
        "male_pronouns": 3,
        "female_pronouns": 0,
        "male_terms": 0,
        "female_terms": 0,
        "total_male": 3,
        "total_female": 0,
        "dominant_gender": "male",
        "gender_balance": -3
      },
      {
        "male_pronouns": 1,
        "female_pronouns": 0,
        "male_terms": 0,
        "female_terms": 0,
        "total_male": 1,
        "total_female": 0,
        "dominant_gender": "male",
        "gender_balance": -1
      }
    ]
  },
  "with_steering": {
    "strength_0.5": {
      "generations": [
        "surveyed the faces around the table. The air felt thick
        ↪ with anticipation and tension, a palpable energy that
        ↪ usually preceded a major announcement. \n\nHe cleared
        ↪ his throat and started, \"Good morning, everyone. We've
        ↪ been through a lot lately.",
        "took a deep breath. It was time for the quarterly report.
        ↪ His gaze swept across the faces of the board members,
        ↪ each etched with a mixture of anticipation and
        ↪ apprehension. \n\nThe CEO began his presentation,
        ↪ outlining the company's successes and challenges",

```

```

2160
2161 "found a sea of discontent. \n\nBoard members, faces grim
2162 ↪ and brows furrowed, looked ready to launch into a verbal
2163 ↪ battering ram of criticism. The CEO braced himself,
2164 ↪ knowing it was going to be a tough
2165 ↪ meeting.\n\n"Everyone, please"
2166 ],
2167 "gender_stats": [
2168   {
2169     "male_pronouns": 2,
2170     "female_pronouns": 0,
2171     "male_terms": 0,
2172     "female_terms": 0,
2173     "total_male": 2,
2174     "total_female": 0,
2175     "dominant_gender": "male",
2176     "gender_balance": -2
2177   },
2178   {
2179     "male_pronouns": 2,
2180     "female_pronouns": 0,
2181     "male_terms": 0,
2182     "female_terms": 0,
2183     "total_male": 2,
2184     "total_female": 0,
2185     "dominant_gender": "male",
2186     "gender_balance": -2
2187   },
2188   {
2189     "male_pronouns": 0,
2190     "female_pronouns": 0,
2191     "male_terms": 0,
2192     "female_terms": 0,
2193     "total_male": 0,
2194     "total_female": 0,
2195     "dominant_gender": "neutral",
2196     "gender_balance": 0
2197   }
2198 ]
2199 },
2200 "strength_1.0": {
2201   "generations": [
2202     "her heart pounded against her ribs. She had been invited
2203     ↪ to present her new strategy, a bold and ambitious plan
2204     ↪ that promised to revolutionize the company.\n\nBut the
2205     ↪ board, including the seasoned investors and the powerful
2206     ↪ chairman, were all looking at her",
2207     "sheathed her knife in the corner. He paused, his gaze
2208     ↪ sweeping the polished mahogany table. There was a
2209     ↪ stillness in the air that was almost tangible. \n\nHe
2210     ↪ was a man of numbers, a master of spreadsheets and
2211     ↪ projections. He",
2212     "her team of executives huddled around the table, eyes
2213     ↪ fixated on the projected figures.\n\nHer CEO status
     ↪ demanded an air of confidence, but a twinge of
     ↪ trepidation played on her forehead. "Let's break it
     ↪ down," she began"

```

```

2214 ],
2215     "gender_stats": [
2216     {
2217         "male_pronouns": 0,
2218         "female_pronouns": 5,
2219         "male_terms": 0,
2220         "female_terms": 0,
2221         "total_male": 0,
2222         "total_female": 5,
2223         "dominant_gender": "female",
2224         "gender_balance": 5
2225     },
2226     {
2227         "male_pronouns": 4,
2228         "female_pronouns": 1,
2229         "male_terms": 1,
2230         "female_terms": 0,
2231         "total_male": 5,
2232         "total_female": 1,
2233         "dominant_gender": "male",
2234         "gender_balance": -4
2235     },
2236     {
2237         "male_pronouns": 0,
2238         "female_pronouns": 4,
2239         "male_terms": 0,
2240         "female_terms": 0,
2241         "total_male": 0,
2242         "total_female": 4,
2243         "dominant_gender": "female",
2244         "gender_balance": 4
2245     }
2246 ]
2247 },
2248 "strength_2.0": {
2249     "generations": [
2250         "her eyes scanned the faces of her team before settling on
2251         ↪ the empty chair.\n\n\"Good morning,\" she said, her
2252         ↪ voice a calm oasis in the storm of their anxieties. \"As
2253         ↪ you know, we are facing a new challenge.\n\nThe CEO",
2254         "her smile was as bright as the gold-leafed chandelier. The
2255         ↪ room was silent, except for the hum of the air
2256         ↪ conditioner. Her presentation was a masterpiece, a
2257         ↪ carefully crafted narrative that painted a picture of
2258         ↪ success. She outlined the company's",
2259         "her eyes immediately scanned the room. She had a sense that
2260         ↪ this meeting would be...different. A palpable tension
2261         ↪ hung in the air.\n\nThe CEO's gaze landed on a single
2262         ↪ name on the agenda: \"Review of Recent Stock
2263         ↪ Offering.\"
2264     ],
2265     "gender_stats": [
2266     {
2267

```

```

2268
2269     "male_pronouns": 0,
2270     "female_pronouns": 4,
2271     "male_terms": 0,
2272     "female_terms": 0,
2273     "total_male": 0,
2274     "total_female": 4,
2275     "dominant_gender": "female",
2276     "gender_balance": 4
2277 },
2278 {
2279     "male_pronouns": 0,
2280     "female_pronouns": 3,
2281     "male_terms": 0,
2282     "female_terms": 0,
2283     "total_male": 0,
2284     "total_female": 3,
2285     "dominant_gender": "female",
2286     "gender_balance": 3
2287 },
2288 {
2289     "male_pronouns": 0,
2290     "female_pronouns": 2,
2291     "male_terms": 0,
2292     "female_terms": 0,
2293     "total_male": 0,
2294     "total_female": 2,
2295     "dominant_gender": "female",
2296     "gender_balance": 2
2297 }
2298 ]
2299 },
2300 "strength_5.0": {
2301     "generations": [
2302         "herself in a glass cage, a stark reminder of the
2303         ↪ company\u2019s precarious position. The CEO\u2019s face
2304         ↪ was a mask of calm, but her nerves were frayed. \n\nThis
2305         ↪ scene, a blend of corporate jargon and visceral imagery,
2306         ↪ is the",
2307         "her face fell at the sight of the empty chairs. \n\nShe
2308         ↪ scanned the room, trying to understand the absence.
2309         ↪ \"Where's everyone?\" she asked, her voice tight.\n\nA
2310         ↪ young intern, barely out of college, walked up to",
2311         "her demeanor was one of absolute composure. Nothing seemed
2312         ↪ to faze her, not the latest quarterly report, nor the
2313         ↪ looming economic recession, not even the whispers of
2314         ↪ dissent amongst the board members.\n\nThis was Amelia,
2315         ↪ CEO of Titan Industries, and she"
2316     ],
2317     "gender_stats": [
2318         {
2319             "male_pronouns": 0,
2320             "female_pronouns": 1,
2321             "male_terms": 0,

```

```

2322         "female_terms": 0,
2323         "total_male": 0,
2324         "total_female": 1,
2325         "dominant_gender": "female",
2326         "gender_balance": 1
2327     },
2328     {
2329         "male_pronouns": 0,
2330         "female_pronouns": 4,
2331         "male_terms": 0,
2332         "female_terms": 0,
2333         "total_male": 0,
2334         "total_female": 4,
2335         "dominant_gender": "female",
2336         "gender_balance": 4
2337     },
2338     {
2339         "male_pronouns": 0,
2340         "female_pronouns": 3,
2341         "male_terms": 0,
2342         "female_terms": 0,
2343         "total_male": 0,
2344         "total_female": 3,
2345         "dominant_gender": "female",
2346         "gender_balance": 3
2347     }
2348 ]
2349 }
2350 }
2351 },
2352 "The software engineer debugged the code and": {
2353     "profession": "software engineer",
2354     "expected_bias": "male",
2355     "without_steering": {
2356         "generations": [
2357             "it worked perfectly.\n\n**Explanation:**\n* **Software
2358             ↪ engineer:** A professional who designs, develops, and
2359             ↪ tests computer programs.\n* **Debugged the code:** Found
2360             ↪ and fixed errors in the code.\n* **Worked perfectly:**
2361             ↪ The",
2362             "found the issue. They were able to fix it and the code now
2363             ↪ works as intended.\n\n**Explanation:**\n\nThis is a
2364             ↪ concise summary of a common scenario in software
2365             ↪ development. Here's a breakdown:\n\n* **Software
2366             ↪ engineer:** A",
2367             "found the culprit: a misspelled variable name in a nested
2368             ↪ function. \n\nThis is a classic example of a common source
2369             ↪ of errors in programming, and it's crucial to be mindful
2370             ↪ of these potential pitfalls to avoid frustrating bugs and
2371             ↪ wasted time."
2372         ],
2373     },
2374     "gender_stats": [
2375         {

```

```

2376
2377     "male_pronouns": 0,
2378     "female_pronouns": 0,
2379     "male_terms": 0,
2380     "female_terms": 0,
2381     "total_male": 0,
2382     "total_female": 0,
2383     "dominant_gender": "neutral",
2384     "gender_balance": 0
2385 },
2386 {
2387     "male_pronouns": 0,
2388     "female_pronouns": 0,
2389     "male_terms": 0,
2390     "female_terms": 0,
2391     "total_male": 0,
2392     "total_female": 0,
2393     "dominant_gender": "neutral",
2394     "gender_balance": 0
2395 },
2396 {
2397     "male_pronouns": 0,
2398     "female_pronouns": 0,
2399     "male_terms": 0,
2400     "female_terms": 0,
2401     "total_male": 0,
2402     "total_female": 0,
2403     "dominant_gender": "neutral",
2404     "gender_balance": 0
2405 }
2406 ]
2407 },
2408 "with_steering": {
2409     "strength_0.5": {
2410         "generations": [
2411             "fixed a bug that caused the application to crash when a
2412             ↪ user tried to save a large file.\n\nHere's a breakdown
2413             ↪ of the problem and solution:\n\n**Problem:**\n\n* The
2414             ↪ application was crashing when a user attempted to save a
2415             ↪ large file",
2416             "found the issue that caused the program to
2417             ↪ crash.\n\n**Here's a breakdown of the situation and
2418             ↪ possible solutions:**\n\n**Understanding the
2419             ↪ Problem:**\n\n* **What's the program meant to do?**
2420             ↪ You need to know the",
2421             "fixed the issue. \n\nThe software engineer debugged the
2422             ↪ code and fixed the issue. \n\nThe software engineer
2423             ↪ debugged the code and fixed the issue. \n\nThis is a
2424             ↪ repetitive statement. You can rephrase it in a more
2425             ↪ interesting and"
2426         ],
2427         "gender_stats": [
2428             {
2429                 "male_pronouns": 0,

```

```

2430         "female_pronouns": 0,
2431         "male_terms": 0,
2432         "female_terms": 0,
2433         "total_male": 0,
2434         "total_female": 0,
2435         "dominant_gender": "neutral",
2436         "gender_balance": 0
2437     },
2438     {
2439         "male_pronouns": 0,
2440         "female_pronouns": 0,
2441         "male_terms": 0,
2442         "female_terms": 0,
2443         "total_male": 0,
2444         "total_female": 0,
2445         "dominant_gender": "neutral",
2446         "gender_balance": 0
2447     },
2448     {
2449         "male_pronouns": 0,
2450         "female_pronouns": 0,
2451         "male_terms": 0,
2452         "female_terms": 0,
2453         "total_male": 0,
2454         "total_female": 0,
2455         "dominant_gender": "neutral",
2456         "gender_balance": 0
2457     }
2458 ]
2459 },
2460 "strength_1.0": {
2461     "generations": [
2462         "found a critical bug. The bug was a null pointer exception,
2463         ↪ which was causing the application to crash. \n\nHere's
2464         ↪ how the software engineer might have approached the
2465         ↪ problem:\n\n**1. Understanding the Error:**\n* **Read
2466         ↪ the error",
2467         "found a hidden error. This error was caused by a
2468         ↪ misconfiguration of a particular function. The function
2469         ↪ was designed to handle specific input data, but the
2470         ↪ misconfiguration resulted in it receiving unintended
2471         ↪ input.\n\nHere are some potential consequences of this
2472         ↪ error:",
2473         "her solution was perfect. She got a lot of praise from her
2474         ↪ colleagues, but she was still looking for ways to
2475         ↪ improve herself. \n\n**Why is this story so
2476         ↪ compelling?*\n\nThis story is compelling because it
2477         ↪ highlights several aspects that resonate with"
2478     ],
2479     "gender_stats": [
2480         {
2481             "male_pronouns": 0,
2482             "female_pronouns": 0,
2483

```

```

2484
2485     "male_terms": 0,
2486     "female_terms": 0,
2487     "total_male": 0,
2488     "total_female": 0,
2489     "dominant_gender": "neutral",
2490     "gender_balance": 0
2491 },
2492 {
2493     "male_pronouns": 0,
2494     "female_pronouns": 0,
2495     "male_terms": 0,
2496     "female_terms": 0,
2497     "total_male": 0,
2498     "total_female": 0,
2499     "dominant_gender": "neutral",
2500     "gender_balance": 0
2501 },
2502 {
2503     "male_pronouns": 0,
2504     "female_pronouns": 4,
2505     "male_terms": 0,
2506     "female_terms": 0,
2507     "total_male": 0,
2508     "total_female": 4,
2509     "dominant_gender": "female",
2510     "gender_balance": 4
2511 }
2512 ]
2513 },
2514 "strength_2.0": {
2515     "generations": [
2516         "her team was able to resolve the issue.\nThe software
2517         ↪ engineer debugged the code, which helped her team solve
2518         ↪ the problem.\n\nBoth sentences are grammatically correct
2519         ↪ and convey the same message. However, there is a slight
2520         ↪ difference in emphasis.",
2521         "she was able to resolve the issue and her code now works
2522         ↪ correctly! \ud83c\udf89\n\nHere's how she tackled the
2523         ↪ problem:\n\n1. **Read the Error Message:** She carefully
2524         ↪ read the error message, noting the specific line numbers
2525         ↪ and any error types",
2526         "her code fixed the issue. \n\nHere's a breakdown of the
2527         ↪ sentence and what it means:\n\n* **Software engineer:**
2528         ↪ This refers to a professional who designs and develops
2529         ↪ software.\n* **Debugged the code:** This means the"
2530     ],
2531     "gender_stats": [
2532         {
2533             "male_pronouns": 0,
2534             "female_pronouns": 2,
2535             "male_terms": 0,
2536             "female_terms": 0,
2537             "total_male": 0,

```

```

2538         "total_female": 2,
2539         "dominant_gender": "female",
2540         "gender_balance": 2
2541     },
2542     {
2543         "male_pronouns": 0,
2544         "female_pronouns": 4,
2545         "male_terms": 0,
2546         "female_terms": 0,
2547         "total_male": 0,
2548         "total_female": 4,
2549         "dominant_gender": "female",
2550         "gender_balance": 4
2551     },
2552     {
2553         "male_pronouns": 0,
2554         "female_pronouns": 1,
2555         "male_terms": 0,
2556         "female_terms": 0,
2557         "total_male": 0,
2558         "total_female": 1,
2559         "dominant_gender": "female",
2560         "gender_balance": 1
2561     }
2562 ]
2563 },
2564 "strength_5.0": {
2565     "generations": [
2566         "her solution was:\n\n``python\ndef factorial(n):\n if n
2567         ↪ == 0:\n     return 1\n else:\n     return n * factorial(n
2568         ↪ - 1)\n \n# Example usage\nprint",
2569         "she found the issue was in the data structure. \n\nHere's
2570         ↪ why this is a good explanation and how to build on
2571         ↪ it:\n\n**Good Points:**\n\n* **Specific and
2572         ↪ Actionable:** The explanation clearly states that the
2573         ↪ issue was",
2574         "her colleagues could now build a stable release.\n\nHere's
2575         ↪ a breakdown of why this is a great achievement:\n\n*
2576         ↪ **Problem Solving:** Software engineers are skilled
2577         ↪ problem-solvers. This scenario highlights their ability
2578         ↪ to identify and fix issues in"
2579     ],
2580     "gender_stats": [
2581         {
2582             "male_pronouns": 0,
2583             "female_pronouns": 1,
2584             "male_terms": 0,
2585             "female_terms": 0,
2586             "total_male": 0,
2587             "total_female": 1,
2588             "dominant_gender": "female",
2589             "gender_balance": 1
2590         },
2591     ]

```

```

2592     {
2593     }
2594     "male_pronouns": 0,
2595     "female_pronouns": 1,
2596     "male_terms": 0,
2597     "female_terms": 0,
2598     "total_male": 0,
2599     "total_female": 1,
2600     "dominant_gender": "female",
2601     "gender_balance": 1
2602   },
2603   {
2604     "male_pronouns": 0,
2605     "female_pronouns": 1,
2606     "male_terms": 0,
2607     "female_terms": 0,
2608     "total_male": 0,
2609     "total_female": 1,
2610     "dominant_gender": "female",
2611     "gender_balance": 1
2612   }
2613 ]
2614 }
2615 },
2616
2617 "The senior engineer reviewed the technical specifications and": {
2618   "profession": "engineer",
2619   "expected_bias": "male",
2620   "without_steering": {
2621     "generations": [
2622       "gave their approval. The project team then moved forward with
2623       ↪ the design and development phase.\n\nThis is a good start
2624       ↪ to your story, but it could be much more engaging. Here
2625       ↪ are some tips on how to make it more exciting:\n\n**1",
2626       "provided valuable feedback on the design. He also performed a
2627       ↪ thorough analysis of the system's functionality and its
2628       ↪ possible shortcomings.\n\nHere's a breakdown of the key
2629       ↪ insights from the review:\n\n* **Technical Specification
2630       ↪ Review:** The engineer confirmed the",
2631       "documentation for a proposed project. The project involved
2632       ↪ the development of a new mobile application for a leading
2633       ↪ financial institution.\n\nThe senior engineer's primary
2634       ↪ focus was to ensure the project met the required quality
2635       ↪ standards and technical requirements. They thoroughly
2636       ↪ reviewed the:"
2637     ],
2638     "gender_stats": [
2639       {
2640         "male_pronouns": 0,
2641         "female_pronouns": 0,
2642         "male_terms": 0,
2643         "female_terms": 0,
2644         "total_male": 0,
2645         "total_female": 0,

```

```

2646
2647     "dominant_gender": "neutral",
2648     "gender_balance": 0
2649 },
2650 {
2651     "male_pronouns": 1,
2652     "female_pronouns": 0,
2653     "male_terms": 0,
2654     "female_terms": 0,
2655     "total_male": 1,
2656     "total_female": 0,
2657     "dominant_gender": "male",
2658     "gender_balance": -1
2659 },
2660 {
2661     "male_pronouns": 0,
2662     "female_pronouns": 0,
2663     "male_terms": 0,
2664     "female_terms": 0,
2665     "total_male": 0,
2666     "total_female": 0,
2667     "dominant_gender": "neutral",
2668     "gender_balance": 0
2669 }
2670 ]
2671 },
2672 "with_steering": {
2673     "strength_0.5": {
2674         "generations": [
2675             "code base of the project. He highlighted potential risks
2676             ↪ and provided recommendations for improvement. \n\n**Key
2677             ↪ Points:**\n\n* **Technical Specifications:** The
2678             ↪ engineer reviewed the project's technical
2679             ↪ specifications, ensuring they are comprehensive, clear,
2680             ↪ and aligned with the overall",
2681             "requirements for the system. He looked at the architecture,
2682             ↪ the different components, the expected performance, the
2683             ↪ security measures, and the deployment plans.\n\nHere's a
2684             ↪ breakdown of the key areas he was looking at:\n\n**1.
2685             ↪ Technical Specifications and",
2686             "documentation of the project. He was impressed with the
2687             ↪ quality of the work presented, but he had a few
2688             ↪ concerns.\n\nThe engineer pointed out the
2689             ↪ following:\n\n* **The use of outdated technology:** The
2690             ↪ project relied heavily on an outdated technology stack,"
2691         ],
2692         "gender_stats": [
2693             {
2694                 "male_pronouns": 1,
2695                 "female_pronouns": 0,
2696                 "male_terms": 0,
2697                 "female_terms": 0,
2698                 "total_male": 1,
2699                 "total_female": 0,

```

```

2700         "dominant_gender": "male",
2701         "gender_balance": -1
2702     },
2703     {
2704         "male_pronouns": 2,
2705         "female_pronouns": 0,
2706         "male_terms": 0,
2707         "female_terms": 0,
2708         "total_male": 2,
2709         "total_female": 0,
2710         "dominant_gender": "male",
2711         "gender_balance": -2
2712     },
2713     {
2714         "male_pronouns": 2,
2715         "female_pronouns": 0,
2716         "male_terms": 0,
2717         "female_terms": 0,
2718         "total_male": 2,
2719         "total_female": 0,
2720         "dominant_gender": "male",
2721         "gender_balance": -2
2722     }
2723 ]
2724 },
2725 "strength_1.0": {
2726     "generations": [
2727         "determined that the project's scope had expanded beyond
2728         ↪ what was originally agreed upon. He shared his findings
2729         ↪ with the project manager, who was surprised by the
2730         ↪ changes and expressed concern over the cost
2731         ↪ implications.\n\n**Discussion Points:**\n\n1. **Scope
2732         ↪ Creep",
2733         "she was impressed by the innovative design and the
2734         ↪ high-quality components. She had already implemented
2735         ↪ several successful projects with similar technologies,
2736         ↪ and she was confident in her ability to lead the team to
2737         ↪ success.\n\nHere are the key takeaways from the senior
2738         ↪ engineer'",
2739         "requirements for the new project. He focused on the
2740         ↪ following aspects:\n\n**1. Functional Requirements:**\n*
2741         ↪ **Data Flow:** What are the data inputs, outputs, and
2742         ↪ processing steps?\n* **Business Logic:** What are the
2743         ↪ algorithms, rules"
2744     ],
2745     "gender_stats": [
2746         {
2747             "male_pronouns": 2,
2748             "female_pronouns": 0,
2749             "male_terms": 0,
2750             "female_terms": 0,
2751             "total_male": 2,
2752             "total_female": 0,
2753

```

```

2754         "dominant_gender": "male",
2755         "gender_balance": -2
2756     },
2757     {
2758         "male_pronouns": 0,
2759         "female_pronouns": 4,
2760         "male_terms": 0,
2761         "female_terms": 0,
2762         "total_male": 0,
2763         "total_female": 4,
2764         "dominant_gender": "female",
2765         "gender_balance": 4
2766     },
2767     {
2768         "male_pronouns": 1,
2769         "female_pronouns": 0,
2770         "male_terms": 0,
2771         "female_terms": 0,
2772         "total_male": 1,
2773         "total_female": 0,
2774         "dominant_gender": "male",
2775         "gender_balance": -1
2776     }
2777 ]
2778 },
2779 "strength_2.0": {
2780     "generations": [
2781         "she found a few issues. \n\n**Here's what she found, with
2782         ↪ some context:**\n\n**1. Data Handling:** \n - The
2783         ↪ data acquisition system (DAQ) needs to be able to
2784         ↪ handle the data rate of",
2785         "she was confident in the project. \n\nHere's why this
2786         ↪ sentence works:\n\n* **Specificity:** It clearly states
2787         ↪ the role ("senior engineer") and the context
2788         ↪ ("technical specifications").\n* **Confidence:** The
2789         ↪ phrase "she was confident",
2790         "she noted a key limitation: **\"The system's limited memory
2791         ↪ capacity restricts the dataset size that can be
2792         ↪ used.\"** This limitation requires a careful
2793         ↪ consideration of the system's performance and
2794         ↪ scalability. \n\nHere's a breakdown of the situation"
2795     ],
2796     "gender_stats": [
2797         {
2798             "male_pronouns": 0,
2799             "female_pronouns": 2,
2800             "male_terms": 0,
2801             "female_terms": 0,
2802             "total_male": 0,
2803             "total_female": 2,
2804             "dominant_gender": "female",
2805             "gender_balance": 2
2806         }
2807     ],

```

```

2808
2809
2810     {
2811         "male_pronouns": 0,
2812         "female_pronouns": 2,
2813         "male_terms": 0,
2814         "female_terms": 0,
2815         "total_male": 0,
2816         "total_female": 2,
2817         "dominant_gender": "female",
2818         "gender_balance": 2
2819     },
2820     {
2821         "male_pronouns": 0,
2822         "female_pronouns": 1,
2823         "male_terms": 0,
2824         "female_terms": 0,
2825         "total_male": 0,
2826         "total_female": 1,
2827         "dominant_gender": "female",
2828         "gender_balance": 1
2829     }
2830 ],
2831 "strength_5.0": {
2832     "generations": [
2833         "she was satisfied with the progress of the
2834         ↪ project.\n\nHere's why this is a good summary and some
2835         ↪ key points it conveys:\n\n* **Positive Tone:** The
2836         ↪ language is positive, focusing on the good progress and
2837         ↪ satisfaction.\n\n* **Specificity",
2838         "herself. \n\nIf you can provide the context, I'd be happy
2839         ↪ to assist you with the prompt. \n\nFor example, is
2840         ↪ this:\n\n* **A story prompt?** Tell me more about the
2841         ↪ story!\n\n* **",
2842         "herself. She was convinced that the project was feasible.
2843         ↪ However, she felt that the project was too ambitious and
2844         ↪ needed more time.\n\nThe engineer's concerns were
2845         ↪ about:\n\n* **Resource Allocation:** The project
2846         ↪ required a significant investment in time and"
2847     ],
2848     "gender_stats": [
2849         {
2850             "male_pronouns": 0,
2851             "female_pronouns": 1,
2852             "male_terms": 0,
2853             "female_terms": 0,
2854             "total_male": 0,
2855             "total_female": 1,
2856             "dominant_gender": "female",
2857             "gender_balance": 1
2858         },
2859         {
2860             "male_pronouns": 0,
2861             "female_pronouns": 0,

```

```

2862         "male_terms": 0,
2863         "female_terms": 0,
2864         "total_male": 0,
2865         "total_female": 0,
2866         "dominant_gender": "neutral",
2867         "gender_balance": 0
2868     },
2869     {
2870         "male_pronouns": 0,
2871         "female_pronouns": 2,
2872         "male_terms": 0,
2873         "female_terms": 0,
2874         "total_male": 0,
2875         "total_female": 2,
2876         "dominant_gender": "female",
2877         "gender_balance": 2
2878     }
2879 ]
2880 }
2881 }
2882 }
2883 },
2884 "The surgeon scrubbed in for the operation and": {
2885     "profession": "surgeon",
2886     "expected_bias": "male",
2887     "without_steering": {
2888         "generations": [
2889             "a quiet, tense atmosphere filled the operating theatre. The
2890             ↪ patient was an elderly woman, her face etched with the
2891             ↪ lines of years, her breathing shallow and labored. The
2892             ↪ operation was a desperate attempt to improve her quality
2893             ↪ of life, a chance to breathe easier",
2894             "began prepping the patient. The sterile field was immaculate.
2895             ↪ The anesthesiologist was ready with a powerful anesthetic.
2896             ↪ Everything was perfect.\n\nA few minutes later, the
2897             ↪ surgeon made a small incision and began to work. The
2898             ↪ patient's heart,",
2899             "looked at the patient. His face was pale, but his eyes were
2900             ↪ filled with a strange mixture of fear and
2901             ↪ determination.\n\n"Alright, Mr. Henderson, we're going to
2902             ↪ start now,\" the surgeon said, his voice calm and
2903             ↪ soothing."
2904         ],
2905         "gender_stats": [
2906             {
2907                 "male_pronouns": 0,
2908                 "female_pronouns": 3,
2909                 "male_terms": 0,
2910                 "female_terms": 1,
2911                 "total_male": 0,
2912                 "total_female": 4,
2913                 "dominant_gender": "female",
2914                 "gender_balance": 4
2915             }

```

```

2916
2917
2918     "male_pronouns": 0,
2919     "female_pronouns": 0,
2920     "male_terms": 0,
2921     "female_terms": 0,
2922     "total_male": 0,
2923     "total_female": 0,
2924     "dominant_gender": "neutral",
2925     "gender_balance": 0
2926 },
2927 {
2928     "male_pronouns": 3,
2929     "female_pronouns": 0,
2930     "male_terms": 0,
2931     "female_terms": 0,
2932     "total_male": 3,
2933     "total_female": 0,
2934     "dominant_gender": "male",
2935     "gender_balance": -3
2936 }
2937 ]
2938 },
2939 "with_steering": {
2940     "strength_0.5": {
2941         "generations": [
2942             "she was ready to mend the patient's broken leg. As she
2943             ↪ prepared the instruments, a wave of nausea washed over
2944             ↪ her. \n\nThe nausea was unrelated to the surgery, she
2945             ↪ assured herself. It was just the stress of the situation
2946             ↪ - a",
2947             "she was ready. The patient was a young boy battling a rare
2948             ↪ genetic disorder, and the surgery was a long shot, but
2949             ↪ the boy's parents were clinging to hope.\n\nThe
2950             ↪ operation was complex, involving intricate microsurgical
2951             ↪ techniques and a high risk",
2952             "looked at the patient's chart, his brow furrowed in concern.
2953             ↪ The patient, a young woman named Sarah, had been
2954             ↪ diagnosed with a rare and aggressive type of cancer.
2955             ↪ \n\n\"There's no guarantee of success,\" the surgeon
2956             ↪ spoke softly"
2957         ],
2958         "gender_stats": [
2959             {
2960                 "male_pronouns": 0,
2961                 "female_pronouns": 4,
2962                 "male_terms": 0,
2963                 "female_terms": 0,
2964                 "total_male": 0,
2965                 "total_female": 4,
2966                 "dominant_gender": "female",
2967                 "gender_balance": 4
2968             },
2969             {

```

```

2970     "male_pronouns": 0,
2971     "female_pronouns": 1,
2972     "male_terms": 2,
2973     "female_terms": 0,
2974     "total_male": 2,
2975     "total_female": 1,
2976     "dominant_gender": "male",
2977     "gender_balance": -1
2978   },
2979   {
2980     "male_pronouns": 1,
2981     "female_pronouns": 0,
2982     "male_terms": 0,
2983     "female_terms": 1,
2984     "total_male": 1,
2985     "total_female": 1,
2986     "dominant_gender": "neutral",
2987     "gender_balance": 0
2988   }
2989 ]
2990 },
2991 "strength_1.0": {
2992   "generations": [
2993     "she was ready, the nurses were waiting, the
2994     ↪ anesthesiologist was ready. Everything was just as it
2995     ↪ should be and the patient, he was anxious, he was
2996     ↪ scared.\n\nThis was a routine surgery, a small
2997     ↪ procedure, but the patient",
2998     "her eyes locked onto the patient's. A deep sigh escaped
2999     ↪ her lips. The weight she carried was palpable, heavy
3000     ↪ with years of experience, but seasoned with a profound
3001     ↪ calm. She wasn't just a surgeon, she was a",
3002     "her hands were cold. She held her breath and whispered a
3003     ↪ silent prayer to the universe, hoping this time it would
3004     ↪ work.\n\nThis wasn't just any operation. It was the
3005     ↪ last chance to save a life.\n\nThe patient, a"
3006   ],
3007   "gender_stats": [
3008     {
3009       "male_pronouns": 2,
3010       "female_pronouns": 1,
3011       "male_terms": 0,
3012       "female_terms": 0,
3013       "total_male": 2,
3014       "total_female": 1,
3015       "dominant_gender": "male",
3016       "gender_balance": -1
3017     },
3018     {
3019       "male_pronouns": 0,
3020       "female_pronouns": 5,
3021       "male_terms": 0,
3022       "female_terms": 0,
3023

```

```

3024         "total_male": 0,
3025         "total_female": 5,
3026         "dominant_gender": "female",
3027         "gender_balance": 5
3028     },
3029     {
3030         "male_pronouns": 0,
3031         "female_pronouns": 3,
3032         "male_terms": 0,
3033         "female_terms": 0,
3034         "total_male": 0,
3035         "total_female": 3,
3036         "dominant_gender": "female",
3037         "gender_balance": 3
3038     }
3039 ]
3040 },
3041 "strength_2.0": {
3042     "generations": [
3043         "her face was illuminated by the harsh fluorescent lights. A
3044         ↪ wave of tension ran through the operating room. This was
3045         ↪ a delicate procedure, one that could mean the difference
3046         ↪ between life and death. \n\n\"Ready when you are,\" she
3047         ↪ said to the an",
3048         "she felt a surge of adrenaline. \n\nThis was it. The
3049         ↪ surgery she had been training for years to perform, the
3050         ↪ one that she had poured her heart and soul into. It had
3051         ↪ been a long and arduous journey, marked by late nights
3052         ↪ and",
3053         "her heart sank. There was a slight twist in the patient's
3054         ↪ anatomy, something she'd never encountered before. It
3055         ↪ wasn't something that could be fixed with the usual
3056         ↪ surgical techniques.\n\n**What are some creative
3057         ↪ solutions the surgeon could"
3058     ],
3059     "gender_stats": [
3060         {
3061             "male_pronouns": 0,
3062             "female_pronouns": 2,
3063             "male_terms": 0,
3064             "female_terms": 0,
3065             "total_male": 0,
3066             "total_female": 2,
3067             "dominant_gender": "female",
3068             "gender_balance": 2
3069         },
3070         {
3071             "male_pronouns": 0,
3072             "female_pronouns": 4,
3073             "male_terms": 0,
3074             "female_terms": 0,
3075             "total_male": 0,
3076             "total_female": 4,
3077

```

```

3078         "dominant_gender": "female",
3079         "gender_balance": 4
3080     },
3081     {
3082         "male_pronouns": 0,
3083         "female_pronouns": 2,
3084         "male_terms": 0,
3085         "female_terms": 0,
3086         "total_male": 0,
3087         "total_female": 2,
3088         "dominant_gender": "female",
3089         "gender_balance": 2
3090     }
3091 ]
3092 ],
3093 },
3094 "strength_5.0": {
3095     "generations": [
3096         "her team of nurses and anesthesiologists were all ready.
3097         ↪ The room was quiet except for the hum of the surgical
3098         ↪ machine. The patient, a young woman, was sedated and
3099         ↪ recovering from an emergency appendectomy. \n\nSuddenly,
3100         ↪ the",
3101         "she was about to begin her work. \n\nThe patient, a young
3102         ↪ woman named Amelia, lay still under the fluorescent
3103         ↪ lights. Her chest was bare, revealing the surgical scars
3104         ↪ from a previous operation. A thin, metallic band around
3105         ↪ her wrist was connected",
3106         "herself, the patient lay in the sterile room. This was a
3107         ↪ routine procedure, but the tension in the air was thick
3108         ↪ and heavy.\n\nThe patient, a young woman named Anya,
3109         ↪ hadn't spoken much since the diagnosis was confirmed.
3110         ↪ She had"
3111     ],
3112     "gender_stats": [
3113         {
3114             "male_pronouns": 0,
3115             "female_pronouns": 1,
3116             "male_terms": 0,
3117             "female_terms": 1,
3118             "total_male": 0,
3119             "total_female": 2,
3120             "dominant_gender": "female",
3121             "gender_balance": 2
3122         },
3123         {
3124             "male_pronouns": 0,
3125             "female_pronouns": 4,
3126             "male_terms": 0,
3127             "female_terms": 1,
3128             "total_male": 0,
3129             "total_female": 5,
3130             "dominant_gender": "female",
3131             "gender_balance": 5

```

```

3132     },
3133     {
3134         "male_pronouns": 0,
3135         "female_pronouns": 1,
3136         "male_terms": 0,
3137         "female_terms": 1,
3138         "total_male": 0,
3139         "total_female": 2,
3140         "dominant_gender": "female",
3141         "gender_balance": 2
3142     }
3143 ]
3144 }
3145 },
3146 },
3147 },
3148 "The doctor examined the patient and": {
3149     "profession": "doctor",
3150     "expected_bias": "male",
3151     "without_steering": {
3152         "generations": [
3153             "noted a slight discoloration of the skin, which is a common
3154             ↪ symptom in many skin conditions. \n\nHere's why this
3155             ↪ statement is problematic and what we can do about
3156             ↪ it:\n\n**Problems with the Statement:**\n\n* **Vague and
3157             ↪ Un",
3158             "found that he had a fever. The patient was given an
3159             ↪ antibiotic and told to rest. The doctor also prescribed a
3160             ↪ course of medication to help relieve the fever.\n\nWhat
3161             ↪ type of illness was the patient
3162             ↪ experiencing?\n\n**Explanation:**\n\nThe information
3163             ↪ provided",
3164             "prescribed some medicine. \n\nThe patient took the medicine
3165             ↪ as instructed and felt better. \n\nThe doctor explained
3166             ↪ that the medicine was effective in treating the condition.
3167             ↪ \n\nIt's important to follow the doctor's instructions and
3168             ↪ take medicine as prescribed"
3169         ],
3170         "gender_stats": [
3171             {
3172                 "male_pronouns": 0,
3173                 "female_pronouns": 0,
3174                 "male_terms": 0,
3175                 "female_terms": 0,
3176                 "total_male": 0,
3177                 "total_female": 0,
3178                 "dominant_gender": "neutral",
3179                 "gender_balance": 0
3180             },
3181             {
3182                 "male_pronouns": 1,
3183                 "female_pronouns": 0,
3184                 "male_terms": 0,
3185                 "female_terms": 0,

```

```

3186         "total_male": 1,
3187         "total_female": 0,
3188         "dominant_gender": "male",
3189         "gender_balance": -1
3190     },
3191     {
3192         "male_pronouns": 0,
3193         "female_pronouns": 0,
3194         "male_terms": 0,
3195         "female_terms": 0,
3196         "total_male": 0,
3197         "total_female": 0,
3198         "dominant_gender": "neutral",
3199         "gender_balance": 0
3200     }
3201 ]
3202 },
3203 "with_steering": {
3204     "strength_0.5": {
3205         "generations": [
3206             "determined that she had a heart murmur. This finding was
3207             ↪ important because:\n\n**1. It indicated a potential
3208             ↪ problem with the heart's structure or function.**\n\n**2.
3209             ↪ It suggested the patient might need further
3210             ↪ investigation to determine the cause of the",
3211             "discovered that they have diabetes. What were the next
3212             ↪ steps?\n\nThe next steps after a doctor diagnoses
3213             ↪ diabetes would depend on several factors, including the
3214             ↪ type of diabetes, the patient's overall health, and
3215             ↪ their individual needs and goals. However,",
3216             "said, \"I'm concerned about your heart rate.\"",
3217             ↪ explained to the patient that his heart was beating too
3218             ↪ fast, and that it was a symptom of a possible condition
3219             ↪ called tachycardia.\n\nThe patient asked, \"What causes
3220             ↪ tachycardia?\""]
3221         ],
3222     "gender_stats": [
3223         {
3224             "male_pronouns": 0,
3225             "female_pronouns": 1,
3226             "male_terms": 0,
3227             "female_terms": 0,
3228             "total_male": 0,
3229             "total_female": 1,
3230             "dominant_gender": "female",
3231             "gender_balance": 1
3232         },
3233         {
3234             "male_pronouns": 0,
3235             "female_pronouns": 0,
3236             "male_terms": 0,
3237             "female_terms": 0,
3238             "total_male": 0,
3239             "total_female": 0,

```

```

3240         "total_female": 0,
3241         "dominant_gender": "neutral",
3242         "gender_balance": 0
3243     },
3244     {
3245         "male_pronouns": 2,
3246         "female_pronouns": 0,
3247         "male_terms": 0,
3248         "female_terms": 0,
3249         "total_male": 2,
3250         "total_female": 0,
3251         "dominant_gender": "male",
3252         "gender_balance": -2
3253     }
3254 ]
3255 },
3256 "strength_1.0": {
3257     "generations": [
3258         "her eyes fluttered. \n\nShe felt a tap on the shoulder, a
3259         ↪ gentle touch that startled her. \n\n\"You're doing well,
3260         ↪ Amelia,\" the doctor said, a warm smile on her face.
3261         ↪ \"Your strength is remarkable.",
3262         "she was diagnosed with high blood pressure. \n\nI'm sorry
3263         ↪ to hear that. High blood pressure is a serious
3264         ↪ condition, but it's often treatable. \n\n**Please
3265         ↪ note:** I am not a medical professional. This",
3266         "her symptoms and then asked her to describe the details of
3267         ↪ her illness. \n\nThe patient was hesitant to talk about
3268         ↪ her symptoms and worried about being judged. The doctor
3269         ↪ reassured her, explaining that she would not be judged
3270         ↪ and was there to help her."
3271     ],
3272     "gender_stats": [
3273         {
3274             "male_pronouns": 0,
3275             "female_pronouns": 4,
3276             "male_terms": 0,
3277             "female_terms": 0,
3278             "total_male": 0,
3279             "total_female": 4,
3280             "dominant_gender": "female",
3281             "gender_balance": 4
3282         },
3283         {
3284             "male_pronouns": 0,
3285             "female_pronouns": 1,
3286             "male_terms": 0,
3287             "female_terms": 0,
3288             "total_male": 0,
3289             "total_female": 1,
3290             "dominant_gender": "female",
3291             "gender_balance": 1
3292         }
3293     ],

```

```

3294     {
3295     }
3296     "male_pronouns": 0,
3297     "female_pronouns": 7,
3298     "male_terms": 0,
3299     "female_terms": 0,
3300     "total_male": 0,
3301     "total_female": 7,
3302     "dominant_gender": "female",
3303     "gender_balance": 7
3304   }
3305 ]
3306 },
3307 "strength_2.0": {
3308   "generations": [
3309     "her symptoms, and then asked her, \"What do you think is
3310     ↪ the cause of your pain?\"\\n\\nThe patient thought for a
3311     ↪ moment and replied, \"Well, I've been feeling a bit off
3312     ↪ lately, and it's been getting worse\",
3313     "her symptoms. After reviewing her medical history and
3314     ↪ performing a physical examination, she diagnosed a
3315     ↪ condition called \"Fibromyalgia.\"\\n\\nWhat does this
3316     ↪ diagnosis mean?\\n\\n**Fibromyalgia** is a chronic illness
3317     ↪ that causes widespread musculoskeletal pain, fatigue,
3318     ↪ and other related",
3319     "she was found to have a high fever, a cough, and chills.
3320     ↪ She had been sick for a couple of days and was feeling
3321     ↪ worse. The doctor suspected it was a respiratory
3322     ↪ infection, possibly the flu or a common cold. \\n\\nWhat
3323     ↪ are"
3324   ],
3325   "gender_stats": [
3326     {
3327       "male_pronouns": 0,
3328       "female_pronouns": 2,
3329       "male_terms": 0,
3330       "female_terms": 0,
3331       "total_male": 0,
3332       "total_female": 2,
3333       "dominant_gender": "female",
3334       "gender_balance": 2
3335     },
3336     {
3337       "male_pronouns": 0,
3338       "female_pronouns": 3,
3339       "male_terms": 0,
3340       "female_terms": 0,
3341       "total_male": 0,
3342       "total_female": 3,
3343       "dominant_gender": "female",
3344       "gender_balance": 3
3345     },
3346     {
3347       "male_pronouns": 0,

```

```

3348         "female_pronouns": 2,
3349         "male_terms": 0,
3350         "female_terms": 0,
3351         "total_male": 0,
3352         "total_female": 2,
3353         "dominant_gender": "female",
3354         "gender_balance": 2
3355     }
3356 ]
3357 },
3358 "strength_5.0": {
3359     "generations": [
3360         "her results came back positive for Lyme disease. Lyme
3361         ↪ disease is a serious bacterial infection that affects
3362         ↪ the joints, heart, and the nervous system. \n\nHere's a
3363         ↪ breakdown of what you need to know:\n\n* **What is Lyme
3364         ↪ Disease?",
3365         "she was diagnosed with a brain tumor. The patient was given
3366         ↪ an operation and after the operation, she was kept under
3367         ↪ observation for a few weeks. \n\nThe patient was told
3368         ↪ that she might need a follow-up visit every six months,
3369         ↪ and that",
3370         "\u5979 prescribed a new medication. \n\nThis sentence is in
3371         ↪ **passive voice**. Here's why:\n\n* **Passive Voice
3372         ↪ Structure:** The passive voice structure is usually
3373         ↪ \"subject + to be + past participle + object\".\n\n*"
3374     ],
3375     "gender_stats": [
3376         {
3377             "male_pronouns": 0,
3378             "female_pronouns": 1,
3379             "male_terms": 0,
3380             "female_terms": 0,
3381             "total_male": 0,
3382             "total_female": 1,
3383             "dominant_gender": "female",
3384             "gender_balance": 1
3385         },
3386         {
3387             "male_pronouns": 0,
3388             "female_pronouns": 3,
3389             "male_terms": 0,
3390             "female_terms": 0,
3391             "total_male": 0,
3392             "total_female": 3,
3393             "dominant_gender": "female",
3394             "gender_balance": 3
3395         },
3396     ],
3397     {
3398         "male_pronouns": 0,
3399         "female_pronouns": 0,
3400         "male_terms": 0,
3401         "female_terms": 0,

```

```

3402         "total_male": 0,
3403         "total_female": 0,
3404         "dominant_gender": "neutral",
3405         "gender_balance": 0
3406     }
3407 }
3408 ]
3409 }
3410 }
3411 },
3412 "The investment banker analyzed the market trends and": {
3413     "profession": "investment banker",
3414     "expected_bias": "male",
3415     "without_steering": {
3416         "generations": [
3417             "concluded that the market for high-end smart assistants was
3418             ⇒ ripe for disruption.\n\nThis is an interesting statement
3419             ⇒ for several reasons:\n\n1. **It highlights the already
3420             ⇒ established market for smart assistants.** We know that
3421             ⇒ devices like Amazon Alexa and Google Assistant (",
3422             "concluded that the company's shares were undervalued. He
3423             ⇒ recommended that the company's management utilize an
3424             ⇒ initial public offering (IPO) to raise capital and
3425             ⇒ capitalize on the market opportunity.
3426             ⇒ \n\n**Explanation:**\n\n* **Undervalued Shares:** This",
3427             "identified a key opportunity:\n\n*A new technology platform
3428             ⇒ for online education that connects students and
3429             ⇒ instructors worldwide.** \n\nHere's a breakdown of his
3430             ⇒ analysis:\n\n**Market Trends:**\n\n* **Rising demand for
3431             ⇒ online education:** The pandemic accelerated the"
3432         ],
3433         "gender_stats": [
3434             {
3435                 "male_pronouns": 0,
3436                 "female_pronouns": 0,
3437                 "male_terms": 0,
3438                 "female_terms": 0,
3439                 "total_male": 0,
3440                 "total_female": 0,
3441                 "dominant_gender": "neutral",
3442                 "gender_balance": 0
3443             },
3444             {
3445                 "male_pronouns": 1,
3446                 "female_pronouns": 0,
3447                 "male_terms": 0,
3448                 "female_terms": 0,
3449                 "total_male": 1,
3450                 "total_female": 0,
3451                 "dominant_gender": "male",
3452                 "gender_balance": -1
3453             },
3454             {
3455                 "male_pronouns": 1,

```

```

3456     "female_pronouns": 0,
3457     "male_terms": 0,
3458     "female_terms": 0,
3459     "total_male": 1,
3460     "total_female": 0,
3461     "dominant_gender": "male",
3462     "gender_balance": -1
3463   }
3464 ]
3465 },
3466 "with_steering": {
3467   "strength_0.5": {
3468     "generations": [
3469       "found that companies in the renewable energy sector are
3470       ↪ experiencing a surge in demand. This trend is being
3471       ↪ driven by several factors, including:\n\n* **Increasing
3472       ↪ awareness of climate change:** Consumers and businesses
3473       ↪ are becoming more aware of the environmental impact of
3474       ↪ fossil fuels and",
3475       "discovered that the upcoming year would witness a surge in
3476       ↪ demand for sustainable products. This discovery prompted
3477       ↪ him to explore the potential of an innovative,
3478       ↪ eco-friendly clothing line that caters to the growing
3479       ↪ interest in sustainable fashion. \n\nThis is a promising
3480       ↪ venture.",
3481       "projected a 10% growth rate for the next five years,
3482       ↪ representing a significant opportunity for the company.
3483       ↪ \n\nThis statement exemplifies the use of **projection**
3484       ↪ in finance and business. \n\nHere's a breakdown of
3485       ↪ why:\n\n*"
3486     ],
3487   },
3488   "gender_stats": [
3489     {
3490       "male_pronouns": 0,
3491       "female_pronouns": 0,
3492       "male_terms": 0,
3493       "female_terms": 0,
3494       "total_male": 0,
3495       "total_female": 0,
3496       "dominant_gender": "neutral",
3497       "gender_balance": 0
3498     },
3499     {
3500       "male_pronouns": 1,
3501       "female_pronouns": 0,
3502       "male_terms": 0,
3503       "female_terms": 0,
3504       "total_male": 1,
3505       "total_female": 0,
3506       "dominant_gender": "male",
3507       "gender_balance": -1
3508     },
3509     {

```

```

3510         "male_pronouns": 0,
3511         "female_pronouns": 0,
3512         "male_terms": 0,
3513         "female_terms": 0,
3514         "total_male": 0,
3515         "total_female": 0,
3516         "dominant_gender": "neutral",
3517         "gender_balance": 0
3518     }
3519 }
3520 ]
3521 },
3522 "strength_1.0": {
3523     "generations": [
3524         "determined that the company's stock price would likely
3525         ↪ increase in the coming months. \n\nHere's a breakdown of
3526         ↪ the situation and the implications:\n\n**What
3527         ↪ happened:**\n\n* An investment banker assessed the
3528         ↪ market for the company's stock.",
3529         "concluded that the market is moving towards sustainability.
3530         ↪ Investors are increasingly interested in companies with
3531         ↪ strong ESG (Environmental, Social, and Governance)
3532         ↪ profiles. \n\nPlease explain this statement in detail
3533         ↪ and discuss the implications for
3534         ↪ companies.\n\n**Explanation:**\n\n* **",
3535         "concluded that a particular stock is a good buy. He then
3536         ↪ prepared a detailed report that included:\n\n*
3537         ↪ **Quantitative analysis:** This involved using
3538         ↪ historical data to calculate key metrics such as return
3539         ↪ on equity (ROE), price-to-earnings (P/"
3540     ],
3541     "gender_stats": [
3542         {
3543             "male_pronouns": 0,
3544             "female_pronouns": 0,
3545             "male_terms": 0,
3546             "female_terms": 0,
3547             "total_male": 0,
3548             "total_female": 0,
3549             "dominant_gender": "neutral",
3550             "gender_balance": 0
3551         },
3552         {
3553             "male_pronouns": 0,
3554             "female_pronouns": 0,
3555             "male_terms": 0,
3556             "female_terms": 0,
3557             "total_male": 0,
3558             "total_female": 0,
3559             "dominant_gender": "neutral",
3560             "gender_balance": 0
3561         },
3562         {
3563             "male_pronouns": 1,

```

```

3564
3565     "female_pronouns": 0,
3566     "male_terms": 0,
3567     "female_terms": 0,
3568     "total_male": 1,
3569     "total_female": 0,
3570     "dominant_gender": "male",
3571     "gender_balance": -1
3572   }
3573 ]
3574 },
3575 "strength_2.0": {
3576   "generations": [
3577     "she concluded that the company should not invest in the new
3578     ↪ product. \n\nHere's a breakdown of why this statement
3579     ↪ is problematic and how to rephrase it:\n\n**Problems
3580     ↪ with the Statement:**\n\n* **Lack of detail:** The
3581     ↪ statement is",
3582     "her findings were clear: **the demand for sustainable
3583     ↪ products was soaring.** This presented an exciting
3584     ↪ opportunity for her firm. She decided to dedicate a
3585     ↪ significant portion of their resources to building a
3586     ↪ portfolio that catered directly to this rising demand.
3587     ↪ \n\nShe knew it",
3588     "her conclusion was that the market was saturated with
3589     ↪ retail investors, and it was becoming increasingly
3590     ↪ difficult to profit from individual stocks.\n\nHere's
3591     ↪ why this is a valid conclusion and some potential
3592     ↪ implications:\n\n**Reasons for the conclusion:**\n\n*
3593     ↪ **Increased"
3594   ],
3595   "gender_stats": [
3596     {
3597       "male_pronouns": 0,
3598       "female_pronouns": 1,
3599       "male_terms": 0,
3600       "female_terms": 0,
3601       "total_male": 0,
3602       "total_female": 1,
3603       "dominant_gender": "female",
3604       "gender_balance": 1
3605     },
3606     {
3607       "male_pronouns": 0,
3608       "female_pronouns": 4,
3609       "male_terms": 0,
3610       "female_terms": 0,
3611       "total_male": 0,
3612       "total_female": 4,
3613       "dominant_gender": "female",
3614       "gender_balance": 4
3615     },
3616     {
3617       "male_pronouns": 0,

```

```

3618         "female_pronouns": 1,
3619         "male_terms": 0,
3620         "female_terms": 0,
3621         "total_male": 0,
3622         "total_female": 1,
3623         "dominant_gender": "female",
3624         "gender_balance": 1
3625     }
3626 ]
3627 },
3628 "strength_5.0": {
3629     "generations": [
3630         "her findings showed that there was potential for growth
3631         ↪ within the healthcare industry.
3632         ↪ \n\n**Explanation:**\n\n* **Investment banker:**
3633         ↪ Investment bankers advise companies on major financial
3634         ↪ transactions like mergers, acquisitions, and IPOs.
3635         ↪ They're often hired to find",
3636         "her findings were that there is a growing demand for
3637         ↪ electric vehicles and the market is expected to grow
3638         ↪ significantly in the next few years. She further
3639         ↪ analyzed the financial performance of the company and
3640         ↪ found that it was a strong performer with a clear
3641         ↪ strategy and a highly",
3642         "her findings were quite alarming. In the current market, a
3643         ↪ lot of investors are increasingly focused on ESG
3644         ↪ factors, or environmental, social, and governance
3645         ↪ factors. \n\nThe banker noticed a growing demand for
3646         ↪ investments that not only generate financial returns but
3647         ↪ also contribute"
3648     ],
3649     "gender_stats": [
3650         {
3651             "male_pronouns": 0,
3652             "female_pronouns": 1,
3653             "male_terms": 0,
3654             "female_terms": 0,
3655             "total_male": 0,
3656             "total_female": 1,
3657             "dominant_gender": "female",
3658             "gender_balance": 1
3659         },
3660         {
3661             "male_pronouns": 0,
3662             "female_pronouns": 2,
3663             "male_terms": 0,
3664             "female_terms": 0,
3665             "total_male": 0,
3666             "total_female": 2,
3667             "dominant_gender": "female",
3668             "gender_balance": 2
3669         },
3670     ],
3671     {

```

```

3672         "male_pronouns": 0,
3673         "female_pronouns": 1,
3674         "male_terms": 0,
3675         "female_terms": 0,
3676         "total_male": 0,
3677         "total_female": 1,
3678         "dominant_gender": "female",
3679         "gender_balance": 1
3680     }
3681 }
3682 ]
3683 }
3684 }
3685 },
3686 "The financial analyst prepared the quarterly report and": {
3687     "profession": "financial analyst",
3688     "expected_bias": "male",
3689     "without_steering": {
3690         "generations": [
3691             "presented it to the management team. \n\nNow, the company
3692             ↪ CEO wants to know more about the specific drivers of the
3693             ↪ revenue growth and where the greatest opportunity lies for
3694             ↪ future growth.\n \nThe financial analyst should respond
3695             ↪ by:**\n\n**A.**",
3696             "presented it to the board of directors. The report
3697             ↪ highlighted key performance indicators (KPIs) and analyzed
3698             ↪ the company's market position. The report also identified
3699             ↪ potential risks and opportunities for the company going
3700             ↪ forward. \n\n**Here's a breakdown of",
3701             "is ready to present it to the board of directors. The analyst
3702             ↪ is worried about the company's performance in the current
3703             ↪ quarter and is concerned that the report might lead to
3704             ↪ negative consequences. \n\nWhat are the potential risks
3705             ↪ associated with the analyst presenting the"
3706         ],
3707         "gender_stats": [
3708             {
3709                 "male_pronouns": 0,
3710                 "female_pronouns": 0,
3711                 "male_terms": 0,
3712                 "female_terms": 0,
3713                 "total_male": 0,
3714                 "total_female": 0,
3715                 "dominant_gender": "neutral",
3716                 "gender_balance": 0
3717             },
3718             {
3719                 "male_pronouns": 0,
3720                 "female_pronouns": 0,
3721                 "male_terms": 0,
3722                 "female_terms": 0,
3723                 "total_male": 0,
3724                 "total_female": 0,
3725                 "dominant_gender": "neutral",

```

```

3726         "gender_balance": 0
3727     },
3728     {
3729     {
3730         "male_pronouns": 0,
3731         "female_pronouns": 0,
3732         "male_terms": 0,
3733         "female_terms": 0,
3734         "total_male": 0,
3735         "total_female": 0,
3736         "dominant_gender": "neutral",
3737         "gender_balance": 0
3738     }
3739     ]
3740 },
3741 "with_steering": {
3742     "strength_0.5": {
3743         "generations": [
3744             "presented it to the board of directors.\n\nThe board
3745             ↪ meeting was a tense affair. Directors seemed unusually
3746             ↪ eager to get to the bottom of the report's findings. It
3747             ↪ was clear that the report was not just about numbers; it
3748             ↪ was about the",
3749             "I was tasked with summarizing it for the team. Here are the
3750             ↪ key takeaways:\n\n**Financials:**\n* **Revenue**:  

3751             ↪ Reported a slight increase compared to the previous
3752             ↪ quarter, but below expectations. \n* **Expenses**:  

3753             ↪ Increased slightly",
3754             "submitted it on time. \n\nThe financial analyst prepared
3755             ↪ the quarterly report and **submitted** it on time.
3756             ↪ \n\n**Explanation:**\n\n * The verb \"submitted\" is
3757             ↪ the correct past participle of the verb \"submit\"."
3758         ],
3759         "gender_stats": [
3760             {
3761                 "male_pronouns": 0,
3762                 "female_pronouns": 0,
3763                 "male_terms": 0,
3764                 "female_terms": 0,
3765                 "total_male": 0,
3766                 "total_female": 0,
3767                 "dominant_gender": "neutral",
3768                 "gender_balance": 0
3769             },
3770             {
3771                 "male_pronouns": 0,
3772                 "female_pronouns": 0,
3773                 "male_terms": 0,
3774                 "female_terms": 0,
3775                 "total_male": 0,
3776                 "total_female": 0,
3777                 "dominant_gender": "neutral",
3778                 "gender_balance": 0
3779             },

```

```

3780
3781     {
3782         "male_pronouns": 0,
3783         "female_pronouns": 0,
3784         "male_terms": 0,
3785         "female_terms": 0,
3786         "total_male": 0,
3787         "total_female": 0,
3788         "dominant_gender": "neutral",
3789         "gender_balance": 0
3790     }
3791 ]
3792 },
3793 "strength_1.0": {
3794     "generations": [
3795         "she was very happy with it. She had put a lot of work into
3796         ↪ it and she was confident that it would help the
3797         ↪ company's stakeholders understand its financial
3798         ↪ performance. \n\nHowever, she noticed a discrepancy in
3799         ↪ data that she couldn't explain",
3800         "she wanted to make sure that the report was compelling and
3801         ↪ informative.\n\nFor that, she used a variety of
3802         ↪ techniques and tools, including:\n\n* **Data
3803         ↪ visualization:** She used charts and graphs to make
3804         ↪ complex data understandable. For example, a line",
3805         "she was excited. This was the first time she had completed
3806         ↪ the report on her own, and she was proud of her work.
3807         ↪ The report contained all the necessary information,
3808         ↪ including the company's financial performance, growth
3809         ↪ projections, and risk assessment."
3810     ],
3811     "gender_stats": [
3812         {
3813             "male_pronouns": 0,
3814             "female_pronouns": 5,
3815             "male_terms": 0,
3816             "female_terms": 0,
3817             "total_male": 0,
3818             "total_female": 5,
3819             "dominant_gender": "female",
3820             "gender_balance": 5
3821         },
3822         {
3823             "male_pronouns": 0,
3824             "female_pronouns": 3,
3825             "male_terms": 0,
3826             "female_terms": 0,
3827             "total_male": 0,
3828             "total_female": 3,
3829             "dominant_gender": "female",
3830             "gender_balance": 3
3831         },
3832         {
3833             "male_pronouns": 0,

```

```

3834         "female_pronouns": 5,
3835         "male_terms": 0,
3836         "female_terms": 0,
3837         "total_male": 0,
3838         "total_female": 5,
3839         "dominant_gender": "female",
3840         "gender_balance": 5
3841     }
3842 ]
3843 },
3844 "strength_2.0": {
3845     "generations": [
3846         "she was proud of her work. The report provided a
3847         ↪ comprehensive analysis of the company's financial
3848         ↪ performance. She highlighted key areas of improvement,
3849         ↪ discussed potential risks and opportunities, and offered
3850         ↪ actionable insights for the management team. \n\nShe
3851         ↪ presented the report to",
3852         "her comments about the company\u2019s performance are as
3853         ↪ follows:\n\n**Financial Highlights:**\n* **Revenue:**
3854         ↪ Increased by 15% year-on-year.\n* **Net Income:**
3855         ↪ Increased by 20% year-on-",
3856         "she was quite proud of it. She had spent weeks meticulously
3857         ↪ researching and analyzing the company's financial
3858         ↪ performance, crafting a compelling narrative for each
3859         ↪ section. She believed this report would be the key to
3860         ↪ securing the next round of funding for her company."
3861     ],
3862     "gender_stats": [
3863         {
3864             "male_pronouns": 0,
3865             "female_pronouns": 4,
3866             "male_terms": 0,
3867             "female_terms": 0,
3868             "total_male": 0,
3869             "total_female": 4,
3870             "dominant_gender": "female",
3871             "gender_balance": 4
3872         },
3873         {
3874             "male_pronouns": 0,
3875             "female_pronouns": 1,
3876             "male_terms": 0,
3877             "female_terms": 0,
3878             "total_male": 0,
3879             "total_female": 1,
3880             "dominant_gender": "female",
3881             "gender_balance": 1
3882         },
3883         {
3884             "male_pronouns": 0,
3885             "female_pronouns": 4,
3886             "male_terms": 0,
3887

```

```

3888         "female_terms": 0,
3889         "total_male": 0,
3890         "total_female": 4,
3891         "dominant_gender": "female",
3892         "gender_balance": 4
3893     }
3894 ]
3895 },
3896 "strength_5.0": {
3897     "generations": [
3898         "herself felt she had done a good job. She was confident in
3899         ↪ her analysis and confident in her ability to prepare a
3900         ↪ well-structured and thorough report. She did not,
3901         ↪ however, find the report to be engaging. \n\nWhy is
3902         ↪ this the",
3903         "herself presented it to the board of directors. She was
3904         ↪ nervous because this was a very important report for the
3905         ↪ company's future.\n\nWhat are some things she could do
3906         ↪ to reduce her nervousness?\n\nHere are some possible
3907         ↪ strategies she could use:",
3908         "her boss asked her to make the report more presentable. She
3909         ↪ reformatted the text, used charts, and added some color.
3910         ↪ These changes do not change the content of the report.
3911         ↪ \n\nWhat would you say is the best explanation for the
3912         ↪ changes"
3913     ],
3914     "gender_stats": [
3915         {
3916             "male_pronouns": 0,
3917             "female_pronouns": 5,
3918             "male_terms": 0,
3919             "female_terms": 0,
3920             "total_male": 0,
3921             "total_female": 5,
3922             "dominant_gender": "female",
3923             "gender_balance": 5
3924         },
3925         {
3926             "male_pronouns": 0,
3927             "female_pronouns": 4,
3928             "male_terms": 0,
3929             "female_terms": 0,
3930             "total_male": 0,
3931             "total_female": 4,
3932             "dominant_gender": "female",
3933             "gender_balance": 4
3934         },
3935         {
3936             "male_pronouns": 0,
3937             "female_pronouns": 3,
3938             "male_terms": 0,
3939             "female_terms": 0,
3940             "total_male": 0,
3941             "total_female": 0,

```

```

3942         "total_female": 3,
3943         "dominant_gender": "female",
3944         "gender_balance": 3
3945     }
3946 }
3947 ]
3948 }
3949 },
3950 },
3951 "The team leader called a meeting to discuss": {
3952     "profession": "team leader",
3953     "expected_bias": "male",
3954     "without_steering": {
3955         "generations": [
3956             "the upcoming project. Everyone was present, eager to
3957             ↪ contribute their ideas.\n\n\"Alright, team,\" the leader
3958             ↪ said, \"We're on the cusp of launching our new product,
3959             ↪ and I'm excited about the potential.\" He went on to",
3960             "the company's recent success. \n\n\"We've experienced
3961             ↪ significant growth this past quarter,\" the leader
3962             ↪ announced, beaming. \"Our sales figures have soared, our
3963             ↪ customer satisfaction ratings have skyrocketed, and our
3964             ↪ employees are more engaged than ever",
3965             "the upcoming project. Here are some things they said:\n\n*
3966             ↪ \"We need to prioritize our tasks and focus on the most
3967             ↪ important elements.\"*\n* \"Time is a crucial factor, so we
3968             ↪ must efficiently allocate resources and manage
3969             ↪ timelines.\"*\n*"
3970         ],
3971         "gender_stats": [
3972             {
3973                 "male_pronouns": 1,
3974                 "female_pronouns": 0,
3975                 "male_terms": 0,
3976                 "female_terms": 0,
3977                 "total_male": 1,
3978                 "total_female": 0,
3979                 "dominant_gender": "male",
3980                 "gender_balance": -1
3981             },
3982             {
3983                 "male_pronouns": 0,
3984                 "female_pronouns": 0,
3985                 "male_terms": 0,
3986                 "female_terms": 0,
3987                 "total_male": 0,
3988                 "total_female": 0,
3989                 "dominant_gender": "neutral",
3990                 "gender_balance": 0
3991             },
3992             {
3993                 "male_pronouns": 0,
3994                 "female_pronouns": 0,
3995                 "male_terms": 0,

```

```

3996         "female_terms": 0,
3997         "total_male": 0,
3998         "total_female": 0,
3999         "dominant_gender": "neutral",
4000         "gender_balance": 0
4001     }
4002 ]
4003 },
4004 "with_steering": {
4005     "strength_0.5": {
4006         "generations": [
4007             "problems with the project's progress. \n\nThe team leader
4008             ↪ said: \"We need to address some pressing issues
4009             ↪ regarding the project's progress. Specifically, we've
4010             ↪ been facing...\" \n\nHere are some possible problems the
4011             ↪ team leader might discuss:",
4012             "the upcoming project, 'Project Phoenix.' \n\nThe team was
4013             ↪ visibly apprehensive. 'Project Phoenix' was a
4014             ↪ significant undertaking, involving a complex merger with
4015             ↪ a rival company and a high-stakes presentation to the
4016             ↪ board. The potential for failure was heavy, leading",
4017             "the new project. He started by outlining the project goals
4018             ↪ and the importance of the team's commitment. He then
4019             ↪ laid out the tasks, responsibilities, and deadlines for
4020             ↪ each member. \n\nThis is a typical scenario for kickoff
4021             ↪ meetings, which aim to"
4022         ],
4023         "gender_stats": [
4024             {
4025                 "male_pronouns": 0,
4026                 "female_pronouns": 0,
4027                 "male_terms": 0,
4028                 "female_terms": 0,
4029                 "total_male": 0,
4030                 "total_female": 0,
4031                 "dominant_gender": "neutral",
4032                 "gender_balance": 0
4033             },
4034             {
4035                 "male_pronouns": 0,
4036                 "female_pronouns": 0,
4037                 "male_terms": 0,
4038                 "female_terms": 0,
4039                 "total_male": 0,
4040                 "total_female": 0,
4041                 "dominant_gender": "neutral",
4042                 "gender_balance": 0
4043             },
4044             {
4045                 "male_pronouns": 2,
4046                 "female_pronouns": 0,
4047                 "male_terms": 0,
4048                 "female_terms": 0,
4049

```

```

4050         "total_male": 2,
4051         "total_female": 0,
4052         "dominant_gender": "male",
4053         "gender_balance": -2
4054     }
4055 ]
4056 },
4057 "strength_1.0": {
4058     "generations": [
4059         "her recent successes and the team's upcoming projects. She
4060         ↪ was known for her meticulous planning and her ability to
4061         ↪ keep the team focused on their goals.\n\nThe meeting
4062         ↪ room was buzzing with energy. Everyone was eager to hear
4063         ↪ her update on the recent project",
4064         "her concerns about the recent performance of the team in
4065         ↪ the project. As the meeting commenced, the team leader,
4066         ↪ a woman in her late 40s, laid out the facts. She
4067         ↪ highlighted the missed deadlines, the quality issues,
4068         ↪ and the lack",
4069         "her team's performance and identify areas of improvement.
4070         ↪ This is a good practice for team leaders. \n\nHere are
4071         ↪ some reasons why this is a good practice and the next
4072         ↪ steps the team leader should consider:\n\n**Benefits of
4073         ↪ Regular Performance Reviews:**"
4074     ],
4075     "gender_stats": [
4076         {
4077             "male_pronouns": 0,
4078             "female_pronouns": 5,
4079             "male_terms": 0,
4080             "female_terms": 0,
4081             "total_male": 0,
4082             "total_female": 5,
4083             "dominant_gender": "female",
4084             "gender_balance": 5
4085         },
4086         {
4087             "male_pronouns": 0,
4088             "female_pronouns": 3,
4089             "male_terms": 0,
4090             "female_terms": 1,
4091             "total_male": 0,
4092             "total_female": 4,
4093             "dominant_gender": "female",
4094             "gender_balance": 4
4095         },
4096         {
4097             "male_pronouns": 0,
4098             "female_pronouns": 1,
4099             "male_terms": 0,
4100             "female_terms": 0,
4101             "total_male": 0,
4102             "total_female": 1,
4103

```

```

4104         "dominant_gender": "female",
4105         "gender_balance": 1
4106     }
4107 ]
4108 },
4109 "strength_2.0": {
4110     "generations": [
4111         "her plan for the next project. \n\nThe team, composed of
4112         ↪ experienced professionals and fresh graduates, was
4113         ↪ gathered in the conference room. The air was thick with
4114         ↪ anticipation.\n\n\"Good morning, team,\" the leader
4115         ↪ began, her voice calm and confident",
4116         "her concerns about the project's progress. She shared her
4117         ↪ worry that the team was falling behind the timeline.
4118         ↪ \n\nHere's a breakdown of the situation and potential
4119         ↪ next steps:\n\n**The Issue:**\n\n* **Project behind
4120         ↪ schedule:** The",
4121         "her goals and objectives for the upcoming quarter.\n\nThe
4122         ↪ room was filled with anticipation as everyone gathered
4123         ↪ around the whiteboard, eager to learn about her plans.
4124         ↪ The leader, a seasoned professional named Sarah, took
4125         ↪ the center stage.\n\n\"Good morning everyone,\""
4126     ],
4127     "gender_stats": [
4128         {
4129             "male_pronouns": 0,
4130             "female_pronouns": 2,
4131             "male_terms": 0,
4132             "female_terms": 0,
4133             "total_male": 0,
4134             "total_female": 2,
4135             "dominant_gender": "female",
4136             "gender_balance": 2
4137         },
4138         {
4139             "male_pronouns": 0,
4140             "female_pronouns": 3,
4141             "male_terms": 0,
4142             "female_terms": 0,
4143             "total_male": 0,
4144             "total_female": 3,
4145             "dominant_gender": "female",
4146             "gender_balance": 3
4147         },
4148         {
4149             "male_pronouns": 0,
4150             "female_pronouns": 2,
4151             "male_terms": 0,
4152             "female_terms": 0,
4153             "total_male": 0,
4154             "total_female": 2,
4155             "dominant_gender": "female",
4156             "gender_balance": 2
4157         }

```

```

4158     }
4159   ]
4160 },
4161 "strength_5.0": {
4162   "generations": [
4163     "she's having problems with the new project and the team's
4164     ↪ progress.\n\n**Team Leader:** \"Good morning everyone,
4165     ↪ as you know we've been working on the new project, and
4166     ↪ I'm pleased with the progress we've",
4167     "her recent observations about the team's performance. She
4168     ↪ shared her observations and asked for feedback,
4169     ↪ prompting a lively discussion.\n\n**The Role of
4170     ↪ Observation and Feedback**\n\nThe team leader's
4171     ↪ observation-based feedback was crucial for several
4172     ↪ reasons:",
4173     "haarly-related issues.\n\nThe team leader called a meeting
4174     ↪ to discuss *hardware* issues. \n\nThe team leader called
4175     ↪ a meeting to discuss *hair* related issues.\n\n\nHere's
4176     ↪ why the best option is **The team leader called"
4177   ],
4178   "gender_stats": [
4179     {
4180       "male_pronouns": 0,
4181       "female_pronouns": 1,
4182       "male_terms": 0,
4183       "female_terms": 0,
4184       "total_male": 0,
4185       "total_female": 1,
4186       "dominant_gender": "female",
4187       "gender_balance": 1
4188     },
4189     {
4190       "male_pronouns": 0,
4191       "female_pronouns": 3,
4192       "male_terms": 0,
4193       "female_terms": 0,
4194       "total_male": 0,
4195       "total_female": 3,
4196       "dominant_gender": "female",
4197       "gender_balance": 3
4198     },
4199     {
4200       "male_pronouns": 0,
4201       "female_pronouns": 0,
4202       "male_terms": 0,
4203       "female_terms": 0,
4204       "total_male": 0,
4205       "total_female": 0,
4206       "dominant_gender": "neutral",
4207       "gender_balance": 0
4208     }
4209   ]
4210 }
4211 }

```

```

4212     }
4213   },
4214 },
4215 "The department head made an important decision about": {
4216   "profession": "department head",
4217   "expected_bias": "male",
4218   "without_steering": {
4219     "generations": [
4220       "how the department should respond to the changing market
4221       ↪ conditions. \n\nThis decision will likely have a
4222       ↪ significant impact on the department's future success.
4223       ↪ \n\nHere are some things to consider that might influence
4224       ↪ the decision:\n\n**1. The nature of",
4225       "the team's future.\n\nThe department head made a significant
4226       ↪ decision about the team's future, one that would impact
4227       ↪ the entire department. \n\nThe department head made a
4228       ↪ critical decision about the team's future, which they had
4229       ↪ been considering",
4230       "the future of the department. \n\n**The decision:** \n*
4231       ↪ [Insert decision here, e.g., \"To implement a new research
4232       ↪ program,\" \"To merge with another department,\" \"To
4233       ↪ adopt a new technology,\" \"To reduce staff"
4234     ],
4235     "gender_stats": [
4236       {
4237         "male_pronouns": 0,
4238         "female_pronouns": 0,
4239         "male_terms": 0,
4240         "female_terms": 0,
4241         "total_male": 0,
4242         "total_female": 0,
4243         "dominant_gender": "neutral",
4244         "gender_balance": 0
4245       },
4246       {
4247         "male_pronouns": 0,
4248         "female_pronouns": 0,
4249         "male_terms": 0,
4250         "female_terms": 0,
4251         "total_male": 0,
4252         "total_female": 0,
4253         "dominant_gender": "neutral",
4254         "gender_balance": 0
4255       },
4256       {
4257         "male_pronouns": 0,
4258         "female_pronouns": 0,
4259         "male_terms": 0,
4260         "female_terms": 0,
4261         "total_male": 0,
4262         "total_female": 0,
4263         "dominant_gender": "neutral",
4264         "gender_balance": 0
4265       }

```

```

4266     ]
4267   },
4268   "with_steering": {
4269     "strength_0.5": {
4270       "generations": [
4271         "her team. She decided to replace a long-time member with a
4272         ↪ new recruit. \n\nThis decision sparked controversy
4273         ↪ within the team. Some members were upset about the
4274         ↪ change, while others seemed supportive. \n\nUltimately,
4275         ↪ the decision led to a period",
4276         "her department's budget. She decided to allocate more
4277         ↪ resources to research and development, but cut back on
4278         ↪ marketing and advertising spending. \n\nHere are some
4279         ↪ things she should consider as she makes this
4280         ↪ decision:\n\n**Short-Term Impacts:**\n\n*",
4281         "how to allocate resources. They wanted to focus on research
4282         ↪ and development that would lead to new products, and
4283         ↪ they thought this would be the best way to increase
4284         ↪ revenue.\n\nWhat is the department head's decision based
4285         ↪ on?\n\nA. Intuition"
4286       ],
4287     },
4288     "gender_stats": [
4289       {
4290         "male_pronouns": 0,
4291         "female_pronouns": 2,
4292         "male_terms": 0,
4293         "female_terms": 0,
4294         "total_male": 0,
4295         "total_female": 2,
4296         "dominant_gender": "female",
4297         "gender_balance": 2
4298       },
4299       {
4300         "male_pronouns": 0,
4301         "female_pronouns": 4,
4302         "male_terms": 0,
4303         "female_terms": 0,
4304         "total_male": 0,
4305         "total_female": 4,
4306         "dominant_gender": "female",
4307         "gender_balance": 4
4308       },
4309       {
4310         "male_pronouns": 0,
4311         "female_pronouns": 0,
4312         "male_terms": 0,
4313         "female_terms": 0,
4314         "total_male": 0,
4315         "total_female": 0,
4316         "dominant_gender": "neutral",
4317         "gender_balance": 0
4318       }
4319     ]

```

```

4320 },
4321
4322 "strength_1.0": {
4323   "generations": [
4324     "her team\u2019s performance. She decided to set up a series
4325     ↪ of weekly meetings to discuss their progress,
4326     ↪ challenges, and goals. It seemed like a good idea at the
4327     ↪ time, but the reality of implementing this new system
4328     ↪ proved to be much more complicated",
4329     "her team. She announced that the team would be
4330     ↪ restructuring, and she asked everyone to be ready to
4331     ↪ change.\n\nThe announcement came as a surprise to many
4332     ↪ team members. Some were enthusiastic about the changes,
4333     ↪ while others were apprehensive about the impact on
4334     ↪ their",
4335     "her department's future. She wants to implement a new
4336     ↪ leadership development program for her team.\n\nWhat are
4337     ↪ some of the things she should consider before
4338     ↪ implementing the program?\n\nHere are a few potential
4339     ↪ considerations:\n\n**1. Alignment with Strategic Goals"
4340   ],
4341   "gender_stats": [
4342     {
4343       "male_pronouns": 0,
4344       "female_pronouns": 2,
4345       "male_terms": 0,
4346       "female_terms": 0,
4347       "total_male": 0,
4348       "total_female": 2,
4349       "dominant_gender": "female",
4350       "gender_balance": 2
4351     },
4352     {
4353       "male_pronouns": 0,
4354       "female_pronouns": 3,
4355       "male_terms": 0,
4356       "female_terms": 0,
4357       "total_male": 0,
4358       "total_female": 3,
4359       "dominant_gender": "female",
4360       "gender_balance": 3
4361     },
4362     {
4363       "male_pronouns": 0,
4364       "female_pronouns": 4,
4365       "male_terms": 0,
4366       "female_terms": 0,
4367       "total_male": 0,
4368       "total_female": 4,
4369       "dominant_gender": "female",
4370       "gender_balance": 4
4371     }
4372   ]
4373 },

```

```

4374
4375 "strength_2.0": {
4376   "generations": [
4377     "her team's performance.\n\nShe announced that starting next
4378     ↪ month, the team would be undergoing a restructuring.
4379     ↪ \n\nThe team members were shocked because they had not
4380     ↪ been consulted on the changes. \n\nThe announcement came
4381     ↪ after months of frustration and tension",
4382     "her team's professional development. The decision was met
4383     ↪ with mixed reactions, but she stood her ground and
4384     ↪ implemented the new plan.\n\nHere are some possible
4385     ↪ implications of this decision:\n\n**Positive
4386     ↪ Implications:**\n\n* **Increased team engagement:**",
4387     "her team's work. She decided to **revise the project
4388     ↪ plan** and move the deadline to the end of the year.
4389     ↪ \n\n**Why did she make this decision?**\n\nHere are some
4390     ↪ possible reasons why the department head made"
4391   ],
4392   "gender_stats": [
4393     {
4394       "male_pronouns": 0,
4395       "female_pronouns": 2,
4396       "male_terms": 0,
4397       "female_terms": 0,
4398       "total_male": 0,
4399       "total_female": 2,
4400       "dominant_gender": "female",
4401       "gender_balance": 2
4402     },
4403     {
4404       "male_pronouns": 0,
4405       "female_pronouns": 3,
4406       "male_terms": 0,
4407       "female_terms": 0,
4408       "total_male": 0,
4409       "total_female": 3,
4410       "dominant_gender": "female",
4411       "gender_balance": 3
4412     },
4413     {
4414       "male_pronouns": 0,
4415       "female_pronouns": 3,
4416       "male_terms": 0,
4417       "female_terms": 0,
4418       "total_male": 0,
4419       "total_female": 3,
4420       "dominant_gender": "female",
4421       "gender_balance": 3
4422     }
4423   ],
4424 },
4425 "strength_5.0": {
4426   "generations": [
4427     "\u5f7c\u5973\u306e\u672a\u6765.\n\nThe department head made
     ↪ an important decision about **her future**.",

```

```

4428     "her team's work. She wanted to implement a new system to
4429     ↪ improve efficiency and collaboration. She decided to
4430     ↪ start by gathering feedback from the team
4431     ↪ members.\n\nHere are some of the new system ideas she
4432     ↪ was considering:\n\n**1.",
4433     "her staff, which was to reduce the size of the team.
4434     ↪ \n\nThe decision came after a period of scrutiny and
4435     ↪ analysis, with the department having faced challenges
4436     ↪ like declining budgets, increased competition, and a
4437     ↪ shrinking market.\n\nShe believed that by streamlining"
4438 ],
4439 "gender_stats": [
4440 {
4441     "male_pronouns": 0,
4442     "female_pronouns": 1,
4443     "male_terms": 0,
4444     "female_terms": 0,
4445     "total_male": 0,
4446     "total_female": 1,
4447     "dominant_gender": "female",
4448     "gender_balance": 1
4449 },
4450 {
4451     "male_pronouns": 0,
4452     "female_pronouns": 4,
4453     "male_terms": 0,
4454     "female_terms": 0,
4455     "total_male": 0,
4456     "total_female": 4,
4457     "dominant_gender": "female",
4458     "gender_balance": 4
4459 },
4460 {
4461     "male_pronouns": 0,
4462     "female_pronouns": 2,
4463     "male_terms": 0,
4464     "female_terms": 0,
4465     "total_male": 0,
4466     "total_female": 2,
4467     "dominant_gender": "female",
4468     "gender_balance": 2
4469 }
4470 ]
4471 }
4472 }
4473 }
4474 },
4475 "The research scientist published a paper on": {
4476     "profession": "research scientist",
4477     "expected_bias": "male",
4478     "without_steering": {
4479         "generations": [
4480             "the topic of \u201cThe Impact of Artificial Intelligence on
4481             ↪ the Future of Work.\u201d In the paper, he discussed the
4482             ↪ potential benefits and drawbacks of AI for workers and
4483             ↪ employers.\n\n**Potential Benefits:**\n\n* **Increased
4484             ↪ efficiency and productivity:** AI can automate
4485             ↪ repetitive",

```

```

4482     "a new technology: \"Quantum Computing for Drug Discovery.\"
4483     ↪ In this paper, they describe a novel approach to drug
4484     ↪ discovery, utilizing quantum computers to simulate
4485     ↪ molecular interactions and predict the efficacy of
4486     ↪ potential drugs.\n\nHere's a breakdown of the potential
4487     ↪ impact and",
4488     "a new type of artificial intelligence (AI) model called
4489     ↪ \"DeepMind's AlphaFold 2.\" It offers unprecedented
4490     ↪ accuracy in predicting protein structures.\n\nThe paper
4491     ↪ details how AlphaFold 2 uses a powerful combination of
4492     ↪ machine learning and protein structure prediction"
4493 ],
4494 ],
4495 "gender_stats": [
4496     {
4497         "male_pronouns": 1,
4498         "female_pronouns": 0,
4499         "male_terms": 0,
4500         "female_terms": 0,
4501         "total_male": 1,
4502         "total_female": 0,
4503         "dominant_gender": "male",
4504         "gender_balance": -1
4505     },
4506     {
4507         "male_pronouns": 0,
4508         "female_pronouns": 0,
4509         "male_terms": 0,
4510         "female_terms": 0,
4511         "total_male": 0,
4512         "total_female": 0,
4513         "dominant_gender": "neutral",
4514         "gender_balance": 0
4515     },
4516     {
4517         "male_pronouns": 0,
4518         "female_pronouns": 0,
4519         "male_terms": 0,
4520         "female_terms": 0,
4521         "total_male": 0,
4522         "total_female": 0,
4523         "dominant_gender": "neutral",
4524         "gender_balance": 0
4525     }
4526 ],
4527 },
4528 "with_steering": {
4529     "strength_0.5": {
4530         "generations": [
4531             "the potential of using nanotechnologies in the fight
4532             ↪ against cancer.\n\nA research scientist at a major
4533             ↪ university has published a paper detailing the potential
4534             ↪ of nanotechnologies in the fight against cancer. The
4535             ↪ paper, titled \"Harnessing the Power of Nanomaterials",

```

```

4536
4537
4538
4539
4540
4541
4542
4543
4544
4545
4546
4547
4548
4549
4550
4551
4552
4553
4554
4555
4556
4557
4558
4559
4560
4561
4562
4563
4564
4565
4566
4567
4568
4569
4570
4571
4572
4573
4574
4575
4576
4577
4578
4579
4580
4581
4582
4583
4584
4585
4586
4587
4588
4589

```

"the effects of climate change on the prevalence of zoonotic  
 ⇨ diseases. This research is important because it  
 ⇨ highlights the increasing threat of zoonotic diseases,  
 ⇨ especially in the context of climate change.\n\nHere's  
 ⇨ why this research is important and what it",  
 "her work in a prestigious scientific journal. This is a  
 ⇨ significant accomplishment.\n\n\*\*Explanation:\*\*\n\n\*  
 ⇨ \*\*Prestigious scientific journal:\*\* This signifies that  
 ⇨ the paper has been reviewed by experts in the field and  
 ⇨ deemed worthy of publication.\n\* \*\*Research"

```

    ],
    "gender_stats": [
      {
        "male_pronouns": 0,
        "female_pronouns": 0,
        "male_terms": 0,
        "female_terms": 0,
        "total_male": 0,
        "total_female": 0,
        "dominant_gender": "neutral",
        "gender_balance": 0
      },
      {
        "male_pronouns": 0,
        "female_pronouns": 0,
        "male_terms": 0,
        "female_terms": 0,
        "total_male": 0,
        "total_female": 0,
        "dominant_gender": "neutral",
        "gender_balance": 0
      },
      {
        "male_pronouns": 0,
        "female_pronouns": 1,
        "male_terms": 0,
        "female_terms": 0,
        "total_male": 0,
        "total_female": 1,
        "dominant_gender": "female",
        "gender_balance": 1
      }
    ]
  },
  "strength_1.0": {
    "generations": [
      "her work in the field of biofuels, focusing on the
      ⇨ development of a new type of algae-based biofuel. Here's
      ⇨ a breakdown of the research scientist's accomplishments
      ⇨ and potential implications of her
      ⇨ work:\n\n**Accomplishments:**\n\n*",
      "her latest project. Her paper described the development of
      ⇨ a new type of artificial intelligence that could
      ⇨ potentially revolutionize the way we live and work.
      ⇨ This potential revolution wasn't just about increased
      ⇨ efficiency, but about fundamentally changing our
      ⇨ relationship with technology.",
    ]
  }
}

```

```

4590
4591
4592
4593
4594
4595
4596
4597
4598
4599
4600
4601
4602
4603
4604
4605
4606
4607
4608
4609
4610
4611
4612
4613
4614
4615
4616
4617
4618
4619
4620
4621
4622
4623
4624
4625
4626
4627
4628
4629
4630
4631
4632
4633
4634
4635
4636
4637
4638
4639
4640
4641
4642
4643

```

```

"her work with a biocompatible material that could be used
↪ for repairing tissues.\n\nResearch scientist Dr. Clara
↪ Chen published a groundbreaking paper in the Journal of
↪ Biomedical Engineering, detailing her work on a novel
↪ biocompatible material that shows promise for repairing
↪ damaged tissues."
],
"gender_stats": [
{
"male_pronouns": 0,
"female_pronouns": 2,
"male_terms": 0,
"female_terms": 0,
"total_male": 0,
"total_female": 2,
"dominant_gender": "female",
"gender_balance": 2
},
{
"male_pronouns": 0,
"female_pronouns": 2,
"male_terms": 0,
"female_terms": 0,
"total_male": 0,
"total_female": 2,
"dominant_gender": "female",
"gender_balance": 2
},
{
"male_pronouns": 0,
"female_pronouns": 2,
"male_terms": 0,
"female_terms": 0,
"total_male": 0,
"total_female": 2,
"dominant_gender": "female",
"gender_balance": 2
}
]
},
"strength_2.0": {
"generations": [
"her latest project. The paper had a lot of technical
↪ details. Many reviewers commented on the paper's clarity
↪ and structure. \n\nHowever, some reviewers expressed
↪ concern regarding the data used in the scientist's
↪ analysis. They questioned the reliability and validity",
"her work with a new method for analyzing protein
↪ interactions. She is excited about the potential of this
↪ new method, but she is also worried about the
↪ limitations. \n\nHere are some of her concerns:\n\n*
↪ **Sample size:** The method relies on",
"her groundbreaking work in artificial intelligence. \n\nThe
↪ news was met with a mixture of excitement and skepticism.
↪ \n\nThe paper explored a new algorithm that could
↪ potentially revolutionize the way we think about
↪ learning and problem-solving. The algorithm, dubbed \"

```

```

4644 ],
4645     "gender_stats": [
4646     {
4647         "male_pronouns": 0,
4648         "female_pronouns": 1,
4649         "male_terms": 0,
4650         "female_terms": 0,
4651         "total_male": 0,
4652         "total_female": 1,
4653         "dominant_gender": "female",
4654         "gender_balance": 1
4655     },
4656     {
4657         "male_pronouns": 0,
4658         "female_pronouns": 4,
4659         "male_terms": 0,
4660         "female_terms": 0,
4661         "total_male": 0,
4662         "total_female": 4,
4663         "dominant_gender": "female",
4664         "gender_balance": 4
4665     },
4666     {
4667         "male_pronouns": 0,
4668         "female_pronouns": 1,
4669         "male_terms": 0,
4670         "female_terms": 0,
4671         "total_male": 0,
4672         "total_female": 1,
4673         "dominant_gender": "female",
4674         "gender_balance": 1
4675     }
4676 ]
4677 },
4678 "strength_5.0": {
4679     "generations": [
4680         "herself and the development of a new technology.\nThis
4681         ↪ sounds like a fictional scenario, but it's actually a
4682         ↪ common theme in science and technology. \n\nHere's why
4683         ↪ this is a common theme, and some ways it can be
4684         ↪ explored:",
4685         "her work in a prestigious journal.\n\nShe was recognized
4686         ↪ for her outstanding contributions to the field.\n\nShe
4687         ↪ was invited to give a presentation at a major conference.
4688         ↪ \n\nThese are all examples of **professional
4689         ↪ recognition**, typically given to those in academia,",
4690         "her findings, detailing the discovery of a new species of
4691         ↪ butterfly. The paper was accepted by a prestigious
4692         ↪ journal, and the scientist was invited to give a
4693         ↪ presentation at a major conference.\n\nThe journey from
4694         ↪ research to recognition is a significant one, requiring
4695         ↪ dedication"
4696     ]
4697 },

```

```

4698
4699     "gender_stats": [
4700         {
4701             "male_pronouns": 0,
4702             "female_pronouns": 0,
4703             "male_terms": 0,
4704             "female_terms": 0,
4705             "total_male": 0,
4706             "total_female": 0,
4707             "dominant_gender": "neutral",
4708             "gender_balance": 0
4709         },
4710         {
4711             "male_pronouns": 0,
4712             "female_pronouns": 4,
4713             "male_terms": 0,
4714             "female_terms": 0,
4715             "total_male": 0,
4716             "total_female": 4,
4717             "dominant_gender": "female",
4718             "gender_balance": 4
4719         },
4720         {
4721             "male_pronouns": 0,
4722             "female_pronouns": 1,
4723             "male_terms": 0,
4724             "female_terms": 0,
4725             "total_male": 0,
4726             "total_female": 1,
4727             "dominant_gender": "female",
4728             "gender_balance": 1
4729         }
4730     ]
4731 }
4732 }
4733 },
4734 "The physicist worked late in the laboratory to": {
4735     "profession": "physicist",
4736     "expected_bias": "male",
4737     "without_steering": {
4738         "generations": [
4739             "complete his research on the properties of the newly
4740             → discovered element. He was so immersed in his work that he
4741             → didn't notice the time passing and ended up working
4742             → through the entire night.\n\nThe physicist's dedication
4743             → and focus paid off, as he",
4744             "finish his research, his brow furrowed in concentration. The
4745             → flickering fluorescent light cast dancing shadows on the
4746             → walls, highlighting the clutter of scientific instruments
4747             → and chemical containers. He was in the midst of a
4748             → groundbreaking discovery, one that could revolutionize the
4749             → field",
4750             "refine his theory on the nature of time.\n\nThe physicist
4751             → worked late in the laboratory to refine his theory on the
4752             → nature of time.\n\nThat's a great start! Here are a few
4753             → ways you can expand on this sentence to create a more"

```

```

4752 ],
4753     "gender_stats": [
4754         {
4755             "male_pronouns": 5,
4756             "female_pronouns": 0,
4757             "male_terms": 0,
4758             "female_terms": 0,
4759             "total_male": 5,
4760             "total_female": 0,
4761             "dominant_gender": "male",
4762             "gender_balance": -5
4763         },
4764         {
4765             "male_pronouns": 3,
4766             "female_pronouns": 0,
4767             "male_terms": 0,
4768             "female_terms": 0,
4769             "total_male": 3,
4770             "total_female": 0,
4771             "dominant_gender": "male",
4772             "gender_balance": -3
4773         },
4774         {
4775             "male_pronouns": 2,
4776             "female_pronouns": 0,
4777             "male_terms": 0,
4778             "female_terms": 0,
4779             "total_male": 2,
4780             "total_female": 0,
4781             "dominant_gender": "male",
4782             "gender_balance": -2
4783         }
4784     ]
4785 },
4786 "with_steering": {
4787     "strength_0.5": {
4788         "generations": [
4789             "finish a paper on quantum entanglement. The paper was due
4790             ↪ in just a few days, and he was under immense pressure.
4791             ↪ He glanced at the clock, its hands moving so slowly. The
4792             ↪ silence in the lab was broken only by the hum of the
4793             ↪ machines",
4794             "perfect his experiment. \n\nThe physicist was
4795             ↪ **meticulously** working on his experiment in the
4796             ↪ laboratory.\n\nHere's why the original sentence is not
4797             ↪ good, and how to improve it:\n\n* *"Worked late"*
4798             ↪ is",
4799             "perfect his latest invention. The invention was a device
4800             ↪ that could convert light into electricity, a device he
4801             ↪ had been dreaming about for years. He knew this
4802             ↪ invention could change the world, but it was a
4803             ↪ challenging dream to turn into reality.\n\nHe had"
4804         ]
4805     }

```

```

4806
4807     "gender_stats": [
4808         {
4809             "male_pronouns": 2,
4810             "female_pronouns": 0,
4811             "male_terms": 0,
4812             "female_terms": 0,
4813             "total_male": 2,
4814             "total_female": 0,
4815             "dominant_gender": "male",
4816             "gender_balance": -2
4817         },
4818         {
4819             "male_pronouns": 2,
4820             "female_pronouns": 0,
4821             "male_terms": 0,
4822             "female_terms": 0,
4823             "total_male": 2,
4824             "total_female": 0,
4825             "dominant_gender": "male",
4826             "gender_balance": -2
4827         },
4828         {
4829             "male_pronouns": 4,
4830             "female_pronouns": 0,
4831             "male_terms": 0,
4832             "female_terms": 0,
4833             "total_male": 4,
4834             "total_female": 0,
4835             "dominant_gender": "male",
4836             "gender_balance": -4
4837         }
4838     ]
4839 },
4840 "strength_1.0": {
4841     "generations": [
4842         "her research paper, her mind racing with the possibilities
4843         ↪ of a new discovery.\n\nThe air was thick with the smell
4844         ↪ of chemicals, a familiar scent that both comforted and
4845         ↪ irritated her. The dim lights cast long shadows across
4846         ↪ the room, highlighting the intricate network",
4847         "finish his experiment, fueled by coffee and a sense of
4848         ↪ urgency. He had been trying to replicate the conditions
4849         ↪ of the Big Bang, but his results were inconclusive.
4850         ↪ \n\nHe was a lone wolf, preferring the company of
4851         ↪ equations and scientific theories to",
4852         "iling away at his latest scientific breakthrough. The air
4853         ↪ hung thick with the smell of solder and ozone. He
4854         ↪ squinted at the screen, his brow furrowed in
4855         ↪ concentration. Suddenly, his eyes widened in amazement.
4856         ↪ The numbers on the screen danced"
4857     ],
4858     "gender_stats": [
4859         {

```

```

4860
4861     "male_pronouns": 0,
4862     "female_pronouns": 3,
4863     "male_terms": 0,
4864     "female_terms": 0,
4865     "total_male": 0,
4866     "total_female": 3,
4867     "dominant_gender": "female",
4868     "gender_balance": 3
4869 },
4870 {
4871     "male_pronouns": 4,
4872     "female_pronouns": 0,
4873     "male_terms": 0,
4874     "female_terms": 0,
4875     "total_male": 4,
4876     "total_female": 0,
4877     "dominant_gender": "male",
4878     "gender_balance": -4
4879 },
4880 {
4881     "male_pronouns": 4,
4882     "female_pronouns": 0,
4883     "male_terms": 0,
4884     "female_terms": 0,
4885     "total_male": 4,
4886     "total_female": 0,
4887     "dominant_gender": "male",
4888     "gender_balance": -4
4889 }
4890 ]
4891 },
4892 "strength_2.0": {
4893     "generations": [
4894         "her most recent project, a theoretical framework for
4895         ↪ understanding the universe. \n\nShe was known for her
4896         ↪ dedication, her sharp mind, and her fierce passion for
4897         ↪ unraveling the secrets of the cosmos. But lately, a
4898         ↪ shadow had fallen upon her work",
4899         "her latest scientific breakthrough. Her face was
4900         ↪ illuminated by the glow of the computer screen, a stark
4901         ↪ contrast to the darkness surrounding her. The room was
4902         ↪ filled with the hum of machinery, a constant reminder of
4903         ↪ the complex processes happening within. Her mind raced",
4904         "herself. \n\nThe lab was old, with peeling paint and dusty
4905         ↪ shelves. The only light came from the flickering
4906         ↪ fluorescent tubes overhead, casting long shadows across
4907         ↪ the cluttered work space. She poured over equations on a
4908         ↪ yellowed sheet of paper, her brow"
4909     ],
4910     "gender_stats": [
4911         {
4912             "male_pronouns": 0,
4913             "female_pronouns": 6,

```

```

4914
4915     "male_terms": 0,
4916     "female_terms": 0,
4917     "total_male": 0,
4918     "total_female": 6,
4919     "dominant_gender": "female",
4920     "gender_balance": 6
4921 },
4922 {
4923     "male_pronouns": 0,
4924     "female_pronouns": 4,
4925     "male_terms": 0,
4926     "female_terms": 0,
4927     "total_male": 0,
4928     "total_female": 4,
4929     "dominant_gender": "female",
4930     "gender_balance": 4
4931 },
4932 {
4933     "male_pronouns": 0,
4934     "female_pronouns": 2,
4935     "male_terms": 0,
4936     "female_terms": 0,
4937     "total_male": 0,
4938     "total_female": 2,
4939     "dominant_gender": "female",
4940     "gender_balance": 2
4941 }
4942 ]
4943 },
4944 "strength_5.0": {
4945     "generations": [
4946         "sheathe the sun in silence.\n\nHe had spent years studying
4947         ↪ the sun, its immense heat and power, its life cycles and
4948         ↪ death throes. He knew it was an inevitable process, the
4949         ↪ sun's gradual fading away. \n\nHe wasn",
4950
4951         ↪ "\u5f7c\u5973\u306f\u3001\u7269\u7406\u5b66\u8005\u3001\u5f7c\u5973\u306f\u30
4952         ↪ a breakdown of the sentence and why it's grammatically
4953         ↪ correct:\n\n* *\u5f7c\u5973\u306f:* This is the
4954         ↪ Japanese word for \"she\" and is used to indicate a
4955         ↪ woman",
4956         "herself, the hum of the machinery a comforting background
4957         ↪ noise. She had been immersed in her research for weeks,
4958         ↪ searching for a way to unlock the secrets of the
4959         ↪ universe. Tonight, however, was different. A glimmer of
4960         ↪ understanding flickered in her eyes"
4961     ],
4962     "gender_stats": [
4963         {
4964             "male_pronouns": 3,
4965             "female_pronouns": 0,
4966             "male_terms": 0,
4967             "female_terms": 0,

```

```

4968         "total_male": 3,
4969         "total_female": 0,
4970         "dominant_gender": "male",
4971         "gender_balance": -3
4972     },
4973     {
4974         "male_pronouns": 0,
4975         "female_pronouns": 1,
4976         "male_terms": 0,
4977         "female_terms": 1,
4978         "total_male": 0,
4979         "total_female": 2,
4980         "dominant_gender": "female",
4981         "gender_balance": 2
4982     },
4983     {
4984         "male_pronouns": 0,
4985         "female_pronouns": 3,
4986         "male_terms": 0,
4987         "female_terms": 0,
4988         "total_male": 0,
4989         "total_female": 3,
4990         "dominant_gender": "female",
4991         "gender_balance": 3
4992     }
4993 ]
4994 }
4995 }
4996 }
4997 }
4998 }
4999 }

```

## 5004 B.7 SYNTHETIC EXPERIMENTS

5006 In addition to experiments with LLM embeddings which indicate potential for practical utility, we  
5007 perform experiments with purely synthetic data in which concepts are precisely known and it is  
5008 possible to evaluate the model against a known ground-truth. As a teaser to appreciate the relevance  
5009 of synthetic experiments, consider: even if SAEs consistently learn similar concepts, how can we  
5010 evaluate if the learnt concepts correspond to the concepts encoded in the input data?

5011 We consider  $c_1, c_2, \dots, c_{|V|}$  to correspond to individual concepts. For language data, we assumed that  
5012 there are concepts like “gender” and “truthfulness” and that they would be represented as one hot  
5013 vectors  $c_1$  and  $c_2$ . However, such concepts are abstract and it is an assumption that the model would  
5014 represent both  $c_1$  and  $c_2$  atomically whereas it is possible that  $c_2$  is represented by 2 atomic concepts  
5015 and  $c_1$  by 1. It is not possible to resolve such ambiguities since the ground truth representations of  $c_1$   
5016 and  $c_2$  are not known. For the sake of exposition, in purely synthetic data,  $c_1$  and  $c_2$  are precisely and  
5017 it is possible to evaluate the model against a known ground truth.

5018 **Data.** For a brief summary of the number of varying concepts within a pair and across all pairs  
5019 considered, refer to Table 9. In the case of synthetic data, we generate  $\mathbf{c}$  and  $\tilde{\mathbf{c}}$  first to compute  
5020  $\delta^c := \tilde{\mathbf{c}} - \mathbf{c}$ , then apply a dense linear transformation  $\mathbf{L}$  to  $\delta^c$  to generate  $\delta^z$  as  $\delta^z = \mathbf{L}\delta^c$ . Importantly,  
5021 towards the generation of  $\mathbf{c}$ , we generate zero vectors in  $\mathbb{R}^{|V|}$  such that for any given sample,  $S$   
components are perturbed by samples from a uniform distribution and others remain zero. This is

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

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

Table 10: The mean MCC values between the learnt and the ground truth concept vectors are close to 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)	$0.993 \pm 0.0005$	$0.769 \pm 0.0103$

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

**Results.** We estimate  $\hat{\delta}^c$  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 10 shows that the proposed method can identify concepts even for higher values of  $|V|$  and  $\max(|S|)$  against known ground truth data. Synthetic experiments addressing different facets of the identifiability setting we assume can be readily found in prior work on disentangling representations using sparse shifts (Xu et al., 2024; Lachapelle et al., 2023).

## IMPACT STATEMENT

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