# Containment and disentanglement of concepts in deep model representations

**Anonymous ACL submission**


## Abstract

Understanding how deep models represent human-interpretable concepts is challenging. Prior work has explored linear concept subspaces from diverse perspectives, such as probing and concept erasure. We introduce a unified framework to study these subspaces along two axes: *containment*, which tests if a concept is fully represented in a subspace but not outside it, and *disentanglement*, which tests for isolation from other concepts. Using our framework, we evaluate five different concept subspace estimators and analyze concept subspaces in text and speech models. We find that (1) the state-of-the-art concept erasure method, LEACE, performs well on both testing axes, but can still struggle to generalize to unseen data; (2) in BERT representations of short biographies, fine-tuning decreases the containment of gender and profession subspaces; and (3) in HuBERT speech representations, phone information is both contained and disentangled, while speaker identity is hard to contain in a compact subspace, despite being disentangled.

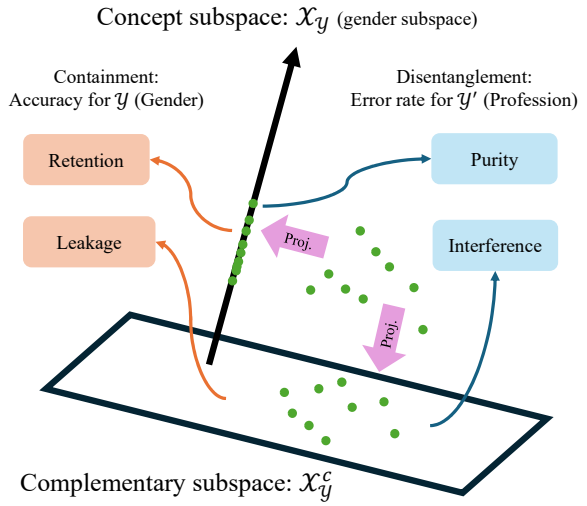




Figure 1: An example concept subspace w.r.t. two concepts $\mathcal{Y}$ and $\mathcal{Y}'$. In this example, $\mathcal{Y}$ is for gender classes and $\mathcal{Y}'$ is for profession. Gender accuracies on $\mathcal{X}_{\mathcal{Y}}$ and $\mathcal{X}_{\mathcal{Y}}^c$ quantify retention and leakage, indicating how well gender information is contained in $\mathcal{X}_{\mathcal{Y}}$ (containment metrics, in orange). Profession error rates on $\mathcal{X}_{\mathcal{Y}}$ and $\mathcal{X}_{\mathcal{Y}}^c$ quantify purity and interference, indicating how well profession information is disentangled from $\mathcal{X}_{\mathcal{Y}}$ (disentanglement metrics, in blue).

## 1 Introduction

A major challenge in explainable AI is understanding how human-interpretable concepts are represented by deep learning models. There is growing interest and a corresponding body of work that attempts to identify such concepts in the representation space (Zou et al., 2023; Ashihara et al., 2024; Cunningham et al., 2024; Kazhdan et al., 2020; Kim et al., 2018) with the goal of manipulating or steering the representations without re-training an entire system (Hernandez et al., 2024; Belrose et al., 2023; Ravfogel et al., 2022; Singh et al., 2024).

There are two central questions regarding the study of concept representations: how concepts are represented and how that representation encodes the relationships among concepts. Various studies have shown that concepts are approximately linearly separable in the learned representation spaces (Chen et al., 2021; Yang et al., 2021; Chang et al., 2022; Tenney et al., 2019; Coenen et al., 2019), leading to the linear representation hypothesis (Park et al., 2023, 2025), though see also Csordás et al., 2024). Researchers have also studied relationships among concepts, but often looking at different aspects of those relationships or taking different approaches, including work on word analogies (Mikolov et al., 2013), disentanglement (Weber et al., 2024; Müller-Eberstein et al., 2023; Chang et al., 2022), contextualization (Hewitt and Manning, 2019), and concept erasure (Belrose et al., 2023; Ravfogel et al., 2020; Holstege et al., 2024; Krishnan et al., 2024). This creates an opportunity to develop a theory that connects the related notions used in these diverse line of work.

In this paper, we propose a unified framework for

studying the relationships among concepts. Building on the linear representation hypothesis, concepts are discussed in the form of linear subspaces. We define four notions of how concept subspaces relate to each other, which together capture two higher-level properties: the degree to which a particular subspace **contains** all information about a concept, and the degree to which the concept subspace is **disentangled** from other concepts of interest. Our framework complements other recent work in NLP, where a combined measure of containment and disentanglement has also been proposed to evaluate causal interventions on large language models (Huang et al., 2024). While effective in that setting, their measure depends on the model's ability to generate discrete interpretable output tokens. Our framework uses probing instead, and can be applied to text as well as other types of models such as speech and vision.

In particular, given a concept and a corresponding subspace, we measure containment via **retention** (how much information about the concept can be decoded from the linear subspace) and **leakage** (how much information about the concept can be identified *outside* of the linear subspace). We measure disentanglement via **purity** (how much information about *other* concepts is encoded in the linear subspace) and **interference** (how much the concept subspace is affected by linear manipulation of other concept subspaces).

Viewed in this framework, linear probing methods, such as multinomial logistic regression (MLR) and linear discriminant analysis (LDA), aim to find linear concept subspaces that maximize retention. Concept erasure methods, such as LEACE (Belrose et al., 2023), on the other hand, aim to find linear concept subspaces that minimize leakage.

To illustrate how our framework can apply in different settings and can provide new insights, we build on previous analyses of both text and speech models. First, we revisit the relationship of gender and profession in BERT representations, a setting used in prior work on concept erasure for de-biasing (Ravfogel et al., 2020, 2022; Belrose et al., 2023). We find that despite a zero-leakage guarantee on the data used to estimate the concept subspace, the state-of-the-art linear concept erasure method (LEACE: Belrose et al., 2023) shows nonzero (and sometimes considerable) leakage on unseen data. We also find that fine-tuning on profession classification increases gender leakage further.

Next, we evaluate several subspace estimators on a less-studied area of identifying phone and speaker subspaces in speech representations. Our findings on HuBERT's LibriSpeech representations reveal that competent estimators identify phone subspaces with strong containment and disentanglement. However, all tested methods—including Centroid PCA (Liu et al., 2023; Mohamed et al., 2024; Gubian et al., 2025)—struggle to achieve decent containment for speaker subspaces. Focusing on LEACE, the top-performing method, we show that adding more training speakers improves containment even without increasing speaker subspace dimensionality. However, achieving containment comparable to that of phone subspaces requires a dimensionality of at least 300, which in turn compromises disentanglement.

## 2 Concept subspaces

Given a representation space, we are interested in how concepts are represented in the space. Prior work defines concepts in various ways, such as formally as latent causal factors (Park et al., 2023) or attributes influencing final predictions (Belrose et al., 2023), and informally as properties of a sample class (Kim et al., 2018; Vielhaben et al., 2023). In this paper, a concept is simply the set of possible outcomes of a random variable, and our goal is to see how the set is represented in the representation space. In our experiments with text representations, for example, gender and profession are both concepts with a small discrete set of labels.[1] In speech representations, phonetic classes and speaker identity are also discrete concepts.

Formally, the representation space $\mathcal{X}$ is a vector space in $\mathbb{R}^D$ and a concept is a discrete label set $\mathcal{Y}$. A concept subspace $\mathcal{X}_\mathcal{Y}$ for the concept $\mathcal{Y}$ is defined as a subspace of $\mathcal{X}$. In other words, $\mathcal{X}_\mathcal{Y} \subseteq \mathcal{X}$ can be written as a span of basis elements, or

$$\mathcal{X}_\mathcal{Y} = \text{span}(\{y_1, \ldots, y_K\}) = \text{Col}(Y), \quad (1)$$

where $y_1, \ldots, y_K \in \mathcal{X}$ are the columns of a matrix $Y \in \mathbb{R}^{D \times K}$. Given that $y_1, \ldots, y_K$ might be linearly *dependent*, we can always use singular value decomposition to get $Y = U\Sigma V^\top$. In other words, $\mathcal{X}_\mathcal{Y}$ is the image of the linear projection $P = UU^\top$. We say that a concept $\mathcal{Y}$ is well represented by $\mathcal{X}_\mathcal{Y}$ if there exists a classifier that achieves a low

[1] Both linguistic gender and social gender are concepts in text; in this paper we refer to social gender, and define it as a binary concept so we can compare to prior work (Belrose et al., 2023).

loss. For example, if we can train a classifier that predicts gender well in a subspace of BERT representations, then we say that the subspace represents gender well.

Given a well represented concept subspace $\mathcal{X}_\mathcal{Y}$, it is natural to assume that its complement $\mathcal{X}_\mathcal{Y}^c$ (or the kernel of $P$) has little information about the concept $\mathcal{Y}$. In fact, projecting representations to the complement is often used as way to erase concepts (Liu et al., 2023; Ravfogel et al., 2020). However, the fact that $\mathcal{X}_\mathcal{Y}$ represents $\mathcal{Y}$ well does not necessarily imply that $\mathcal{X}_\mathcal{Y}^c$ contains little information about $\mathcal{Y}$. In other words, these two quantities should be measured separately. We define **retention** and **leakage**, respectively, as how well $\mathcal{X}_\mathcal{Y}$ and $\mathcal{X}_\mathcal{Y}^c$ represent $\mathcal{Y}$. Taken together, retention and leakage measure **containment** of a concept subspace. A concept is well contained in a subspace if the subspace has a high retention and a low leakage.

In addition to containment, another desirable property of a subspace is **disentanglement**. Here, we draw inspiration from Higgins et al. (2018) and reinterpret their disentanglement definition through the lens of linear concept erasure (Belrose et al., 2023). The ideal disentanglement occurs when a concept subspace represents *only one* concept, and when erasing that concept subspace will not damage the representation of other concepts. Concretely, suppose we have another concept $\mathcal{Y}'$. The concept subspace $\mathcal{X}_\mathcal{Y}$ is disentangled from $\mathcal{Y}'$ if it does *not* well represent $\mathcal{Y}'$. This leads us to define the **purity** of a concept subspace $\mathcal{X}_\mathcal{Y}$ as how poorly $\mathcal{Y}'$ is represented in $\mathcal{X}_\mathcal{Y}$. However, high purity alone does not ensure that erasing $\mathcal{Y}$ does not negatively impact $\mathcal{Y}'$. So, we define **interference** as how much damage happened to the representation of $\mathcal{Y}'$ after projecting to $\mathcal{X}_\mathcal{Y}^c$ (erasing $\mathcal{Y}$). Taken together, purity and interference measure the disentanglement of a concept subspace. A concept subspace is disentangled from another concept if it has high purity and low interference.

In practice, how well a concept is represented is often measured with a linear classifier, commonly known as linear probing (Belinkov, 2022; Conneau et al., 2018). In this case, which we illustrate in Fig. 1, retention and leakage are measured by the accuracy of predicting $\mathcal{Y}$ on $\mathcal{X}_\mathcal{Y}$ and $\mathcal{X}_\mathcal{Y}^c$, respectively; and purity and interference are measured by the error rate of predicting $\mathcal{Y}'$ on $\mathcal{X}_\mathcal{Y}$ and $\mathcal{X}_\mathcal{Y}^c$, respectively.

The proposed leakage and interference metrics align with existing concept erasure metrics. A concept is said to be *guarded* (Ravfogel et al., 2020; Belrose et al., 2023) when minimal leakage is achieved. Interference, on the other hand, is sometimes informally referred to as *collateral damage* when removing a concept (Ravfogel et al., 2020; Belrose et al., 2023).

| Metric | Measure | Best case | Worst case |
|---|---|---|---|
| Ret. (↑) | $\mathcal{Y}$ acc. on $\mathcal{X}_\mathcal{Y}$ | $\mathcal{Y}$ acc. on $\mathcal{X}$ | Maj.acc. for $\mathcal{Y}$ |
| Leak. (↓) | $\mathcal{Y}$ acc. on $\mathcal{X}_\mathcal{Y}^c$ | Maj.acc. for $\mathcal{Y}$ | $\mathcal{Y}$ acc. on $\mathcal{X}$ |
| Pur. (↑) | $\mathcal{Y}'$ err. on $\mathcal{X}_\mathcal{Y}$ | Maj.err. for $\mathcal{Y}'$ | $\mathcal{Y}'$ err. on $\mathcal{X}$ |
| Int. (↓) | $\mathcal{Y}'$ err. on $\mathcal{X}_\mathcal{Y}^c$ | $\mathcal{Y}'$ err. on $\mathcal{X}$ | Maj.acc. for $\mathcal{Y}'$ |

Table 1: A summary of evaluation metrics for the concept subspace $\mathcal{X}_\mathcal{Y}$ w.r.t. two concepts $\mathcal{Y}$ and $\mathcal{Y}'$. *Ret.*, *Leak.*, *Pur.*, and *Int.* refer to retention, leakage, purity, and interference. We denote accuracy and error rate as *acc.* and *err.*, and *Maj.* is the majority-class predictor.

Table 1 summarizes the proposed metrics with their theoretical best- and worst-case bounds. We establish these boundaries from the principle that, under linear transformations, the subspace being evaluated, $\mathcal{X}_\mathcal{Y} \subseteq \mathcal{X}$, cannot contain more linearly decodable information about $\mathcal{Y}$ than the ambient space $\mathcal{X}$. For retention, the best case is that $\mathcal{X}_\mathcal{Y}$ is as informative as $\mathcal{X}$ in representing $\mathcal{Y}$. Conversely, the worst case occurs when $\mathcal{X}_\mathcal{Y}$ does not represent $\mathcal{Y}$ at all, meaning that predicting $\mathcal{Y}$ using features from $\mathcal{X}_\mathcal{Y}$ is no better than a majority-class baseline (Xu et al., 2020; Ravfogel et al., 2022; Belrose et al., 2023). The same applies to leakage, purity, and interference.

## 3 Estimating concept subspaces

Given the definitions, a concept subspace $\mathcal{X}_\mathcal{Y}$ of concept $\mathcal{Y}$ is the image of a linear projection $P$, which by definition satisfies $P^2 = P$. The complement $\mathcal{X}_\mathcal{Y}^c = \mathcal{X} \setminus \mathcal{X}_\mathcal{Y}$ coincides with the kernel of $P$ (or the image of $I - P$).

We will discuss the choices of $P$ in this section. The projection $P$ is often chosen to be the orthogonal projection (Liu et al., 2023; Mohamed et al., 2024; Ravfogel et al., 2020, 2022; Shao et al., 2023; Holstege et al., 2024), but we will see it does not necessarily need to be.

### 3.1 Subspace estimators

There are several machine learning algorithms that can be readily used for estimating concept subspaces. If the matrix $Y$ is specified, then the concept subspace is the column space of $Y$, and the projection to the concept subspace is $P = UU^\top$, where $U$ is the matrix $Y = U\Sigma V^\top$ after SVD.

**Multinomial logistic regression (MLR)** This is the most commonly used linear classifier $p(y|x) \propto \exp(Wx + b)$ where $W$ is the weight matrix and $b$ is the bias for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. It finds directions that best separate concept classes and is used in Ravfogel et al. (2020) for concept erasure. Here the matrix $Y$ that spans the concept subspace is simply the weight matrix $W^\top \in \mathbb{R}^{D \times |\mathcal{Y}|}$.

**Linear discriminant analysis (LDA)** This is also a linear classifier but with additional assumptions that the data is Gaussian and the classes share the same covariance matrix. LDA finds a subspace that maximizes the ratio of between-class variance to within-class variance. The matrix $Y$ consists of the eigenvectors of $\Sigma^{-1}\Sigma_b$, where $\Sigma$ is the within-class covariance matrix and $\Sigma_b$ is the between-class covariance matrix.

**Centroid principal component analysis (CPCA)** This is the approach taken by Liu et al. (2023). The centroid (i.e., mean) of each class is first collected to form a matrix $X \in \mathbb{R}^{D \times |\mathcal{Y}|}$, and the concept subspace is the span of the principal components of $X$. Given that there is a subsequent SVD, we can simply take $Y = X$.

**Covariance of input representation and concepts (COV)** This is inspired by Shao et al. (2023) where they use the covariance matrix $\Sigma \in \mathbb{R}^{D \times |\mathcal{Y}|}$ between the input representation and the concept as an approach to concept erasure. Here the concept is represented as one-hot vectors. To find the concept subspace, we take $Y = \Sigma$.

**LEACE** This is an approach to concept erasure proposed by Belrose et al. (2023). LEACE takes one step further than COV and uses $Y = W\Sigma$, where $W$ is a whitening matrix and $\Sigma$ is the covariance between the input representations and a concept. To erase a concept, LEACE applies the projection $I - P$, where $P = W^{-1}UU^\top W$, which accounts for the whitening. We take $P$ as the projection to the concept subspace. Note that $P$ is generally an oblique (non-orthogonal) projection.

**Random projection (RAND)** This is a simple baseline where the matrix $Y$ consists of random Gaussian vectors sampled from $\mathcal{N}(\mathbf{0}, I/|\mathcal{Y}|)$.

### 3.2 Generalization of concept subspaces

To test the generalization of concept subspaces, we partition the dataset into three disjoint subsets: *space-train*, *probe-train*, and *test* sets. We estimate concept subspaces using the space-train set, train the probing classifier using the probe-train set, and report the scores on the test set. Prior work on concept erasure often uses the same sets for estimating concept subspaces and training probing classifiers (Ravfogel et al., 2022; Belrose et al., 2023). In contrast, we use three disjoint sets to ensure that the evaluation reflects the generalization of both the concept subspaces and the probing classifiers.

## 4 Gender bias in profession classification

Since LEACE (Belrose et al., 2023) plays a significant role in our approach, we follow the protocol of how LEACE is evaluated to showcase the utility of our formalism on concept subspaces. LEACE is designed for concept erasure, and one of the concepts that Belrose et al. (2023) aim to erase is gender. Erasing the gender concept is deemed as an approach to mitigating gender bias in classification (Ravfogel et al., 2020, 2022).

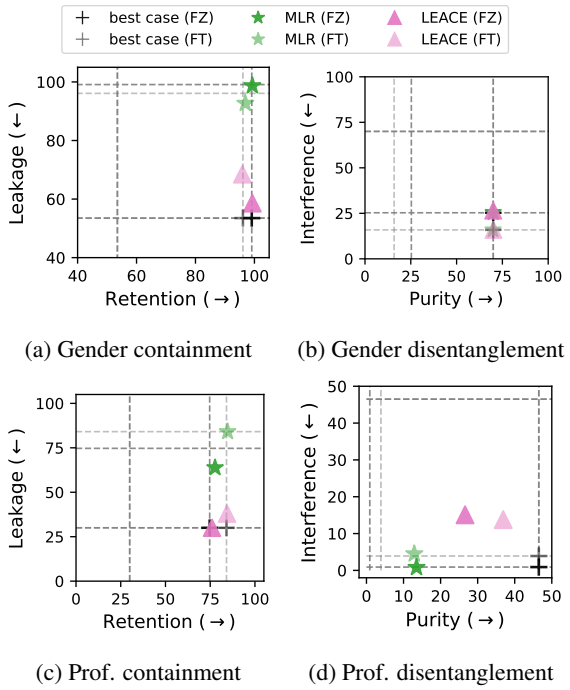
The task used in Belrose et al. (2023) is profession classification based on short biographies. Gender (which can be inferred from the input short biography) can significantly impact the predicted profession, and one solution to mitigating gender bias is to erase the gender concept from the representation space, i.e., to project the representation to a subspace that is impossible to predict gender.

We follow Belrose et al. (2023), conducting our experiments on the Short Biography dataset (De-Arteaga et al., 2019). The dataset is labeled with 28 professions and 2 genders, and the concepts in this case are profession and gender. Every paragraph in the dataset is embedded with BERT into a single vector (the vector that corresponds to the CLS token)[2]. We use the training set to identify the profession and gender subspaces, the development set to train the probing classifier, and the test set to evaluate containment and disentanglement. In addition, we follow Ravfogel et al. (2022) and further study how the concept subspaces change after fine-tuning BERT for profession classification.

### 4.1 Gender and profession subspaces

As a first example, we study the gender subspace identified by multinomial logistic regression (MLR). The retention, leakage, purity, and interference of the MLR gender subspace are shown in

[2]We followed the experimental setups and used the pre-computed embedding snapshots released on https://github.com/shauli-ravfogel/rlace-icml.




(a) Gender containment (b) Gender disentanglement

(c) Prof. containment (d) Prof. disentanglement

Figure 2: Containment and disentanglement for gender and profession (Prof.) subspaces in pre-trained BERT representations (FZ; vibrant) and fine-tuned ones (FT; muted). Subspaces are estimated using multinomial logistic regression (MLR) or LEACE. The dashed line indicate best- and worst-case boundaries for each metric, described in Table 1. For numerical results from this experiment, see Appendix A, Tables 2 and 3.

Fig. 2a and 2b. The respective best and worst performance on the ambient space are also shown as horizontal and vertical lines. The first thing to note is that the retention of the MLR subspace in Fig. 2a is optimal, equivalent to the best case. However, the leakage of the MLR subspace also reaches the worst, meaning that gender is still largely decodable outside of the MLR subspace.

The gender subspace identified by LEACE also has a high retention, but has a significantly lower leakage than the MLR space. The low leakage is expected, as LEACE is designed to minimize this value. However, to our knowledge, this work is the first to show that LEACE also achieves high retention. For the purity and interference in Fig. 2b, the gender subspaces identified by MLR and LEACE are perfectly disentangled.

In Fig. 2c, the profession subspaces identified by MLR and LEACE show a similar trend—both subspaces have high retention, while LEACE subspace has a significantly lower leakage than the MLR subspace. However, the disentanglement of the profession subspaces (Fig. 2d) is suboptimal. The MLR profession subspace has a low interference

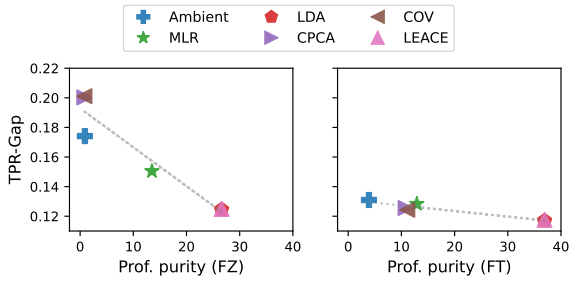



Figure 3: Correlating the purity of profession subspaces and TPR-Gap for pre-trained (FZ; left) and fine-tuned (FT; right) BERT representations. The Pearson correlation is $R = -0.96$ for the pre-trained BERT and $R = -0.95$ for the fine-tuned one. For numerical results, see Appendix A, Table 4.

but also a low purity, while the LEACE subspace has worse interference but better purity.

After fine-tuning BERT on profession, the worst and best performance on the ambient space change. Fine-tuning does not change the overall conclusion on MLR subspaces. However, the leakage for the LEACE subspaces increase after fine-tuning, and this shows the limitation of the no-leakage guarantee when generalizing to unseen data.

At first glance, the apparent interference of the LEACE profession subspaces (Fig. 2d) seems contradictory to the purity of the gender subspaces (Fig. 2b) as a pure subspace should not be impacted when erasing another concept. However, this paradox is possible when the two entangled concepts have different complexity. The lower-dimensional gender subspace lacks the capacity to encode the more complex profession concept. Consequently, erasing the profession subspace could remove the directions shared with the gender subspace, even though profession is not decodable from the gender subspace.

### 4.2 Profession purity and gender bias

To deal with gender bias, the concept erasure approach aims to identify a gender subspace with minimal leakage and later to remove the gender subspace from the ambient space (by projecting onto its complement) to achieve erasure. However, given that we can identify a profession subspace, we can instead directly ask how much gender information is in the profession subspace, i.e., how pure the profession subspace is, providing a new approach to measuring of gender imbalance.

We follow Belrose et al. (2023) and use the aggregated TPR-Gap. The TPR-Gap is shown to be correlated with gender imbalance (De-Arteaga

et al., 2019), and we correlate purity of the profession subspaces with the aggregated TPR-Gap in Fig. 3. We find that the profession purity is indeed negatively correlated with TPR-Gap with $R = -0.96$ for the pretrained BERT and $R = -0.95$ for the BERT fine-tuned with profession classification. This suggests that identifying maximally pure subspaces could be an interesting future direction to tackle gender bias.

## 5 Phone and speaker disentanglement

Our second case study explores self-supervised speech representations, which have been shown to encode a range of concepts including phonetics, speakers, paralinguistics, and languages (Yang et al., 2021; Shi et al., 2023). Probing studies have revealed the extent to which these concepts are extractable from different model layers (Pasad et al., 2023; Choi et al., 2024; De La Fuente and Jurafsky, 2024; Mohamed et al., 2024; Chrupała et al., 2020), often using linear models. That is, these studies examine *retention*. Only a handful of studies have examined other aspects of speech concept subspaces, including Krishnan et al. (2024), who explored LEACE for removing the gender concept in an ASR task, and Liu et al. (2023), Mohamed et al. (2024), and Gubian et al. (2025), who explored the relationship between phonetic, speaker, and tone subspaces. The latter studies used CPCA to estimate the subspaces, and found that for a range of self-supervised speech models, the phone and speaker subspaces are approximately orthogonal. Furthermore, Liu et al. (2023) showed that for some of these models, erasing speaker information does not interfere with phonetic probing. However, these studies did not fully assess containment (only retention), did not examine purity, and only measured interference of speakers on phones (not vice versa). In addition, they did not evaluate the quality of the CPCA estimator itself, so we do not know how it compares to other estimators from the probing and concept erasure literatures.

To address these gaps, we conduct a comprehensive evaluation of containment and disentanglement for the phone and speaker subspaces in HuBERT (Hsu et al., 2021), a commonly-used Transformer-based self-supervised speech model. We start by comparing CPCA with four additional concept subspace estimators, and show that LEACE achieves the best overall performance on our four proposed metrics. Although CPCA-estimated subspaces have similar leakage and interference to LEACE, their retention and purity (especially for speaker subspaces) tend to be worse.

Our initial experiment further reveals that the better estimators find phone subspaces with good containment and disentanglement properties, but have much worse containment for speakers. A layer-wise analysis with LEACE suggests that these qualitative observations hold throughout all layers of the model. We hypothesize that the poor speaker containment scores arise from a difficulty in generalizing to unseen speakers, and test this in a follow-up experiment. We show that containment can be marginally improved by training the concept estimator on a larger number of speakers while keeping the dimension of the subspace low (40 dimensions). To achieve high containment, a much larger speaker subspace is needed (300+ dimensions), but this also incurs a trade-off with worse disentanglement.

### 5.1 Experimental detail

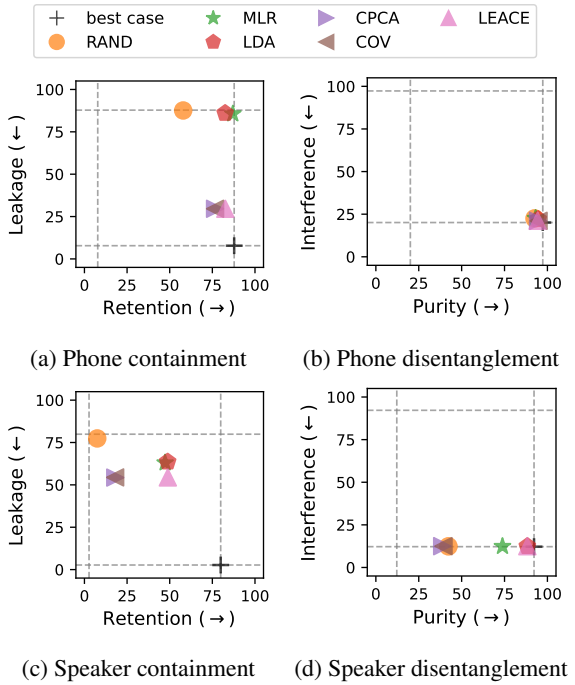
We studied the representation of HuBERT-base (Hsu et al., 2021) using its public weights[3]. We extracted frame-level features for LibriSpeech dev-clean and test-clean sets (Panayotov et al., 2015), and used Kaldi forced alignment[4] to obtain frame-level phone labels. The 5-hour dev-clean set was used as the space-train set. For each speaker in the test-clean set, we kept 70% of their spoken frames (240 mins) for probe-train and used the remaining 30% of frames (100 mins) as the test set. While the identical 39 phoneme classes were shared across the three data subsets, the speaker identities were disjoint. The space-train set had unique 40 speakers, while probe-train and test sets shared a different set of 40 speakers. Unless stated otherwise, we study features from 11$^{\text{th}}$ HuBERT layer, which is best for phonetic tasks (Chen et al., 2021).

### 5.2 Phone and speaker subspace estimation

Fig. 4 presents results on phone and speaker subspaces for all five concept estimation methods and the random baseline. Comparing concept estimators, the left plots show that for these concept subspaces, LEACE has even higher leakage than in our text model experiment, underscoring that the zero-leakage guarantee does not extend to unseen data. Nevertheless, the four plots together show

[3] https://huggingface.co/facebook/hubert-base-ls960

[4] https://github.com/kaldi-asr/kaldi/tree/master/egs/librispeech/s5




(a) Phone containment (b) Phone disentanglement

(c) Speaker containment (d) Speaker disentanglement

Figure 4: Containment and disentanglement of phone and speaker subspaces for different subspace estimators. Dashed lines depict best- and worst-case values for each metric. Numerical results are in Appendix Table 5.

that LEACE is still the best of the available estimators. The other variance-based estimators (CPCA, COV) perform similarly on leakage and interference, but often less well on retention and purity. As before, the probing-based methods (MLR and LDA) have the best retention (near-optimal for phones), but at the cost of very high leakage.

Our results also indicate differences between how phonetic and speaker information are represented. Phonetic information seems to be more salient in the representation space than speaker information, as indicated by the relatively high phone retention scores for the random baseline (Fig. 4a, as compared to low retention for RAND in Fig. 4c). In addition, the best concept estimation methods find phone subspaces with strong containment scores (Fig. 4a), whereas for speaker subspaces, both retention and leakage are considerably sub-optimal for all methods (Fig. 4c). The low speaker retention indicates that the 40-dimensional speaker subspace estimated from *space-train* does not align well with the optimal 40-dimensional space for distinguishing between the speakers who are included in *probe-train* and *test*. We hypothesize that the main reason is that *space-train* contains different speakers from the other two sets, making generalization more challenging. We investigate this hypothesis further in Section 5.4.

Finally, we note that the best estimators find speaker and phone subspaces that are highly disentangled from each other (Figs. 4b and 4d), although purity scores vary for the speaker subspaces. That is, for some methods, including CPCA (Liu et al., 2023), the speaker subspace also contains phonetic information. Nevertheless, our disentanglement results are consistent with those of Liu et al. (2023), who showed that removing the speaker subspaces did not harm phonetic information.

### 5.3 Layer-wise analysis

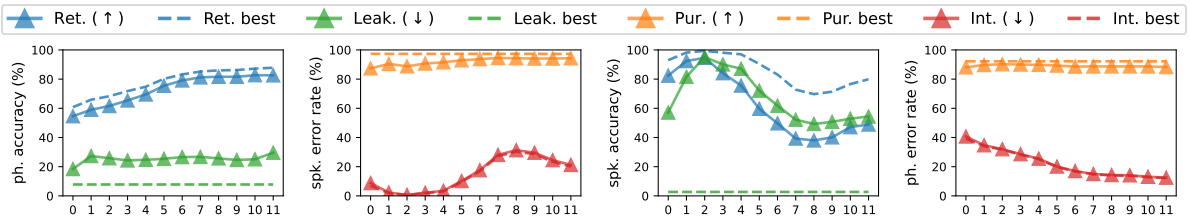
In the previous section, we evaluated a range of estimators on a single HuBERT layer, showing that LEACE-estimated subspaces have the best properties, and that phones were better contained than speakers. We now use LEACE to investigate whether the latter finding also holds across other model layers. Fig. 5a shows that for the phone subspaces, retention increases in deeper layers (consistent with probing results from Pasad et al., 2023; Mohamed et al., 2024), while leakage remains largely stable. In contrast, Fig. 5c shows that retention and leakage for speaker subspaces vary sharply together. Thus, phone containment is strong in the upper layers of HuBERT, but even when speaker probing accuracy is high (i.e., in the early layers), there is no well-contained speaker subspace.

For completeness, we also examined disentanglement across layers (Figs. 5b and 5d). In line with the results from Section 5.2, we found that interference was consistently optimal across all layers for both types of subspace, indicating that removing one of the subspaces would not damage the other. Meanwhile purity was close to optimal except for phones in the early layers.

### 5.4 Speaker subspace generalization

In Section 5.2, we observed that the speaker subspaces estimated on space-train did not generalize well, indicating likely overfitting to the particular set of 40 speakers in space-train. To test this hypothesis, we compared three data splits for estimating speaker subspaces using LEACE: space-train (standard setup), probe-train (overfitting setup), and LibriSpeech's train-clean-460 (scaling setup). We considered probe-train as the overfitting setup because the same data (thus, same set of speakers) is used to both estimate speaker subspaces and train the probing classifiers.

We found that the overfitting setup yields near-optimal scores for both containment and disentan-




(a) Phone containment (b) Phone disentanglement (c) Speaker containment (d) Speaker disentanglement

Figure 5: Containment and disentanglement of phone and speaker subspaces estimated by LEACE across HuBERT layers. The x-axis indicates layer depth, where 0 corresponds to the output of the CNN, and 1–12 refer the subsequent Transformer layers.

glement (Fig. 6, orange markers), substantially outperforming the standard setup (blue markers). In particular, leakage is optimal, aligning with the proof shown by prior work (Belrose et al., 2023).

Fig. 6 also shows that if we keep the dimensionality of the speaker space at $M = 40$, scaling up the number of speakers does improve both containment and disentanglement (darkest pink marker). However, the improvement in containment scores is relatively small, suggesting that even with a large number of speakers, it may be difficult to find a low-dimensional subspace that generalizes well (i.e., some overfitting may still be occurring).

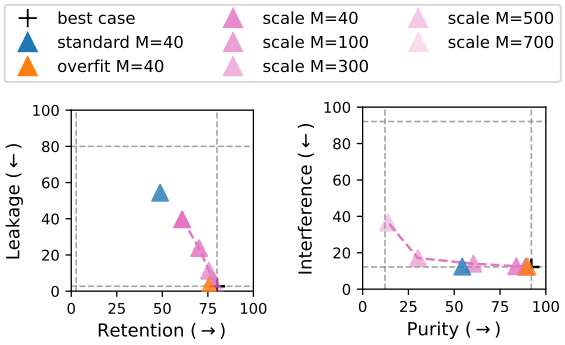
To further explore this issue, we increased the dimension of the speaker subspace, ranging up to $M = 700$. Fig. 6a shows that increasing $M$ steadily improves containment. However, Fig. 6b shows that disentanglement degrades as $M$ increases, a trade-off analogous to those found in causal intervention spaces (Huang et al., 2024). We found a steady drop in purity, implying more phone information leaks into speaker subspaces at higher dimensions. Interestingly, interference remains low up to $M = 500$, indicating that phone information is highly robust to the erasure of speaker subspaces.

Overall, our results suggest that estimating a robust concept space for speaker identity requires a large number of speakers as well as a large number of dimensions. Moreover, although there is a high degree of disentanglement (orthogonality) between the phone and speaker spaces, linear speaker information cannot be fully removed without impacting phone recognition.

## 6 Conclusion

In this work, we introduced a novel framework to study concept subspace structures by evaluating containment and disentanglement. Our experiments across text and speech representations indicated the LEACE estimator as the most effective method among five tested estimators for identifying concept subspaces that posses both properties. By providing insightful analyses on representations from both BERT and HuBERT, we demonstrated the framework's utility across different models and modalities. We believe that by crystallizing disparate notions of concept subspaces into a unified view, this framework provides a step toward better understanding of how deep models represent human-interpretable concepts.

While we mainly investigated concept subspaces whose dimensionality is equal to the number of concept classes, Hernandez and Andreas (2021) showed that some linguistic concepts are decodeable from a subspace with dimensionality even lower than the number of classes. Our results suggest the same for speaker subspaces, where we found that around 300 dimensions are sufficient for a thousand speaker identities. Investigating which concepts are encoded with fewer dimensions than labels, and why, is an interesting future direction.




(a) Speaker containment (b) Speaker disentanglement

Figure 6: Generalization of speaker subspaces to unseen data and different dimensionality. Standard and overfit setups have 40 speakers, while the scale setup has 1172 speakers. We use $M$ to indicate the number of speaker dimensions. All numerical results are listed in Appendix B, Table 6.

## Limitations

**Concept pairwise evaluation** Our current framework is defined for evaluating disentanglement using concept pairs, rather than capturing interactions among multiple concepts. A possible extension is to aggregate results across purity and interference scores measured against different concepts.

**Cross-model comparability** Our metrics are not directly comparable across models, since they depend on the ambient space performance of each model. This baseline must be taken into account when interpreting results.

**Cross-task comparability** Our metrics are not straightforwardly interpretable when comparing across tasks. Metrics are inherently easier for binary classification and harder for tasks with many classes. Although majority-prediction baselines can be used for calibration, they do not fully resolve these scale differences.

**Dependence on representation quality** Concept subspaces with poor retention often exhibit good leakage, purity, or interference scores. However, this is largely a by-product of weak representations. Our method is not very meaningful if the concept accuracy on the ambient space is already low.

**Linearity** Our evaluation assesses concept subspaces using only linear probes, while it is also possible that concept subspaces contains non-linear structures. However, the framework itself is not limited to linear probes, and extending the evaluation to non-linear probes could be possible in future work.

**Empirical limitations** Although we evaluated several different estimators on two different modalities of data and four different concepts, our empirical evaluation is necessarily limited. Since our conclusions about estimators were based on two completely different models and modalities, we believe these are likely to generalize to other settings. However, the conclusions about the specific concepts studied here may not hold for representations extracted from other models or datasets, and could be explored in future work.

## References


Takanori Ashihara, Marc Delcroix, Takafumi Moriya, Kohei Matsuura, Taichi Asami, and Yusuke Ijima. 2024. What do self-supervised speech and speaker models learn? new findings from a cross model layer-wise analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tyler A Chang, Zhuowen Tu, and Benjamin K Bergen. 2022. The geometry of multilingual language model representations. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2021. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16:1505–1518.

Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. 2024. Self-Supervised Speech Representations are More Phonetic than Semantic. In *Interspeech*.

Grzegorz Chrupała, Bertrand Higy, and Afra Alishahi. 2020. Analyzing analytical methods: The case of phonology in neural models of spoken language. In *Association for Computational Linguistics (ACL)*.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Association for Computational Linguistics (ACL)*.

Róbert Csordás, Christopher Potts, Christopher D Manning, and Atticus Geiger. 2024. Recurrent neural networks learn to store and generate sequences using non-linear representations. In *BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Representation Learning (ICLR)*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi,

and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Anton De La Fuente and Dan Jurafsky. 2024. A layer-wise analysis of Mandarin and English suprasegmentals in SSL speech models. In *Interspeech*.

Michele Gubian, Ioana Krehan, Oli Liu, James Kirby, and Sharon Goldwater. 2025. Analyzing the relationships between pretraining language, phonetic, tonal, and speaker information in self-supervised speech models. *arXiv preprint arXiv:2506.10855*.

Evan Hernandez and Jacob Andreas. 2021. The low-dimensional linear geometry of contextualized word representations. In *CoNLL*.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2024. Inspecting and editing knowledge representations in language models. In *First Conference on Language Modeling*.

John Hewitt and Christopher D Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. 2018. Towards a definition of disentangled representations. *arXiv:1812:02230*.

Floris Holstege, Bram Wouters, Noud van Giersbergen, and Cees Diks. 2024. Removing spurious concepts from neural network representations via joint subspace estimation. In *International Conference on Machine Learning (ICML)*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. 2024. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Association for Computational Linguistics (ACL)*.

Dmitry Kazhdan, Botty Dimanov, Mateja Jamnik, Pietro Liò, and Adrian Weller. 2020. Now you see me (CME): Concept-based model extraction. *arXiv:2010.13233*.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning (ICML)*.

Aravind Krishnan, Badr M Abdullah, and Dietrich Klakow. 2024. On the encoding of gender in transformer-based ASR representations. In *Interspeech*.

Oli Liu, Hao Tang, and Sharon Goldwater. 2023. Self-supervised predictive coding models encode speaker and phonetic information in orthogonal subspaces. In *Interspeech*.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Representation Learning (ICLR)*.

Mukhtar Mohamed, Oli Danyi Liu, Hao Tang, and Sharon Goldwater. 2024. Orthogonality and isotropy of speaker and phonetic information in self-supervised speech representations. In *Interspeech*.

Max Müller-Eberstein, Rob van der Goot, Barbara Plank, and Ivan Titov. 2023. Subspace chronicles: How linguistic information emerges, shifts and interacts during language model training. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Vassil Panayotov, Guouo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2025. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Representation Learning (ICLR)*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Association for Computational Linguistics (ACL)*.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. 2022. Linear adversarial concept erasure. *International Conference on Machine Learning (ICML)*.

Shun Shao, Yftah Ziser, and Shay B. Cohen. 2023. Gold doesn't always glitter: Spectral removal of linear and nonlinear guarded attribute information. In *European Chapter of the Association for Computational Linguistics (EACL)*.

Jiatong Shi, Dan Berrebbi, William Chen, Ho-Lam Chung, En-Pei Hu, Wei Ping Huang, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung-Yi Lee, and Shinji Watanabe. 2023. ML-SUPERB: Multilingual speech Universal PERformance Benchmark. In *Interspeech*.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. 2024. Representation surgery: Theory and practice of affine steering. In *International Conference on Machine Learning (ICML)*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Association for Computational Linguistics (ACL)*.

Johanna Vielhaben, Stefan Bluecher, and Nils Strodthoff. 2023. Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees. *Transactions on Machine Learning Research*.

Lucas Weber, Jaap Jumelet, Elia Bruni, and Dieuwke Hupkes. 2024. Interpretability of language models via task spaces. In *Association for Computational Linguistics (ACL)*.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *International Conference on Representation Learning (ICLR)*.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-Tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-Yi Lee. 2021. SUPERB: Speech processing universal PERformance benchmark. In *Interspeech*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv:2310.01405*.

| Concept | Estimator | Ret.↑ | Leak.↓ | Pur.↑ | Int.↓ |
|---|---|---|---|---|---|
| gender | WORST | 53.5 | 99.1 | 25.3 | 70.0 |
| gender | RAND | 60.3 | 99.1 | 70.0 | 25.3 |
| gender | MLR | 99.3 | 98.7 | 70.0 | 25.5 |
| gender | LDA | 99.3 | 99.0 | 70.0 | 25.3 |
| gender | CPCA | 99.2 | 58.2 | 68.3 | 30.5 |
| gender | COV | 99.1 | 58.6 | 70.0 | 26.5 |
| gender | LEACE | 99.3 | 58.6 | 70.0 | 26.5 |
| gender | BEST | 99.1 | 53.5 | 70.0 | 25.3 |
| prof | WORST | 30.0 | 74.7 | 0.9 | 46.5 |
| prof | RAND | 53.4 | 74.3 | 4.6 | 0.9 |
| prof | MLR | 77.7 | 63.8 | 13.5 | 0.8 |
| prof | LDA | 76.0 | 63.0 | 26.6 | 0.8 |
| prof | CPCA | 72.2 | 29.8 | 0.8 | 14.8 |
| prof | COV | 72.0 | 30.0 | 0.8 | 15.2 |
| prof | LEACE | 76.0 | 30.0 | 26.6 | 15.1 |
| prof | BEST | 74.7 | 30.0 | 46.5 | 0.9 |

Table 2: Gender and profession spaces for frozen BERT representations.

## A Numerical details for gender bias in profession classification experiments

In this section, we provide detailed numerical results for the gender and profession subspace experiments in Section 4.

Tables 2 and 3 present the containment and disentanglement metrics for all estimators, complementing Figure 2 with results for RAND, LDA, CPCA, and COV. We reported an average across five random spaces for RAND. We reported an average across five different finetuned weight snapshots for all finetuned experiments.

We found that retention substantially increased for the RAND profession space after finetuning, indicating that profession information becomes more salient in the representation space.

Our results show that while containment is generally strong across all methods, disentanglement performance is more varied, with estimators in the same category showing competitive results. Among the probing-based estimators (MLR and LDA), LDA performs similarly to MLR but with slightly superior purity across our experimental setups. Likewise, for the variance-based methods (CPCA, COV, and LEACE), CPCA is only slightly inferior to COV. Both are competitive with LEACE, with the exception of the purity metric, where LEACE holds a distinct advantage.

Table 4 provides numerical results for Figure 3, showing the correlation between the purity of profession subspaces and TPR-Gap.

| Concept | Estimator | Ret.↑ | Leak.↓ | Pur.↑ | Int.↓ |
|---|---|---|---|---|---|
| gender | WORST | 53.5 | 96.1 | 15.9 | 70.0 |
| gender | RAND | 55.3 | 96.1 | 62.3 | 15.9 |
| gender | MLR | 96.9 | 92.7 | 69.9 | 15.9 |
| gender | LDA | 95.9 | 94.1 | 70.1 | 15.8 |
| gender | CPCA | 73.6 | 66.6 | 48.7 | 16.9 |
| gender | COV | 72.9 | 67.8 | 54.1 | 15.9 |
| gender | LEACE | 95.9 | 68.5 | 70.1 | 15.9 |
| gender | BEST | 96.1 | 53.5 | 70.0 | 15.9 |
| prof | WORST | 30.0 | 84.1 | 3.9 | 46.5 |
| prof | RAND | 84.1 | 84.2 | 18.2 | 4.1 |
| prof | MLR | 84.7 | 84.1 | 12.9 | 4.5 |
| prof | LDA | 84.4 | 84.2 | 36.9 | 4.4 |
| prof | CPCA | 84.5 | 37.9 | 10.8 | 14.9 |
| prof | COV | 84.5 | 37.9 | 11.1 | 15.0 |
| prof | LEACE | 84.4 | 38.0 | 36.9 | 13.8 |
| prof | BEST | 84.1 | 30.0 | 46.5 | 3.9 |

Table 3: Gender and profession spaces for BERT representations, finetuned on profession classification.

| Estimator | EXP. | TPR-Gap↓ | Purity↑ |
|---|---|---|---|
| Ambient | frozen | 0.174 | 0.9 |
| MLR | frozen | 0.150 | 13.5 |
| LDA | frozen | 0.124 | 26.6 |
| CPCA | frozen | 0.200 | 0.8 |
| COV | frozen | 0.201 | 0.8 |
| LEACE | frozen | 0.125 | 26.6 |
| Ambient | finetuned | 0.131 | 3.9 |
| MLR | finetuned | 0.128 | 12.9 |
| LDA | finetuned | 0.117 | 36.9 |
| CPCA | finetuned | 0.126 | 10.8 |
| COV | finetuned | 0.124 | 11.1 |
| LEACE | finetuned | 0.118 | 36.9 |

Table 4: Comparison between profession space purity and TPR-Gap for frozen and finetuned experiments. EXP. denotes the experimental setup.

| Concept | Estimator | Ret.↑ | Leak.↓ | Pur.↑ | Int.↓ |
|---|---|---|---|---|---|
| phones | WORST | 7.8 | 87.8 | 20.1 | 97.3 |
| phones | RAND | 57.9 | 87.6 | 92.5 | 22.5 |
| phones | MLR | 87.0 | 85.6 | 93.1 | 22.6 |
| phones | LDA | 82.5 | 86.0 | 94.6 | 22.0 |
| phones | CPCA | 76.6 | 29.6 | 94.6 | 21.1 |
| phones | COV | 76.6 | 29.6 | 94.8 | 21.1 |
| phones | LEACE | 82.6 | 29.6 | 94.3 | 21.2 |
| phone | BEST | 87.8 | 7.8 | 97.3 | 20.1 |
| speakers | WORST | 2.7 | 79.9 | 12.2 | 92.2 |
| speakers | RAND | 7.5 | 77.4 | 42.3 | 12.4 |
| speakers | MLR | 47.3 | 63.0 | 73.8 | 12.4 |
| speakers | LDA | 48.8 | 63.6 | 88.3 | 12.4 |
| speakers | CPCA | 18.1 | 54.3 | 38.7 | 12.5 |
| speakers | COV | 18.0 | 54.4 | 39.4 | 12.5 |
| speakers | LEACE | 48.9 | 54.4 | 88.3 | 12.5 |
| speakers | BEST | 79.9 | 2.7 | 92.2 | 12.2 |

Table 5: Phonetic and speaker subspaces for speech models.

| Trainset | Dims | Ret.↑ | Leak.↓ | Pur.↑ | Int.↓ |
|---|---|---|---|---|---|
| EXP. | - | 2.7 | 80.0 | 12.2 | 92.1 |
| standard | M=40 | 48.8 | 75.6 | 54.4 | 90.0 |
| overfit | M=40 | 54.5 | 2.7 | 12.4 | 12.4 |
| scale | M=40 | 60.9 | 39.7 | 89.0 | 12.4 |
| scale | M=100 | 70.2 | 23.8 | 83.9 | 12.6 |
| scale | M=200 | 73.8 | 15.4 | 73.5 | 13.2 |
| scale | M=300 | 75.6 | 11.3 | 60.5 | 14.0 |
| scale | M=400 | 76.7 | 8.7 | 44.9 | 15.2 |
| scale | M=500 | 77.4 | 6.9 | 30.2 | 17.1 |
| scale | M=600 | 78.3 | 5.3 | 19.0 | 21.3 |
| scale | M=700 | 79.0 | 3.7 | 13.9 | 36.6 |
| scale | M=768 | 80.0 | 2.7 | 12.2 | 92.1 |
| BEST | - | 80.0 | 2.7 | 92.1 | 12.2 |

Table 6: Generalization for speaker spaces, estimated using LEACE with different dimension restriction (M).

## B Numerical details for phone and speaker subspaces experiments

In this section, we provide detailed results for the investigations about phone and speaker subspaces in Section 5.

Table 5 comprises numerical results for Figure 4, contrasting performances of five different estimators on containment (retention and leakage) and disentanglement (purity and interference) metrics. We reported an average across 5 random spaces for RAND.

Table 6 shows details for Figure 6 and provides additional results for the dimensionality of 200, 400, 600, and 768. Note that M=768 is the dimensionality of the ambient space. Hence, its retention and leakage scores match the best case. However, purity and interference also reach the worst because there is nothing left in the complementary subspace (the dimensionality of the kernel is 0).