PREDICTING THE COLLATERAL COST OF LINEAR CONCEPT ERASURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear concept erasure removes a target attribute from a frozen representation and guarantees linear guardedness: the optimal affine least-squares predictor of that attribute can no longer improve on the constant predictor. The guarantee says nothing about what the removal costs other attributes, and that cost is normally established only after the fact, by retraining a probe for every retained attribute. We show that the corresponding loss in linear accessibility can instead be computed in closed form beforehand. For any linear map that guards the concept set, the post-erasure accessibility matrix of the retained concepts is bounded above by a Schur complement of the accessible task covariance A_0, whose diagonal entries are the linear R^2 of each concept and whose off-diagonal entries are the linearly accessible coupling between concept pairs, with equality for minimal-kernel erasers including LEACE. The Schur term is therefore the minimum collateral loss that guarding makes unavoidable, and for LEACE the leverage profile of the erased subspace additionally locates the edit. Across 69 frozen representations and four datasets, both closed forms hold to within numerical error. Erasing every concept of every dataset in turn, the accessible coupling read off A_0 predicts held-out AUROC degradation of retrained probes better than squared label correlation does, and screening all 455 CelebA triples from A_0 alone separates the cheapest erasures from the most damaging before any eraser is run.

1 INTRODUCTION

Some practical tasks require removing a single attribute from a frozen representation while leaving the rest of it usable. A protected attribute is removed before a downstream classifier is trained, so that a linear readout of that attribute is no longer available to it. A shortcut attribute is removed to test whether a model was relying on it. Throughout, a concept means one binary attribute of an input, supplied as a label column $Y_j \in \{0, 1\}$, and linearly accessible means recoverable from the representation by a linear map. The standard way to perform the removal is linear erasure: iterative nullspace projection (Ravfogel et al., 2020) repeatedly trains a linear classifier for the concept and projects the representation onto that classifier's nullspace, while LEACE (Belrose et al., 2023) replaces the iteration by a closed form and is the minimum-distortion affine map under which the optimal affine least-squares predictor of the target concept cannot improve on the constant predictor.

These methods guarantee linear guardedness in a second-order sense: after the edit, the optimal affine least-squares predictor of the target cannot improve on the constant predictor. They make no statement about the effect on the other attributes. That effect matters because the directions carrying the target also carry part of the linear signal for every attribute correlated with it, and it is what has to be weighed before the edit is applied. It is currently established only after the erasure has been performed, by running the eraser and retraining a probe for each remaining attribute of interest, and the natural predictor available beforehand is the correlation between the labels. Label correlation is a weak descriptor of that effect: it is a property of the annotation rather than of the representation, and Section 3 shows that it constrains that geometry only in the Loewner order, without determining or entrywise bounding the pairwise couplings that govern the damage.

This paper characterises that effect in advance, from second-order statistics computed before any eraser is run. We collect these quantities in the accessible task covariance A_0, whose diagonal entry for a concept is the linear R^2 of that concept and whose off-diagonal entry for a pair is the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

linearly accessible coupling between them. Linear guardedness is a constraint on the kernel of the erasing map (Holstege et al., 2025), and that constraint alone fixes an upper bound on what can remain: for any linear map that guards a concept set, the retained concepts are left at or below a Schur complement of A_0, with equality for the minimal guarding kernel and in particular for LEACE. The Schur term is therefore the minimum collateral loss that guarding makes unavoidable, not merely the cost of one eraser. LEACE is one such map, the one with minimum distortion (Belrose et al., 2023), and for it the per-coordinate leverage of the removed subspace additionally says where in the representation the edit falls. For scale, our 321-run evaluation required $6{,}615$ post-erasure logistic-probe fits. Once A_0 is computed, the corresponding retained losses in linear accessibility can instead be screened directly through Schur complements, without applying the eraser or retraining probes. The degradation of a regularised downstream probe is not fixed by the kernel and remains an empirical, eraser-dependent value, so we test it separately.

Contributions. **C1. Accessible task geometry.** We use the accessible task covariance A_0, whose off-diagonal carries the cross-concept coupling that a per-concept score cannot capture (Definition 1). **C2. The unavoidable collateral cost of linear guardedness.** For every linear map that guards a concept set S, retained accessibility is bounded above by $A_{TT} - A_{TS}A_{SS}^{\dagger}A_{ST}$ in the Loewner order, with equality for minimal-kernel maps including LEACE (Theorem 1), so the Schur term is the minimum loss that guarding requires and is governed by accessible coupling rather than by label correlation (Corollary 2). **C3. The LEACE realisation: edit rank and footprint.** For LEACE the squared Mahalanobis edit cost is $\mathrm{rank}(A_{SS})$ (Corollary 3) and the leverage profile of the removed subspace gives the per-coordinate footprint (Proposition 1). Both are realisation-specific rather than kernel invariants. **C4. Empirical downstream validation.** Across 69 frozen representations and four datasets the identities hold to near-machine precision on three datasets and to a mean absolute error of 9.9×10^{-6} on EUR-Lex, whose covariance is near-degenerate, and over 122 erased concepts the accessible coupling read off A_0 predicts held-out AUROC degradation of retrained probes better than squared label correlation does, on three of the four corpora. **C5. Screening candidate erasure sets.** Ranking all 455 CelebA triples from A_0 alone separates the cheapest erasures from the most damaging before any eraser is run (Section 5.3).

The four datasets are CelebA (Liu et al., 2015) for face attributes, Bias-in-Bios (De-Arteaga et al., 2019) for occupation prediction alongside a protected gender attribute, EUR-Lex (Chalkidis et al., 2019) for legal topic classification, and GoEmotions (Demszky et al., 2020) for emotion classification.

2 Related work

Linear concept erasure. Methods for removing a concept from a frozen representation by a linear map divide into two groups by how the removed subspace is obtained. Iterative methods estimate it by repeated supervised fitting: iterative nullspace projection (Ravfogel et al., 2020) trains a linear classifier for the concept, projects the representation onto that classifier's nullspace, and repeats until no classifier succeeds, so the removed subspace is the span of the classifiers that were fitted along the way. Closed-form methods obtain it from second-order statistics in one step: LEACE (Belrose et al., 2023) projects, in whitened coordinates, onto the orthogonal complement of the accessible subspace of the target set. The two differ in guarantee as well as in cost. Iterative projection removes whatever its classifiers found and gives no bound on the number of iterations or on the distortion. LEACE achieves linear guardedness at minimum mean squared edit distance, and by Theorem 4.2 of Belrose et al. (2023) that minimum is attained simultaneously for every positive-semidefinite inner product. The eraser is determined up to its action on the null space of Σ_Z, which carries no variance, so it is unique in distribution.

Both guarantees are stated for the target set only. Neither makes any statement about the concepts outside that set. The effect on the untargeted concepts is therefore established only after the erasure has been run, by retraining probes, and attributed beforehand to the correlation between the labels.

Same-kernel erasers and SPLINCE. Holstege et al. (2025) make the structure of this family explicit: linear guardedness is a constraint on the kernel of the erasing map, and LEACE, SAL and SPLINCE can share that kernel while choosing different ranges. They prove that same-kernel projections give identical predictions after unregularised refitting, and use the remaining freedom in the

range to design an eraser that preserves a specified task covariance, showing that the invariance need not hold under regularisation or a frozen readout. Our analysis complements their result. We identify the common retained value explicitly, as the Schur complement $A_{TT} - A_{TS}A_{SS}^{\dagger}A_{ST}$ computed from pre-erasure geometry, and obtain it jointly for the whole retained-concept accessibility matrix, which is what makes the cross-concept coupling visible. We then show that this same quantity upper-bounds retained accessibility over all linear guarding maps in the Loewner order (Theorem 1), not only over different ranges sharing one kernel, so it is the minimum unavoidable collateral loss rather than merely a shared value.

Naowarat et al. (2026) study concept representations through subspace structure and report that linear concept erasure can generalise imperfectly to unseen representations.

Scoring what a representation makes linearly available. A separate literature summarises a frozen representation by a scalar, either against one labelled task (Bao et al., 2019; You et al., 2021; Shao et al., 2022) or from the embedding spectrum alone (Garrido et al., 2023; Ghosh et al., 2022; Tsitsulin et al., 2023; Thilak et al., 2024), with no single score dominating across settings (Abou Baker & Handmann, 2024). $(A_0)_{jj}$ exactly recovers the binary H-score under the setup of Equation (1). Unlike these scalars, A_0 is a matrix, and its off-diagonal entries encode the cross-concept accessible coupling that governs the effect of an erasure on a retained concept, which no scalar per-task score contains. Appendix D compares against the label-free spectral scores, which by construction cannot respond to a change that moves the concept signal at fixed spectrum. Schur complements, subspace leverage (Mohri & Talwalkar, 2011) and mutual coherence (Tropp, 2004) are established tools. Here they are combined to determine the outcome of a linear concept-erasure intervention before the eraser is run.

3 Predicting Linear Erasure Outcomes

Proofs of every statement in this section are in Appendix A.

3.1 Accessible task covariance and LEACE

Let $Z \in \mathbb{R}^d$ be a centred frozen representation with $\mathbb{E}Z = 0$ and finite second moment, and $Y \in \{0,1\}^m$ a multi-label concept vector with prevalences $p_j = \Pr(Y_j = 1) \in (0,1)$, not assumed mutually exclusive. Write

$$\Sigma_Z = \mathrm{Cov}(Z), \qquad C = \mathrm{Cov}(Z,Y), \qquad D_\Gamma = \mathrm{diag}(\gamma_1,\dots,\gamma_m), \quad \gamma_j = p_j(1-p_j),$$

let $c_j = \mathrm{Cov}(Z,Y_j)$ be the jth column of C, let $A^{\dagger}$ denote the Moore–Penrose pseudo-inverse, and let $W = \Sigma_Z^{\dagger/2}$ be the whitening operator. Everything below is computed from these second-order statistics: no higher moment of Z enters.

Definition 1 (Accessible task covariance)**.**

$$A_0 = D_\Gamma^{-1/2} C^\top \Sigma_Z^{\dagger} C\, D_\Gamma^{-1/2} = B_0^\top B_0, \qquad B_0 = \Sigma_Z^{\dagger/2} C\, D_\Gamma^{-1/2} = W C\, D_\Gamma^{-1/2} \in \mathbb{R}^{d\times m}.$$

Where an estimator has to be regularised we write A_λ for the ridge form, which replaces $\Sigma_Z^{\dagger}$ by $(\Sigma_Z + \lambda I)^{-1}\Sigma_Z(\Sigma_Z + \lambda I)^{-1}$.

The diagonal is the linear R^2 of each concept. The mean-square optimal linear predictor of Y_j from Z has squared multiple correlation $c_j^\top \Sigma_Z^{\dagger} c_j/\gamma_j$, which also equals the binary H-score $H_j = \mathrm{tr}(\Sigma_Z^{\dagger}\Sigma_b^{(j)})$ of Bao et al. (2019) with $\Sigma_b^{(j)} = \mathrm{Cov}(\mathbb{E}[Z \mid Y_j])$:

$$(A_0)_{jj} = \frac{c_j^\top \Sigma_Z^{\dagger} c_j}{\gamma_j} = R_j^2 = H_j \in [0,1], \tag{1}$$

and we call this the accessibility of concept j. The off-diagonal $(A_0)_{ij}$ is the linearly accessible coupling between concepts i and j, and it is the entry that carries every result below.

Behaviour of the plug-in estimator near $n \approx d$. The unregularised plug-in attains the bound $(A_0)_{jj} \leq 1$ implied by Equation (2): whenever $n \leq d$ it returns exactly 1.0000 even for labels statistically independent of the features, because the sample covariance is rank-deficient and the regression on the spanned subspace is an interpolation. Ridge suppresses this without removing it. A split-block estimator that combines $\hat{\Sigma}_Z$ from one sample with $\hat{C}$ and $\hat{\Gamma}$ from another does not inherit the bound, because these blocks no longer form a single positive semidefinite joint sample covariance, and the estimate is then unbounded above. All pairs used here sit at $n/d \geq 42$, the smallest ratio over the 117 pairs. Appendix C gives the sweep.

Write $B_{0,S}$ for the columns of B_0 indexed by an erased set $S \subseteq [m]$, $T = [m] \setminus S$ for the retained concepts, and A_{SS}, A_{ST}, A_{TS}, A_{TT} for the corresponding blocks of A_0.

Lemma 1 (The erased subspace)**.** *In whitened coordinates the LEACE eraser r_S is the orthogonal projection onto the complement of $\mathcal{S} = \operatorname{colsp}(B_{0,S}) = \operatorname{colsp}(WC_S)$, and its kernel is $\operatorname{colsp}(C_S)$, the smallest kernel any linear map guarding S can have. Since $\operatorname{colsp}(C) \subseteq \operatorname{colsp}(\Sigma_Z)$, whitening is injective on $\operatorname{colsp}(C_S)$, so*

$$\dim \mathcal{S} = \operatorname{rank}(C_S) = \operatorname{rank}(B_{0,S}) = \operatorname{rank}(A_{SS}) =: r.$$

The left singular vectors of $B_{0,S}$ furnish an orthonormal basis of $\mathcal{S}$.

Terminology. We call the loss in accessibility over T caused by erasing S the collateral damage, the squared Mahalanobis norm $\mathbb{E}\|Z - r_S(Z)\|^2_{\Sigma_Z^\dagger}$ of the edit the edit cost, and the coordinatewise distribution of that cost the edit footprint. Where two erasures are compared we also report the erasure depth, the loss in accessibility on the erased concepts themselves, so that targets removed to a similar degree can be placed side by side.

The label-correlation constraint. Positive semidefiniteness of the joint covariance of (Z, Y) gives

$$0 \preceq A_0 \preceq R^Y, \qquad R^Y = D_\Gamma^{-1/2} \Gamma D_\Gamma^{-1/2}, \quad \Gamma = \operatorname{Cov}(Y), \tag{2}$$

and in particular $0 \leq (A_0)_{jj} \leq 1$. The bound is in the Loewner order: R^Y upper-bounds A_0 as a positive-semidefinite matrix. The label-correlation matrix therefore provides a representation-independent Loewner constraint on A_0, but it neither determines nor entrywise bounds the off-diagonal accessible couplings, which are the quantities that govern collateral damage in Theorem 1.

3.2 The unavoidable collateral cost of guarding

Say that a linear map P guards S when $PC_S = 0$, so that the optimal affine least-squares predictor of Y_S from PZ cannot improve on the constant predictor (Belrose et al., 2023). This is a second-order criterion: it constrains covariances, not the full conditional law, and Section 6 records what it leaves untouched. LEACE is one such map, but the following statement is about all of them.

Projector intuition. Write b_j for the jth column of B_0, so that $(A_0)_{ij} = b_i^\top b_j$ and $(A_0)_{jj} = \|b_j\|_2^2$. In whitened coordinates, minimal guarding removes the subspace $\mathcal{S} = \operatorname{colsp}(B_{0,S})$. If P_S is the orthogonal projector onto $\mathcal{S}$, a retained concept direction decomposes as

$$b_j = P_S b_j + (I - P_S) b_j.$$

The first component is necessarily lost when S is erased, so its squared length $\|P_S b_j\|_2^2$ is the collateral loss in accessibility of concept j. Collecting these projections over all retained concepts gives

$$B_{0,T}^\top P_S B_{0,T} = A_{TS} A_{SS}^\dagger A_{ST},$$

which is exactly the Schur term below.

Theorem 1 (Optimal retained accessibility under linear guarding)**.** *Let P be any linear map with $PC_S = 0$. Then, in the Loewner order,*

$$A_0(PZ, Y_T) \;\preceq\; A_{TT} - A_{TS} A_{SS}^\dagger A_{ST},$$

equivalently

$$A_{TT} - A_0(PZ, Y_T) \;\succeq\; A_{TS} A_{SS}^\dagger A_{ST}.$$

Equality holds whenever the effective erased kernel $\ker(P) \cap \mathrm{colsp}(\Sigma_Z)$ *equals the minimal guarding kernel* $\mathrm{colsp}(C_S)$*, and in particular for LEACE, whose post-erasure accessibility we write* $A^{\mathrm{post}} = A_0(r_S(Z), Y)$*:*

$$A_{TT}^{\mathrm{post}} = A_{TT} - A_{TS}A_{SS}^{\dagger}A_{ST}. \tag{3}$$

The Schur residual is thus the maximum retained accessibility compatible with guarding, and the Schur term $A_{TS}A_{SS}^{\dagger}A_{ST}$ is the minimum unavoidable collateral loss that guarding S imposes. Both are computed from pre-erasure geometry alone. Every minimal-kernel guarding map shares the value in Equation (3), independently of its range, by the same-kernel invariance of Holstege et al. (2025). What Theorem 1 adds is this explicit matrix-valued common value and the fact that it upper-bounds every guarding map. The exact equality condition is Theorem 5 in Appendix A.

Screening candidate erasure sets. Compute A_0 once for the full concept set. For any candidate erasure set S, the retained accessibility follows from the Schur residual $A_{TT}^{\mathrm{post}} = A_{TT} - A_{TS}A_{SS}^{\dagger}A_{ST}$ without applying the eraser or retraining downstream probes. Evaluating a new candidate set requires only extracting the corresponding blocks of A_0 and forming a small Schur complement in concept space, rather than applying an eraser and retraining a probe for every retained concept, so exhaustive or constrained screening of candidate sets is feasible before any representation-level intervention is run. The quantity screened is the predicted loss in linear accessibility. How a regularised downstream probe degrades is not fixed by the kernel and is tested separately in Section 5.2, and Section 5.3 checks the screening itself against real erasures.

Two consequences of Equation (3) are worth stating separately, because they are what the experiments use. Both are for the minimal-kernel case.

Corollary 1 (Single-concept collateral loss)**.** *For a single erased concept* j *with* $(A_0)_{jj} > 0$ *and a retained concept* j'*,*

$$(A_0)_{j'j'} - A_{j'j'}^{\mathrm{post}} = \frac{(A_0)_{j'j}^2}{(A_0)_{jj}}.$$

The loss to a retained concept is set by its accessible coupling to what is removed, discounted by the accessibility of the removed concept itself, since the denominator is $(A_0)_{jj}$ and not $(A_0)_{j'j'}$. The coupling that appears is the one in A_0 and not the one in R^Y, so two concept pairs with the same label correlation can suffer very different losses in two representations. Normalising by $\tilde{A}_0 = D_A^{-1/2}A_0D_A^{-1/2}$ with $D_A = \mathrm{diag}(A_0)$, and writing $\mu_{\mathrm{task}} = \max_{i \neq j} |(\tilde{A}_0)_{ij}|$, the same statement becomes a relative loss with a single-scalar bound.

Corollary 2 (Normalised collateral fraction)**.** *Assume* $(A_0)_{jj} > 0$ *for all* j*. For single-concept erasure the relative accessibility lost by* j' *when* j *is erased is exactly*

$$\frac{(A_0)_{j'j'} - A_{j'j'}^{\mathrm{post}}}{(A_0)_{j'j'}} = (\tilde{A}_0)_{j'j}^2 \;\leq\; \mu_{\mathrm{task}}^2.$$

For a single erased concept, $(\tilde{A}_0)_{j'j}^2$ is the squared cosine between the two accessible directions $b_{j'}$ and b_j in whitened coordinates.

3.3 Edit cost and edit footprint: the LEACE realisation

Theorem 1 fixes how much linear signal guarding must lose, but not how a particular realisation, meaning a particular map with that kernel, pays the cost. This subsection considers LEACE, which Belrose et al. (2023) show is the minimum-distortion guarding map. Let $P_S = U_S U_S^{\top}$ be the orthogonal projector onto $\mathcal{S}$, with $U_S \in \mathbb{R}^{d \times r}$ the left singular vectors of $B_{0,S}$.

Corollary 3 (Edit cost)**.** *For the LEACE map* r_S*,*

$$\mathbb{E}\big\|r_S(Z) - Z\big\|_{\Sigma_Z^{\dagger}}^2 = \mathrm{rank}(B_{0,S}) = \mathrm{rank}(A_{SS}) = r,$$

which equals $|S|$ *exactly when* $\mathrm{rank}(C_S) = |S|$*. Any other guarding map with the same kernel has whitened edit energy at least* r *(Proposition 2).*

Proposition 1 (Footprint equals leverage). *For the LEACE map, the leverage of whitened coordinate i is $\ell_i = (P_S)_{ii} \in [0, 1]$, and the whitened edit $\tilde{\delta} = P_S \tilde{Z}$ satisfies*

$$\mathbb{E}[\tilde{\delta}_i^2] = \ell_i, \qquad \sum_{i=1}^{d} \ell_i = r,$$

so $\{\ell_i\}$ is the exact per-coordinate decomposition of the edit cost of Corollary 3.

To summarise a footprint by one number we use the erasure coherence $\mu_{\mathrm{er}} = (d/r) \max_i \ell_i \in [1, d/r]$, for which

$$d/\mu_{\mathrm{er}} \in [r, d] \tag{4}$$

counts the effective number of edited coordinates in the whitened basis, not in raw neurons or in the original representation dimensions. It equals r for a concentrated edit falling on exactly r coordinates and d for a distributed one with uniform leverage.

Both quantities in this subsection are properties of the LEACE realisation and not kernel invariants: under an oblique map with the same kernel the coordinatewise energies can exceed one and sum to more than r, so they are no longer leverage scores (Proposition 2). The Schur term of Theorem 1 says how much linear accessibility guarding must destroy. The footprint says where in the representation the LEACE realisation pays that cost.

Limits of invariance. Theorem 1 concerns unregularised linear refitting. If two surviving representations are related by an invertible but non-orthogonal map F taking the first to the second, their unregularised linear hypothesis classes coincide, but an isotropic penalty does not transform isotropically: $\lambda\|\theta\|_2^2$ in the second becomes $\lambda\, \eta^\top (F^\top F)^{-1} \eta$ in the first, with $\eta = F^\top \theta$. Regularised downstream performance can therefore depend on the eraser's range, and so can the behaviour of a frozen readout. It follows that the ℓ_2-regularised logistic probes of Section 5.2 are not algebraically determined by Theorem 1: they constitute a separate empirical test of whether accessible coupling predicts downstream collateral degradation for the LEACE realisation used here.

4 Experimental setup

Datasets. We evaluate on four multi-label datasets, chosen to vary the number of concepts m, the per-concept prevalence, and the label-correlation structure among concepts, which provides the Loewner-order reference matrix in Equation (2). Concept counts run from $m = 15$ to $m = 50$ and median prevalences from 0.018 to 0.384. Table 2 in Appendix B gives the per-dataset counts, split sizes and prevalence range. CelebA (Liu et al., 2015) carries $m = 15$ face attributes, all but `Smiling` in four clusters coupled through known mechanisms, which makes its erasure results interpretable rather than only numerical: gender presentation, facial hair, age and appearance, and a set of gender-correlated controls with their own visual signature (Table 20). Bias-in-Bios (De-Arteaga et al., 2019) pairs a protected attribute with a downstream task, 28 profession indicators plus a binary `gender` column, so erasing `gender` is the fairness reading of collateral damage. EUR-Lex (Chalkidis et al., 2019) and GoEmotions (Demszky et al., 2020) are the sparse-label, high-m cases. EUR-Lex is truncated to 128 tokens, leaving shared boilerplate on which every encoder is more anisotropic, so its geometry-derived quantities carry the caveat in Appendix D.

Representations. We evaluate 45 vision backbones and 24 text encoders, 69 representations in total, giving 117 representation $\times$ dataset pairs. The vision backbones span 11 pretraining families, from language-supervised (CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023)) through self-distilled (DINO (Caron et al., 2021), DINOv2 (Oquab et al., 2023)) and masked-autoencoding (MAE (He et al., 2022)) to dense-prediction (SAM (Kirillov et al., 2023), SAM 2 (Ravi et al., 2025)). The text encoders range from raw masked-language-model controls (BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019)) to dedicated embedding models (E5 (Wang et al., 2022), BGE (Xiao et al., 2024), Sentence-BERT (Reimers & Gurevych, 2019)). Complete checkpoints, extraction settings and model-level results are in Appendices B and F.

Protocol. No network is fine-tuned: representations are extracted once and frozen, so every number compared on a given dataset comes from identical embeddings. LEACE is fit on the training

split. Algebraic identities are evaluated on that fitting split, while the fixed train-fitted eraser is applied to validation for downstream probe evaluation. Appendix B states which split is used at each stage. After each erasure, probes are retrained from scratch on the erased representation. These probes are ℓ_2-regularised logistic regressions at fixed $C = 1.0$. Nine erasure sets were run across the four datasets, 321 runs in total. Three further panels support Sections 5.2 and 5.3: an expanded single-concept panel erasing every concept of every corpus on ten representations per corpus (122 targets, 1,220 runs), a reduced panel varying the probe family, and a CelebA screening panel of ten triples selected from the pre-erasure ranking before any validation erasure was run, on ten representations.

Scope of the numerical claims. We distinguish exact algebraic identities from empirical downstream consequences. The identities of Theorem 1, Corollary 3, and Proposition 1 are evaluated by numerical error, reported as mean and worst-case absolute deviation. Because the same empirical covariance is used to construct the eraser and to evaluate A^{post}, these are exact finite-sample identity checks on the fitting split rather than held-out measurements. The non-algebraic claim, that A_0 predicts what an erasure will cost in practice, is evaluated on probes retrained after the erasure, primarily through the AUROC drop and secondarily through the balanced-accuracy drop. Those probes are ℓ_2-regularised, and by the argument at the end of Section 3.3 their behaviour is not algebraically determined by Theorem 1.

5 Results

5.1 Representation-specific variation in collateral cost

Variation across representation–dataset settings. The cost of guarding varies widely across settings: for the two gender erasures, mean retained R^2 loss is 0.1523 on CelebA against 0.0046 on Bias-in-Bios at comparable target erasure depth, a factor of roughly 33 (Table 10). The comparison is descriptive rather than causal. The closed form does not explain why the two settings differ, but it identifies where the difference sits: in the off-diagonal block A_{TS}. On CelebA the gendered appearance attributes share much of their linear signal with `Male`. The settings also differ in dataset and modality, so this does not attribute the gap; it shows that the gap is visible in A_0 before erasure.

The same asymmetry is visible in the coupling structure. On CelebA the accessible coupling exceeds the label correlation in magnitude on 98.7% of concept pairs (Table 6). The largest couplings, μ_{task} and $\mu_{\mathrm{label}} = \max_{i \neq j} |R^Y_{ij}|$ ($0.97/0.80$ on CelebA, $0.66/0.21$ on GoEmotions), order the datasets almost identically, so the difference is entrywise, which a label-only audit does not reveal.

Erasure depth alone does not predict collateral cost: among the three multi-concept CelebA edits, `age_appearance` removes the least of its own target (0.411) yet causes the greatest damage to the retained concepts (0.165), while `gender_makeup` removes the most (0.753) and causes the least (0.106). On GoEmotions both erasures are shallow and nearly free, and on EUR-Lex a four-concept cluster is removed to depth 0.8578 for 0.0274 of collateral damage. How much signal an eraser removes from its target is therefore not a substitute for the Schur residual.

Localisation of the edit. The effective number of edited coordinates d/μ_{er} separates erasures that act on a handful of whitened directions from erasures that are spread across the representation. At the concentrated end, `Wearing_Lipstick`, `Male` and `Heavy_Makeup` on CelebA concentrate their edits on 2.3 to 2.4 effective coordinates against a mean whitened dimension of 946, and Bias-in-Bios `gender` on 4.2. At the distributed end the rare GoEmotions categories `pride` and `embarrassment` spread over roughly 36 coordinates of a mean whitened dimension of 783. Both numbers are available from one thin singular value decomposition before the eraser runs, and they answer a different question from the collateral damage: not how much the edit costs, but where in the representation it falls. Figure 8 in Appendix E.8 plots the two profiles.

Scope	rows	ρ coupling	ρ label corr.	$\Delta\rho$	95% CI
Pooled	42280	0.459	0.212	$+0.247$	$[+0.201, +0.294]$
CelebA	2100	0.770	0.743	$+0.027$	$[-0.077, +0.109]$
Bias-in-Bios	8120	0.435	0.106	$+0.329$	$[+0.167, +0.481]$
EUR-Lex	24500	0.325	0.219	$+0.106$	$[+0.064, +0.146]$
GoEmotions	7560	0.566	0.100	$+0.466$	$[+0.370, +0.568]$

Table 1: Expanded single-concept panel: accessible coupling $(\tilde{A}_0)^2_{ij}$ against squared label correlation $(R^Y_{ij})^2$ as predictors of the held-out AUROC drop of probes retrained after erasure, over all 122 erased concepts and 40 representations. Paired bootstraps with 10,000 resamples, stratified by corpus in the pooled row. Every interval shown is the two-way one, which resamples erased targets and representations together and is the widest of the three schemes. Tables 13 and 15 give the target-cluster and representation-cluster intervals, Table 14 the balanced-accuracy version. $\Delta\rho > 0$ favours accessible coupling.

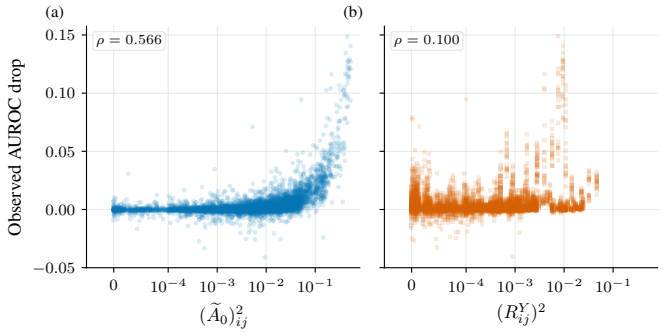

Figure 1: GoEmotions: pre-erasure predictors of held-out AUROC drop. Each point is one (retained concept, erased concept, representation) triple. Both horizontal axes are symmetric-log with a linear segment at zero. **(a)**: normalised accessible coupling $(\tilde{A}_0)^2_{ij}$ against the held-out AUROC drop of a probe retrained after erasure. **(b)**: the same drop against squared label correlation $(R^Y_{ij})^2$. The rank association is substantially stronger for accessible coupling ($\rho = 0.566$) than for squared label correlation ($\rho = 0.100$). Negative drops are cases in which held-out AUROC improves after erasure.

5.2 Pre-erasure accessible coupling as a predictor of held-out AUROC degradation

We compare two pre-erasure predictors of the damage to a retained concept j when concept i is erased: the accessible coupling $(\tilde{A}_0)^2_{ij}$, which is exactly the relative loss of Corollary 2, and the squared label correlation $(R^Y_{ij})^2$. Only single-concept erasures are scored, since multi-concept sets need an aggregation the theory does not dictate (Appendix E). Every concept of every corpus is erased in turn on ten representations per corpus: 122 targets over 40 representations, 1,220 runs and 42,280 retained-concept rows.

Accessible coupling is the stronger predictor of held-out AUROC degradation: pooled it reaches Spearman 0.459 against 0.212, a difference of $+0.247$ whose interval excludes zero under every resampling scheme (Table 1). The advantage is largest where label correlation is weakest, $+0.466$ on GoEmotions against $+0.027$ on CelebA, the one corpus left unresolved, and scoring each concept on its own it leads on 115 of 122 targets (Figure 1 shows the GoEmotions triples). The balanced-accuracy comparison is weaker throughout, at a pooled $+0.098$ (Appendix E.4).

Neither the probe family nor the unregularised estimator drives the result. On a reduced panel the margin stays between $+0.230$ and $+0.242$ across logistic regression at $C \in \{0.1, 1, 10\}$ and a linear support vector machine, and replacing A_0 by the ridge accessibility A_λ moves the pooled margin to $+0.284$, while the dependence on the predictor ridge does not mirror the regularisation of the downstream probe (Appendix E.5).

Comparison on the R^2 drop. A third measure of collateral damage is the drop in A_0 recomputed on the erased representation. Corollary 2 makes $(\tilde{A}_0)_{ij}^2$ the relative loss, the fraction of concept j's own accessibility that erasing i removes, so scoring it against an absolute R^2 drop discards the factor $(A_0)_{jj}$. Matched to the relative loss, the agreement is the identity of Corollary 2 and holds to a mean absolute error of 1.7×10^{-13} over the $42{,}280$ rows, which certifies the implementation and is not evidence for the predictor. When compared with the absolute R^2 loss, the same accessible-coupling score still leads pooled at 0.908 against 0.721, and on every corpus except CelebA (Table 19). The retrained-probe metrics above remain the part of the comparison that the identities do not determine.

5.3 Screening candidate erasure sets

Section 3 claims that candidate sets can be ranked from A_0 alone. We test that on CelebA by scoring all $\binom{15}{3} = 455$ triples by predicted retained loss, fixing the ranking before any validation erasure was run, and then erasing the five lowest-scoring and five highest-scoring sets on ten representations each. Triples match the size of the multi-concept CelebA interventions in Table 9, holding set cardinality fixed while varying its composition. Once A_0 is available, scoring one candidate triple takes 0.10 ms. A complete held-out validation run for one triple takes $3{,}339$ s in the measured implementation, of which $3{,}215$ s is spent fitting the twelve retained-concept probes (Appendix E.6).

The ranking survives the test. The five sets predicted to be cheapest lose 0.0245 of held-out retained AUROC on average against 0.1289 for the five predicted to be most damaging, a factor of 5.26, and the predicted order agrees with the measured one at Spearman 0.794 (Table 18). The sample is ten interventions, so this establishes that the ordering is usable, not that the predicted magnitudes are calibrated.

5.4 Numerical verification of the closed forms

We ran LEACE on all 117 representation $\times$ dataset pairs with nine erasure sets, 321 runs in total. Over $5{,}880$ retained-concept rows, one per concept $\times$ representation $\times$ erasure set, the Schur residual of Theorem 1 reproduces the measured post-erasure R^2 to near-machine precision on three datasets, with a worst case below 5×10^{-13}. Over the $3{,}564$ footprints, one per representation and erased set with every concept also profiled on its own (so single-concept sets appear twice), the leverage profile of Proposition 1 reproduces the per-coordinate mean squared edit to a worst case of 4.8×10^{-13}, and the squared Mahalanobis norm of the edit equals $|S|$ in all 321 runs, which is Corollary 3 when $\operatorname{rank}(C_S) = |S|$. EUR-Lex is the one loose case, at mean absolute error 9.9×10^{-6}, its covariance being near-degenerate under 128-token truncation (Appendix D). Table 11 gives the per-dataset errors.

6 Discussion and Limitations

Both the size and location of collateral loss can be computed before erasure. The Schur term gives the minimum loss in linear accessibility imposed by guarding, attained by LEACE, while the LEACE leverage profile localises the edit. These costs vary across representations through A_0 and are not determined by label correlations. For held-out probes, accessible coupling better predicts AUROC degradation over 122 erased concepts, though the advantage is smaller for balanced accuracy.

Limitations. The framework is linear. Nonlinear concept signal is invisible to A_0 and survives LEACE. On XOR, accessibility is 9.7×10^{-4} while a multi-layer perceptron reaches AUROC 1.00. The unregularised plug-in estimator is unreliable near $n \approx d$ and can saturate even for independent labels. Empirically, all erasures use LEACE although Theorem 1 holds for every linear guarding map. Edit rank and footprint are LEACE-specific (Appendix A.9). The expanded panel uses ten representations per corpus rather than all 69, and on CelebA its two-way interval includes zero. Finally, the exact checks are fitting-split identities computed from the covariance defining the eraser. Transport to new samples is a separate question (Appendix A.10) on which Naowarat et al. (2026) report imperfect generalisation.

AI USE STATEMENT

Generative AI agents were used to assist with writing and revising code, and to review the manuscript for spelling, grammar, and presentation. All AI-assisted code and manuscript changes were reviewed by the authors, who take responsibility for the final content of the paper and associated artifacts.

ETHICS STATEMENT

This work studies the removal of labelled attributes from frozen representations, including gender-related attributes in CelebA and Bias-in-Bios. We collect no new human-subject data and use existing benchmark datasets for the empirical analyses. Linear guardedness is a second-order property and should not be interpreted as demographic fairness, privacy, or removal of sensitive information against nonlinear predictors. Although concept-erasure methods can be used to audit or modify representations, their use in consequential applications requires separate evaluation of downstream harms, dataset-specific biases, and the information retained by nonlinear predictors. We therefore do not interpret successful linear erasure as establishing fairness or privacy.

REPRODUCIBILITY STATEMENT

The paper separates the information needed to reproduce the theoretical and empirical results across the main text and appendices. Appendix A gives complete proofs and assumptions. Appendix B specifies the datasets, data splits, model checkpoints, representation extraction, and evaluation protocol. Appendix C documents estimator behaviour and controlled diagnostics. Appendices D and E provide the representation diagnostics, robustness analyses, bootstrap schemes, probe configurations, screening procedure, and numerical verification of the closed forms. Appendix F reports the model inventories and per-representation results. The main text also distinguishes fitting-split algebraic identity checks from held-out downstream evaluation.

REFERENCES

Nermeen Abou Baker and Uwe Handmann. One size does not fit all in evaluating model selection scores for image classification. *Scientific Reports*, 14(1):30239, 2024.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE international conference on image processing (ICIP)*, pp. 2309–2313. IEEE, 2019.

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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6314–6322, 2019.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128, 2019.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4040–4054, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of*

the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pp. 4171–4186, 2019.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, pp. 10929–10974. PMLR, 2023.

Arna Ghosh, Arnab Kumar Mondal, Kumar Krishna Agrawal, and Blake Richards. Investigating power laws in deep representation learning. *arXiv preprint arXiv:2202.05808*, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Floris Holstege, Shauli Ravfogel, and Bram Wouters. Preserving task-relevant information under linear concept removal. In *Advances in Neural Information Processing Systems*, volume 38, pp. 30563–30603, 2025. doi: 10.52202/085713-0907.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Mikhail Kuznetsov, Ivan Butakov, Marina Munkhoeva, Alexey Frolov, and Ivan Oseledets. Information-theoretic unsupervised embedding quality evaluation. In *ICLR 2026 Workshop on Representational Alignment (Re-Align)*, 2026. URL `https://openreview.net/forum?id=EwaNoLlEXT`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Mehryar Mohri and Ameet Talwalkar. Can matrix coherence be efficiently and accurately estimated? In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 534–542. JMLR Workshop and Conference Proceedings, 2011.

Burin Naowarat, Hao Tang, and Sharon Goldwater. A framework for analyzing concept representations in neural models. In *Proceedings of the 30th Conference on Computational Natural Language Learning*, pp. 574–587, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7237–7256, 2020.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Wenqi Shao, Xun Zhao, Yixiao Ge, Zhaoyang Zhang, Lei Yang, Xiaogang Wang, Ying Shan, and Ping Luo. Not all models are equal: Predicting model transferability in a self-challenging fisher space. In *European Conference on Computer Vision*, pp. 286–302. Springer, 2022.

Vimal Thilak, Chen Huang, Omid Saremi, Laurent Dinh, Hanlin Goh, Preetum Nakkiran, Joshua Susskind, and Etai Littwin. LiDAR: Sensing linear probing performance in joint embedding SSL architectures. In *International Conference on Learning Representations*, volume 2024, pp. 56726–56765, 2024.

Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004.

Anton Tsitsulin, Marina Munkhoeva, and Bryan Perozzi. Unsupervised embedding quality evaluation. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pp. 169–188. PMLR, 2023.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 641–649, 2024.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *International conference on machine learning*, pp. 12133–12143. Pmlr, 2021.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

A PROOFS

Every result stated in Section 3 is proved here.

A.1 THE SCHUR-COMPLEMENT BOUND

Theorem 2 (Label-correlation Loewner constraint)**.** *With notation as above,*

$$0 \preceq C^\top \Sigma_Z^\dagger C \preceq \Gamma,$$

and consequently

$$0 \preceq A_0 \preceq R^Y, \quad \textit{where } R^Y = D_\Gamma^{-1/2} \Gamma D_\Gamma^{-1/2} \textit{ is the label correlation matrix.}$$

In particular, every diagonal entry satisfies $0 \le (A_0)_{jj} \le 1$.

Proof. The block matrix

$$M = \begin{pmatrix} \Sigma_Z & C \\ C^\top & \Gamma \end{pmatrix} \succeq 0$$

because it is the covariance of (Z, Y). The required range condition $\operatorname{colsp}(C) \subseteq \operatorname{colsp}(\Sigma_Z)$ holds automatically: any $v \perp \operatorname{colsp}(\Sigma_Z)$ satisfies $\operatorname{Var}(v^\top Z) = v^\top \Sigma_Z v = 0$, so $v^\top Z$ is constant and $v^\top C = \operatorname{Cov}(v^\top Z, Y) = 0$, hence $v \perp \operatorname{colsp}(C)$. The generalised Schur complement of the $(1,1)$-block is therefore well-defined and gives $\Gamma - C^\top \Sigma_Z^\dagger C \succeq 0$. The lower bound follows from $\Sigma_Z^\dagger \succeq 0$. Conjugating by $D_\Gamma^{-1/2}$ preserves the order, and since R^Y is a correlation matrix its diagonal is 1. □

For a single concept j the bound specialises to $0 \le R_j^2 \le 1$. For the off-diagonals, positive semidefiniteness of A_0 together with $(A_0)_{jj} \le 1$ gives, by Cauchy–Schwarz, $|(A_0)_{ij}| \le \sqrt{(A_0)_{ii}(A_0)_{jj}} \le 1$. The corresponding ridge bound $0 \preceq A_\lambda \preceq A_0 \preceq R^Y$ follows from spectral shrinkage.

A.2 THE DIAGONAL IS THE BINARY H-SCORE

Theorem (H-score is per-task accessibility; Equation (1))**.** *Let* $Y_j \in \{0, 1\}$ *be a single binary task with prevalence* p_j*, and let* $H_j = \operatorname{tr}(\Sigma_Z^\dagger \Sigma_b^{(j)})$ *denote its H-score, where* $\Sigma_b^{(j)} = \operatorname{Cov}(\mathbb{E}[Z \mid Y_j])$*. Then*

$$H_j = (A_0)_{jj} = R_j^2.$$

Proof. For binary Y_j, $\mathbb{E}[Z \mid Y_j = c] = \mu_{j,c}$, so the random variable $\mathbb{E}[Z \mid Y_j]$ takes value $\mu_{j,1}$ with probability p_j and $\mu_{j,0}$ with probability $1 - p_j$. Its covariance is

$$\Sigma_b^{(j)} = p_j(1 - p_j)(\mu_{j,1} - \mu_{j,0})(\mu_{j,1} - \mu_{j,0})^\top = \gamma_j \Delta_j \Delta_j^\top.$$

Therefore

$$H_j = \operatorname{tr}\big(\Sigma_Z^\dagger \gamma_j \Delta_j \Delta_j^\top\big) = \gamma_j \, \Delta_j^\top \Sigma_Z^\dagger \Delta_j.$$

Using $c_j = \gamma_j \Delta_j$,

$$\frac{c_j^\top \Sigma_Z^\dagger c_j}{\gamma_j} = \gamma_j \Delta_j^\top \Sigma_Z^\dagger \Delta_j = H_j.$$

The left side is precisely $(A_0)_{jj}$ by Definition 1. □

Here $\mu_{j,c} = \mathbb{E}[Z \mid Y_j = c]$, $\Delta_j = \mu_{j,1} - \mu_{j,0}$ and $c_j = \operatorname{Cov}(Z, Y_j) = \gamma_j \Delta_j$, as in Section 3.

A.3 FACTORISATION AND THE ERASED SUBSPACE

Lemma (Factorisations). *For every $\lambda > 0$, $A_\lambda = B_\lambda^\top B_\lambda$ with $B_\lambda = \Sigma_Z^{1/2}(\Sigma_Z + \lambda I)^{-1} C D_\Gamma^{-1/2}$, and $A_0 = B_0^\top B_0$ with $B_0 = \Sigma_Z^{\dagger/2} C D_\Gamma^{-1/2}$. In particular, the nonzero eigenvalues of A_λ (resp. A_0) are the squared singular values of B_λ (resp. B_0).*

Proof. Using the spectral identity

$$(\Sigma_Z + \lambda I)^{-1}\Sigma_Z(\Sigma_Z + \lambda I)^{-1} = \left[\Sigma_Z^{1/2}(\Sigma_Z + \lambda I)^{-1}\right]^2,$$

we obtain

$$A_\lambda = D_\Gamma^{-1/2} C^\top \left[\Sigma_Z^{1/2}(\Sigma_Z + \lambda I)^{-1}\right]^2 C\, D_\Gamma^{-1/2} = B_\lambda^\top B_\lambda.$$

Setting $\lambda = 0$ gives $A_0 = (\Sigma_Z^{\dagger/2} C D_\Gamma^{-1/2})^\top (\Sigma_Z^{\dagger/2} C D_\Gamma^{-1/2}) = B_0^\top B_0$. The eigenvalue claim follows from the singular value decomposition. □

Lemma (Lemma 1, restated). *Let $\mathcal{S} = \operatorname{colsp}(WC) \subseteq \mathbb{R}^d$ in whitened activation coordinates. Then:*

1. $\dim \mathcal{S} = \operatorname{rank}(WC) = \operatorname{rank}(C) = \operatorname{rank}(A_\lambda) = \operatorname{rank}(A_0)$ *for any finite λ.*
2. *In whitened coordinates, the LEACE eraser is the orthogonal projection onto $\mathcal{S}^\perp$.*
3. *$\mathcal{S} = \operatorname{colsp}(B_0)$, and the* left *singular vectors of B_0 furnish an orthonormal basis for $\mathcal{S} \subseteq \mathbb{R}^d$. The* right *singular vectors of B_0 live in $\mathbb{R}^m$ and describe task mixtures; they are the eigenvectors of $A_0 = B_0^\top B_0$.*
4. *For $\lambda > 0$, $\operatorname{colsp}(B_\lambda)$ is generally tilted relative to $\operatorname{colsp}(B_0)$, because B_λ differs from B_0 by the spectral shrinker $T_\lambda = \Sigma_Z(\Sigma_Z + \lambda I)^{-1}$ on whitened coordinates. The two subspaces coincide precisely when $\operatorname{colsp}(B_0)$ is invariant under this shrinker. This includes the special case in which $\operatorname{colsp}(B_0)$ is spanned by eigenvectors of Σ_Z. Generically, however, a mixed subspace is tilted.*

Proof. (1) $\operatorname{rank}(WC) = \operatorname{rank}(C)$ because W is full rank on $\operatorname{colsp}(\Sigma_Z) \supseteq \operatorname{colsp}(C)$. The equality with $\operatorname{rank}(A_\lambda)$ follows from the factorisation lemma: multiplying by the invertible $D_\Gamma^{-1/2}$ preserves rank, and $\Sigma_Z^{1/2}(\Sigma_Z + \lambda I)^{-1}$ is invertible on $\operatorname{colsp}(\Sigma_Z)$ for finite λ.

(2) Writing $\tilde{z} = W(z - \mathbb{E}Z)$, the eraser sends $\tilde{z} \mapsto (I - P_{WC})\tilde{z}$, an orthogonal projection onto $\operatorname{colsp}(WC)^\perp = \mathcal{S}^\perp$.

(3) $B_0 = WC\, D_\Gamma^{-1/2}$, and $D_\Gamma^{-1/2}$ is an invertible right factor, so $\operatorname{colsp}(B_0) = \operatorname{colsp}(WC) = \mathcal{S}$. If $B_0 = U\Sigma V^\top$ is its compact SVD with rank r, then $U \in \mathbb{R}^{d\times r}$ spans $\operatorname{colsp}(B_0) = \mathcal{S}$, while $V \in \mathbb{R}^{m\times r}$ consists of eigenvectors of $B_0^\top B_0 = A_0$.

(4) In whitened coordinates, $B_\lambda = \Sigma_Z^{1/2}(\Sigma_Z + \lambda I)^{-1} C D_\Gamma^{-1/2} = T_\lambda\, \Sigma_Z^{\dagger/2} C D_\Gamma^{-1/2} = T_\lambda B_0$ on $\operatorname{colsp}(\Sigma_Z)$. Hence $\operatorname{colsp}(B_\lambda) = T_\lambda \operatorname{colsp}(B_0)$, which equals $\operatorname{colsp}(B_0)$ precisely when $\operatorname{colsp}(B_0)$ is invariant under T_λ. Since T_λ is diagonal in the eigenbasis of Σ_Z, eigenvector-spanned accessible subspaces are invariant, generic mixed subspaces not.

(Kernel.) For an erased subset S, a linear map P guards S only if $\operatorname{Cov}(PZ, Y_S) = PC_S = 0$, so $\operatorname{colsp}(C_S) \subseteq \ker P$ for every guarding map. The linear part of LEACE in the original coordinates is $I - W^\dagger P_{\mathcal{S}} W$ with $\mathcal{S} = \operatorname{colsp}(WC_S)$. If $v \in \operatorname{colsp}(C_S) \subseteq \operatorname{colsp}(\Sigma_Z)$, then $Wv \in \mathcal{S}$ and $W^\dagger W v = v$, so v is in the kernel. Conversely, $v = W^\dagger P_{\mathcal{S}} W v$ places v in $W^\dagger\mathcal{S} = \operatorname{colsp}(W^\dagger WC_S) = \operatorname{colsp}(C_S)$. The kernel of LEACE is therefore exactly $\operatorname{colsp}(C_S)$, the smallest kernel a guarding map can have. □

Lemma 2 (Ridge tilt). *For $\lambda > 0$, $\operatorname{colsp}(B_\lambda)$ is generally tilted relative to $\operatorname{colsp}(B_0)$, because B_λ differs from B_0 by the spectral shrinker $T_\lambda = \Sigma_Z(\Sigma_Z + \lambda I)^{-1}$ on whitened coordinates. The two subspaces coincide precisely when $\operatorname{colsp}(B_0)$ is invariant under this shrinker, which includes the case in which $\operatorname{colsp}(B_0)$ is spanned by eigenvectors of Σ_Z; a generic mixed subspace is tilted. The tilt is small when λ is much smaller than the smallest nonzero eigenvalue of Σ_Z, and the exact identity is recovered in the $\lambda \to 0$ limit. This is why the subspace statements of Lemma 1 are stated for A_0 and not for A_λ.*

Proof. In whitened coordinates, $B_\lambda = \Sigma_Z^{1/2}(\Sigma_Z + \lambda I)^{-1} C D_\Gamma^{-1/2} = T_\lambda \Sigma_Z^{\dagger/2} C D_\Gamma^{-1/2} = T_\lambda B_0$ on $\operatorname{colsp}(\Sigma_Z)$, hence $\operatorname{colsp}(B_\lambda) = T_\lambda \operatorname{colsp}(B_0)$, which equals $\operatorname{colsp}(B_0)$ precisely when $\operatorname{colsp}(B_0)$ is T_λ-invariant. Since T_λ is diagonal in the eigenbasis of Σ_Z, eigenvector-spanned accessible subspaces are invariant and generic mixed subspaces are not. □

A.4 THE DUALITY THEOREM

Throughout this subsection the erased set is all of $[m]$. The subset versions used in Corollary 3 follow by restricting to the columns of B_0 indexed by S.

Theorem 3 (Duality of accessibility and erasability)**.** *Let Z be a centered embedding, allowing singular Σ_Z by restriction to $\operatorname{colsp}(\Sigma_Z)$. Let $B_0 = \Sigma_Z^{\dagger/2} C D_\Gamma^{-1/2}$ and let $\mathcal{S} = \operatorname{colsp}(B_0) \subseteq \mathbb{R}^d$ in whitened activation coordinates. Then:*

1. ***Same subspace, exact for A_0.*** *The LEACE eraser, in whitened coordinates, is exactly the orthogonal projection onto $\mathcal{S}^\perp$. The left singular vectors of B_0 form an orthonormal basis of $\mathcal{S}$; its right singular vectors are the eigenvectors of A_0.*
2. ***Edit cost in Mahalanobis distance.*** *The squared LEACE edit cost in the Mahalanobis inner product $M = \Sigma_Z^\dagger$ equals the dimension of the accessible activation subspace:*

$$\mathbb{E}\big\|r(Z) - Z\big\|_{\Sigma_Z^\dagger}^2 = \operatorname{rank}(C) = \operatorname{rank}(A_0) = \operatorname{rank}(A_\lambda)$$

   for every finite λ.
3. ***Total R^2 removed.*** *Every linear-probe R^2-quality vanishes after LEACE: $R_j^2(r(Z), Y_j) = 0$ for every j. Before LEACE, $R_j^2(Z, Y_j) = (A_0)_{jj}$. Summed over j, the exact total linear R^2 removed is $\sum_j (A_0)_{jj} = \operatorname{tr}(A_0)$.*
4. ***Ridge distinction.*** *For $\lambda > 0$, $\operatorname{colsp}(B_\lambda)$ is generically tilted relative to $\mathcal{S}$, so A_λ identifies the LEACE-removed activation subspace only in the limit $\lambda \to 0$. The hard-rank and Mahalanobis edit-cost statements remain valid for all finite λ. The exact total linear R^2 removed remains $\operatorname{tr}(A_0)$; $\operatorname{tr}(A_\lambda) \leq \operatorname{tr}(A_0)$ is a ridge-shrunk surrogate.*

Proof. (1) is the subspace identity of Lemma 1, proved above.

(2) Let $\tilde{z} = W(z - \mathbb{E}Z)$ so $\operatorname{Cov}(\tilde{Z}) = \Pi$, the orthogonal projector onto $\operatorname{colsp}(\Sigma_Z)$. The LEACE edit in original coordinates is $\delta = W^\dagger P_{WC} \tilde{Z}$. Its Mahalanobis squared norm is $\|\delta\|_{\Sigma_Z^\dagger}^2 = \tilde{Z}^\top P_{WC}^\top (W^\dagger)^\top \Sigma_Z^\dagger W^\dagger P_{WC} \tilde{Z}$. On $\operatorname{colsp}(\Sigma_Z)$, $W = W^\top = \Sigma_Z^{\dagger/2}$, $W^\dagger = \Sigma_Z^{1/2}$ and $\Sigma_Z^\dagger = W^\top W$, so $(W^\dagger)^\top \Sigma_Z^\dagger W^\dagger = \Sigma_Z^{1/2} \Sigma_Z^\dagger \Sigma_Z^{1/2} = \Pi$. Combined with $P_{WC} \Pi P_{WC} = P_{WC}$, which holds because $\operatorname{colsp}(WC) \subseteq \operatorname{colsp}(\Sigma_Z)$, this gives $\|\delta\|_{\Sigma_Z^\dagger}^2 = \tilde{Z}^\top P_{WC} \tilde{Z}$. Taking expectations, $\mathbb{E}\|\delta\|_{\Sigma_Z^\dagger}^2 = \operatorname{tr}(P_{WC}\Pi) = \operatorname{tr}(P_{WC}) = \dim \mathcal{S} = \operatorname{rank}(C)$, which equals $\operatorname{rank}(A_0) = \operatorname{rank}(A_\lambda)$ by the rank identity of Lemma 1.

(3) Linear guardedness from Belrose et al. (2023) gives $\operatorname{Cov}(r(Z), Y) = 0$, equivalent to zero linear predictability of every Y_j from $r(Z)$. Before erasure, $R_j^2(Z, Y_j) = (A_0)_{jj}$ by Definition 1. Summing over j gives $\operatorname{tr}(A_0)$.

(4) is the ridge-tilt statement of Lemma 2 plus the observation that hard rank is preserved by finite ridge shrinkage on $\operatorname{colsp}(\Sigma_Z)$, whereas trace is spectrally shrunk. □

Corollary 3 is item (2) of Theorem 3 restricted to the columns of B_0 in S, using $\operatorname{rank}(B_{0,S}) = \operatorname{rank}(A_{SS})$ from Lemma 1.

A.5 THE MINIMAL-KERNEL SCHUR RESIDUAL

The equality half of Theorem 1 is the case the experiments use. The universal bound and the exact equality condition are proved in Appendix A.8, as Corollary 4 and Theorem 5.

Theorem (Equality case of Theorem 1)**.** *Partition tasks into an erased set S and a retained set T. A guarding map whose effective erased kernel is the minimal one,* $\mathrm{colsp}(C_S)$*, and in particular LEACE, leaves $A^{\mathrm{post}}_{TT} = A_{TT} - A_{TS}A^{\dagger}_{SS}A_{ST}$, the Schur complement of the S-block of A_0; for a single erased task j, $A^{\mathrm{post}}_{j'j'} = (A_0)_{j'j'} - (A_0)^2_{j'j}/(A_0)_{jj}$, which is Corollary 1.*

Proof. Let B_S be the columns of B_0 in S, $P_S = B_S A^{\dagger}_{SS} B_S^{\top}$ the projector onto $\mathrm{colsp}(B_S)$ that LEACE removes. With $\bar{Y}_j = Y_j/\sqrt{\gamma_j}$, $\mathrm{Cov}(\tilde{Z}, \bar{Y}_j) = b_j$ and $(A_0)_{ij} = b_i^{\top} b_j$. After erasure $\tilde{Z}' = (I - P_S)\tilde{Z}$. Since $b_{j'} \in \mathrm{colsp}(\Sigma_Z)$, $\Pi b_{j'} = b_{j'}$ and $\mathrm{Cov}(\tilde{Z}', \bar{Y}_{j'}) = (I - P_S)b_{j'} = (\Pi - P_S)b_{j'}$. Since $P_S\Pi = \Pi P_S = P_S$, $\mathrm{Cov}(\tilde{Z}') = (I - P_S)\Pi(I - P_S) = \Pi - P_S$, an orthogonal projector and hence its own pseudo-inverse. The erased representation is $r(Z) - \mathbb{E}r(Z) = W^{\dagger}\tilde{Z}'$, and $W^{\dagger}$ is injective on $\mathrm{colsp}(\Sigma_Z)$, which contains $\tilde{Z}'$, so it does not change linear accessibility. Therefore $(A^{\mathrm{post}}_{TT})_{j'j''} = b_{j'}^{\top}(\Pi - P_S)(\Pi - P_S)^{\dagger}(\Pi - P_S)b_{j''} = b_{j'}^{\top}(\Pi - P_S)b_{j''} = (A_0)_{j'j''} - A_{j'S}A^{\dagger}_{SS}A_{Sj''}$, the last step because $P_S = B_S A^{\dagger}_{SS} B_S^{\top}$. The single-task case is $S = \{j\}$. □

Here $b_j = \Sigma_Z^{\dagger/2} c_j/\sqrt{\gamma_j}$ is the jth column of B_0, so that $(A_0)_{ij} = \langle b_i, b_j\rangle$ and $(A_0)_{jj} = \|b_j\|^2 = R_j^2$, and Π is the orthogonal projector onto $\mathrm{colsp}(\Sigma_Z)$ as above.

Corollary (Corollary 2, restated with its set form)**.** *Assume $(A_0)_{jj} > 0$ for all j. For single-task erasure, the relative accessibility lost by j' when j is erased is exactly*

$$\frac{(A_0)_{j'j'} - A^{\mathrm{post}}_{j'j'}}{(A_0)_{j'j'}} = (\tilde{A}_0)^2_{j'j} \;\le\; \mu^2_{\mathrm{task}}.$$

For a set S with $(|S| - 1)\mu_{\mathrm{task}} < 1$, the total relative collateral over T obeys $\mathrm{tr}(\tilde{A}_{TS}\tilde{A}^{\dagger}_{SS}\tilde{A}_{ST}) \le |S||T|\mu^2_{\mathrm{task}}/(1 - (|S| - 1)\mu_{\mathrm{task}})$.

Proof. The single-task identity is Theorem 1 divided by $(A_0)_{j'j'}$. The set version conjugates the Schur complement by $D_{A,T}^{-1/2}$; under $(|S|-1)\mu_{\mathrm{task}} < 1$, Gershgorin gives $\lambda_{\min}(\tilde{A}_{SS}) \ge 1 - (|S| - 1)\mu_{\mathrm{task}} > 0$, hence $\mathrm{tr}(\tilde{A}_{TS}\tilde{A}^{-1}_{SS}\tilde{A}_{ST}) \le \|\tilde{A}_{ST}\|_F^2/\lambda_{\min}(\tilde{A}_{SS})$ and $\|\tilde{A}_{ST}\|_F^2 \le |S||T|\mu^2_{\mathrm{task}}$. □

Both quantities are read off $B_0 = \Sigma_Z^{\dagger/2} C D_{\Gamma}^{-1/2}$ alone, with no eraser run: the leverage profile of the left factor is the edit footprint, its incoherence μ_{er} the delocalisation of the edit, and the mutual coherence μ_{task} of the columns an upper bound on the collateral damage. The relevant object is the *accessible* mutual coherence $\max_{i\neq j} |(\tilde{A}_0)_{ij}|$, not the raw label coherence $\max_{i\neq j} |R^Y_{ij}|$.

A.6 The footprint and its coherence

Proposition (Proposition 1, restated)**.** *The whitened LEACE edit $\tilde{\delta} = P_{\mathcal{S}}\tilde{Z}$ satisfies $\mathbb{E}[\tilde{\delta}_i^2] = \ell_i$ and $\sum_{i=1}^d \ell_i = r$. Moreover $\tilde{\delta} = W\delta$ for the original-space edit $\delta = Z - r(Z)$, so $\{\ell_i\}$ is the exact per-coordinate decomposition of the Mahalanobis edit cost $\mathbb{E}\|\delta\|^2_{\Sigma_Z^{\dagger}} = r$ of Theorem 3.*

Proof. From $\Pi P_{\mathcal{S}} = P_{\mathcal{S}}$, $\mathrm{Cov}(\tilde{\delta}) = P_{\mathcal{S}}\Pi P_{\mathcal{S}} = P_{\mathcal{S}}$, whose diagonal is ℓ_i and trace is r. For the identity, $\delta = \Sigma_Z^{1/2} P_{\mathcal{S}}\tilde{Z}$, hence $W\delta = \Pi P_{\mathcal{S}}\tilde{Z} = \tilde{\delta}$ and $\|\tilde{\delta}\|_2^2 = \delta^{\top}\Sigma_Z^{\dagger}\delta = \|\delta\|^2_{\Sigma_Z^{\dagger}}$; expectations give $\sum_i \ell_i = \mathbb{E}\|\delta\|^2_{\Sigma_Z^{\dagger}} = r$. □

Proposition 1 is exact in whitened (Mahalanobis) coordinates, and a whitened coordinate is a row of $\Sigma_Z^{\dagger/2}$, not a neuron. In raw neuron coordinates $\mathrm{Cov}(Z - r(Z)) = \Sigma_Z^{1/2} P_{\mathcal{S}} \Sigma_Z^{1/2}$, so the per-neuron edit energy is $(\Sigma_Z^{1/2} P_{\mathcal{S}} \Sigma_Z^{1/2})_{ii}$ and the conserved budget is the removed variance $\mathrm{tr}(P_{\mathcal{S}}\Sigma_Z)$ rather than the dimension r. The analysis below applies unchanged to those energies normalised by that budget.

Write $\hat{\ell}_i = \ell_i/r$, so $\sum_i \hat{\ell}_i = 1$, and recall $H_\alpha(\hat{\ell}) = \frac{1}{1-\alpha}\log\sum_i \hat{\ell}_i^{\alpha}$ with $H_0 = \log\|\hat{\ell}\|_0$, $H_1 = -\sum_i \hat{\ell}_i \log\hat{\ell}_i$ and $H_\infty = -\log\max_i \hat{\ell}_i$.

Definition (Erasure coherence)**.** $\mu_{\mathrm{er}} = \frac{d}{r}\max_{1\le i\le d}\ell_i \in [1, d/r]$, the standard coherence parameter (Mohri & Talwalkar, 2011) of $\mathcal{S}$.

Theorem (Coherence is the $\alpha=\infty$ effective footprint dimension; Equation (4))**.** $e^{H_\infty(\hat\ell)} = d/\mu_{\mathrm{er}}$, *and*

$$r \;\le\; \underbrace{\frac{d}{\mu_{\mathrm{er}}}}_{e^{H_\infty}} \;\le\; \underbrace{\frac{r^2}{\sum_i \ell_i^2}}_{e^{H_2}} \;\le\; e^{H_1(\hat\ell)} \;\le\; \underbrace{\|\operatorname{diag} P_{\mathcal{S}}\|_0}_{e^{H_0}} \;\le\; d.$$

Thus d/μ_{er} is the min-entropy effective number of edited coordinates: r for an axis-aligned (concentrated) edit, d for the uniformly delocalised edit $\ell_i \equiv r/d$.

Proof. $\max_i \hat\ell_i = \mu_{\mathrm{er}}/d$ by definition, so $e^{H_\infty} = d/\mu_{\mathrm{er}}$. The interior inequalities are monotonicity of H_α in α, with $e^{H_2} = r^2/\sum_i \ell_i^2$ and $e^{H_0} = \|\operatorname{diag} P_{\mathcal{S}}\|_0$. Since $\ell_i \le 1$, $\max_i \hat\ell_i \le 1/r$, so $e^{H_\infty} \ge r$; and $e^{H_0} \le d$. Axis-aligned $\mathcal{S}$ gives r unit entries and $d-r$ zeros, so $\mu_{\mathrm{er}} = d/r$ and $e^{H_\alpha} = r$. $\mu_{\mathrm{er}} = 1$ forces $\ell_i \equiv r/d$, giving $e^{H_\alpha} = d$. □

This chain is the spectral-metric ladder $\mathrm{NESum} \le d_2 \le \mathrm{erank} \le \mathrm{rank}$, the Rényi family at $\alpha = \infty, 2, 1, 0$, applied to the edit footprint $\hat\ell$ instead of to the eigenvalues of Σ_Z. Here erank is the effective rank of Roy & Vetterli (2007): for a positive semidefinite matrix with eigenvalues λ_i normalised to $\hat\lambda_i = \lambda_i/\sum_k \lambda_k$, $\mathrm{erank} = \exp\big(-\sum_i \hat\lambda_i \log \hat\lambda_i\big)$, the exponential of the Shannon entropy of the normalised spectrum, which is the $\alpha = 1$ member of the same family. NESum and d_2 are its $\alpha = \infty$ and $\alpha = 2$ counterparts, and Kuznetsov et al. (2026) study this family as unsupervised embedding-quality measures. The two ladders are independent: the spectral one describes how variance spreads over eigendirections of Σ_Z, the footprint one how the fixed budget r spreads over coordinates. Whitening flattens $\mathrm{Cov}(\tilde Z) = \Pi$, so an unconditional coherence is degenerate and only the label-selected subspace $\mathcal{S}$ carries a meaningful incoherence.

A.7 The Fisher coupling identity

Assume every within-class covariance $\Sigma_W^{(j)}$ below is invertible, and $S_j > 0$. For task j, define the Fisher signal-to-noise ratio $S_j = \Delta_j^\top (\Sigma_W^{(j)})^{-1}\Delta_j$ with $\Sigma_W^{(j)} = p_j\Sigma_{j,1} + (1-p_j)\Sigma_{j,0}$ the within-class covariance, the asymmetric cross-Mahalanobis form $M_{ij} = \Delta_i^\top (\Sigma_W^{(i)})^{-1}\Delta_j$, the optimal Fisher direction $w_j^\star = (\Sigma_W^{(j)})^{-1}\Delta_j$, and the logit $L_j^\star = (w_j^\star)^\top Z$.

Theorem (Heteroskedastic Fisher cross-correlation)**.**

$$\mathrm{Corr}(L_i^\star, L_j^\star)^2 = \frac{M_{ij}M_{ji}}{S_iS_j}.$$

Proof. By the binary law of total covariance, $\Sigma_Z = \Sigma_W^{(j)} + \gamma_j\Delta_j\Delta_j^\top$. Hence $\mathrm{Cov}(L_i^\star, L_j^\star) = (w_i^\star)^\top\Sigma_W^{(j)}w_j^\star + \gamma_j(w_i^\star)^\top\Delta_j\Delta_j^\top w_j^\star$. Using $\Sigma_W^{(j)}w_j^\star = \Delta_j$ and the definition of M_{ij}, $\mathrm{Cov}(L_i^\star, L_j^\star) = M_{ij}(1+\gamma_jS_j)$. Symmetric expansion in task i gives $\mathrm{Cov}(L_i^\star, L_j^\star) = M_{ji}(1+\gamma_iS_i)$. The variance is $\mathrm{Var}(L_j^\star) = S_j(1+\gamma_jS_j)$, so

$$\mathrm{Corr}(L_i^\star, L_j^\star)^2 = \frac{M_{ij}M_{ji}(1+\gamma_iS_i)(1+\gamma_jS_j)}{S_i(1+\gamma_iS_i)S_j(1+\gamma_jS_j)} = \frac{M_{ij}M_{ji}}{S_iS_j}.$$

□

In the homoskedastic case ($\Sigma_W^{(0)} = \Sigma_W^{(1)} =: \Sigma_W$) the binary law of total covariance gives $\Sigma_Z = \Sigma_W + \gamma_j\Delta_j\Delta_j^\top$, and Sherman–Morrison yields $\Delta_j^\top\Sigma_Z^{-1}\Delta_j = S_j/(1+\gamma_jS_j)$ with $S_j = \Delta_j^\top\Sigma_W^{-1}\Delta_j$. Thus $R_j^2 = \gamma_jS_j/(1+\gamma_jS_j)$, and inverting gives the prevalence correction $S_j = R_j^2/(\gamma_j(1-R_j^2))$ of Equation (5), monotone in R_j^2.

A.8 Proof of Theorem 1 for arbitrary linear maps

This subsection proves Theorem 1 in full generality: arbitrary linear maps rather than idempotent projections, singular Σ_Z, rank-deficient C_S, and Moore–Penrose pseudo-inverses throughout. Write Π for the orthogonal projector onto $\mathcal{E} = \operatorname{colsp}(\Sigma_Z)$, and for a linear map $P : \mathbb{R}^d \to \mathbb{R}^k$ define the *effective kernel* and its whitened image

$$\mathcal{K}(P) = \ker(P) \cap \operatorname{colsp}(\Sigma_Z), \qquad \widetilde{\mathcal{K}}(P) = \Sigma_Z^{\dagger/2}\,\mathcal{K}(P).$$

Directions of $\ker(P)$ outside $\operatorname{colsp}(\Sigma_Z)$ carry no variance, so they are invisible in distribution. Accessibility therefore depends on P only through $\mathcal{K}(P)$.

Lemma 3 (Accessibility as the covariance of optimal linear predictions)**.** *Let $G = PZ$ and let $\widehat{Y}_j$ be the population least-squares prediction of Y_j from G, that is, the minimiser of $\mathbb{E}(Y_j - a - b^\top G)^2$ over a and b. Then*

$$A_0(G, Y_T)_{ij} = \frac{\operatorname{Cov}(\widehat{Y}_i, \widehat{Y}_j)}{\sqrt{\gamma_i \gamma_j}}.$$

The diagonal is the optimal linear R^2 and the off-diagonal entries are normalised covariances between optimal fitted values, which is what makes A_0 the natural matrix-valued object rather than a collection of per-concept scores.

Proof. The population least-squares prediction of Y_j from G is $\widehat{Y}_j = \mathbb{E}Y_j + \operatorname{Cov}(G, Y_j)^\top \Sigma_G^\dagger (G - \mathbb{E}G)$. Taking the covariance of two such predictions and using $\operatorname{Cov}(G) = \Sigma_G$,

$$\operatorname{Cov}(\widehat{Y}_i, \widehat{Y}_j) = \operatorname{Cov}(G, Y_i)^\top \Sigma_G^\dagger \Sigma_G \Sigma_G^\dagger \operatorname{Cov}(G, Y_j) = \operatorname{Cov}(G, Y_i)^\top \Sigma_G^\dagger \operatorname{Cov}(G, Y_j),$$

the second equality being the Moore–Penrose identity $\Sigma_G^\dagger \Sigma_G \Sigma_G^\dagger = \Sigma_G^\dagger$. Dividing by $\sqrt{\gamma_i \gamma_j}$ gives $A_0(G, Y_T)_{ij}$ as defined in Definition 1, and the diagonal case $i = j$ is the optimal linear R^2. □

Theorem 4 (Accessibility is kernel-determined)**.** *For every linear map $P : \mathbb{R}^d \to \mathbb{R}^k$, square or rectangular, idempotent or not,*

$$A_0(PZ, Y_T) = B_{0,T}^\top \big(\Pi - P_{\widetilde{\mathcal{K}}(P)}\big) B_{0,T},$$

where $P_{\widetilde{\mathcal{K}}(P)}$ is the orthogonal projector onto $\widetilde{\mathcal{K}}(P)$. Consequently $A_0(PZ, Y_T)$ depends on P only through $\mathcal{K}(P)$: two maps with the same effective kernel leave the same unregularised retained accessibility.

Proof. By Lemma 3, $A_0(PZ, Y_T)$ is the Gram matrix, in the whitened metric, of the components of the retained task directions that survive P. Write $\tilde{Z} = \Sigma_Z^{\dagger/2} Z$ and $Q = P\Sigma_Z^{1/2}$, so that PZ and $Q\tilde{Z}$ span the same linear predictions and therefore carry the same accessible geometry. Least-squares prediction of $\bar{Y}_j = Y_j/\sqrt{\gamma_j}$ from $Q\tilde{Z}$ has fitted-value covariance $B_{0,T}^\top Q^\top (QQ^\top)^\dagger Q B_{0,T}$, and $Q^\top (QQ^\top)^\dagger Q$ is the orthogonal projector onto $(\ker Q)^\perp$. Restricted to $\mathcal{E}$, the kernel of Q is $\Sigma_Z^{\dagger/2}\mathcal{K}(P) = \widetilde{\mathcal{K}}(P)$, so that projector equals $\Pi - P_{\widetilde{\mathcal{K}}(P)}$ and the covariance of the fitted values is $B_{0,T}^\top(\Pi - P_{\widetilde{\mathcal{K}}(P)})B_{0,T}$. Rank deficiency of C_S and singularity of Σ_Z are absorbed by the pseudo-inverse, because $\operatorname{colsp}(C) \subseteq \operatorname{colsp}(\Sigma_Z)$ makes $\Sigma_Z^{\dagger/2}$ injective on $\operatorname{colsp}(C)$. □

Corollary 4 (Theorem 1)**.** *If $PC_S = 0$ then $\operatorname{colsp}(C_S) \subseteq \ker(P)$, so $\mathcal{S} = \operatorname{colsp}(WC_S) \subseteq \widetilde{\mathcal{K}}(P)$ and $P_{\widetilde{\mathcal{K}}(P)} \succeq P_{\mathcal{S}}$. Substituting into Theorem 4,*

$$A_0(PZ, Y_T) = B_{0,T}^\top \big(\Pi - P_{\widetilde{\mathcal{K}}(P)}\big) B_{0,T} \preceq B_{0,T}^\top \big(\Pi - P_{\mathcal{S}}\big) B_{0,T} = A_{TT} - A_{TS} A_{SS}^\dagger A_{ST},$$

the last step being the standard identity $P_{\mathcal{S}} = B_{0,S} A_{SS}^\dagger B_{0,S}^\top$. Equality holds when $\widetilde{\mathcal{K}}(P) = \mathcal{S}$, which is the minimal guarding kernel, and in particular for LEACE, whose kernel is $\operatorname{colsp}(C_S)$ by Lemma 1.

Theorem 5 (Exact equality condition)**.** *Let P guard S, write $\widetilde{\mathcal{K}} = \widetilde{\mathcal{K}}(P)$, and let*

$$\Delta = \widetilde{\mathcal{K}} \ominus \mathcal{S}, \qquad \mathcal{B}_T = \mathrm{colsp}\big((\Pi - P_{\mathcal{S}})B_{0,T}\big)$$

be the surplus erased subspace and the retained accessible subspace that survives minimal erasure, where $\ominus$ denotes the orthogonal complement of $\mathcal{S}$ inside $\widetilde{\mathcal{K}}$. Then

$$A_0(PZ, Y_T) = A_{TT} - A_{TS}A_{SS}^{\dagger}A_{ST} \quad \Longleftrightarrow \quad \Delta \perp \mathcal{B}_T.$$

Proof. $P_{\widetilde{\mathcal{K}}} = P_{\mathcal{S}} + P_{\Delta}$ since the two are orthogonal by construction, so the deficit relative to the Schur residual is $B_{0,T}^{\top}P_{\Delta}B_{0,T}$, which vanishes precisely when Δ is orthogonal to $\mathrm{colsp}((\Pi - P_{\mathcal{S}})B_{0,T}) = \mathcal{B}_T$. □

A minimal kernel is therefore sufficient for equality but not necessary. A map that erases more than it must incurs no extra collateral loss whenever the surplus directions carry no retained accessible signal.

Proposition 2 (Edit energy of an oblique same-kernel realisation)**.** *For a square guarding map P let $\tilde{E} = \Sigma_Z^{\dagger/2}(I - P)\Sigma_Z^{1/2}$. In an orthonormal basis adapted to $\mathcal{S} \oplus (\mathcal{S}^{\perp} \cap \mathcal{E})$,*

$$\tilde{E} = \begin{pmatrix} I_r & N' \\ 0 & M \end{pmatrix}, \qquad \mathbb{E}\|\tilde{\delta}\|_2^2 = r + \|N'\|_F^2 + \|M\|_F^2 \;\geq\; r, \qquad \mathrm{Cov}(\tilde{\delta}) = \tilde{E}\tilde{E}^{\top} \succeq P_{\mathcal{S}}.$$

Equality holds exactly when the map agrees on $\mathrm{colsp}(\Sigma_Z)$ with the LEACE edit, which is the minimum-distortion realisation of Belrose et al. (2023); null-space maps that differ only outside $\mathrm{colsp}(\Sigma_Z)$ still tie in distribution. Under an oblique same-kernel map the coordinatewise energies $\mathrm{Cov}(\tilde{\delta})_{ii}$ can exceed one and sum to more than r, so they are no longer leverage scores and Proposition 1 does not transfer.

Proof. Guarding gives $PC_S = 0$, hence $(I - P)C_S = C_S$, so in whitened coordinates $\tilde{E}v = v$ for every $v \in \mathcal{S}$. In a basis adapted to $\mathcal{S} \oplus (\mathcal{S}^{\perp} \cap \mathcal{E})$ this fixes both left-hand blocks at once: the $\mathcal{S}$-component of $\tilde{E}|_{\mathcal{S}}$ is the identity, giving I_r, and its $\mathcal{S}^{\perp}$-component vanishes, giving the zero block below. The blocks N' and M describe the action on $\mathcal{S}^{\perp} \cap \mathcal{E}$, are unconstrained by guarding, and are what distinguishes one realisation from another. Since $\mathrm{Cov}(\tilde{Z}) = \Pi$, $\mathbb{E}\|\tilde{\delta}\|_2^2 = \|\tilde{E}\|_F^2 = r + \|N'\|_F^2 + \|M\|_F^2 \geq r$. Moreover $\mathrm{Cov}(\tilde{\delta}) = \tilde{E}\tilde{E}^{\top}$, and in the same basis

$$\tilde{E}\tilde{E}^{\top} - P_{\mathcal{S}} = \begin{pmatrix} N'N'^{\top} & N'M^{\top} \\ MN'^{\top} & MM^{\top} \end{pmatrix} = \begin{pmatrix} N' \\ M \end{pmatrix}\begin{pmatrix} N' \\ M \end{pmatrix}^{\top} \succeq 0,$$

so $\mathrm{Cov}(\tilde{\delta}) \succeq P_{\mathcal{S}}$. Equality in both forces $N' = 0$ and $M = 0$, which is the LEACE edit on $\mathrm{colsp}(\Sigma_Z)$. □

A.9 Regimes beyond unregularised same-distribution refitting

The guarding kernel determines which linear signal can survive. The range determines how the surviving representation is positioned. Two maps with the same minimal guarding kernel therefore leave identical retained accessibility while differing in how far they displace the representation, in their coordinatewise footprint, in how a frozen downstream head behaves, and in regularised downstream performance.

Theorem 1 is a statement about unregularised linear refitting under one distribution. Following the analysis of Holstege et al. (2025), the invariance is exact for an unregularised linear predictor refitted under the same distribution. It fails for a ridge predictor and for a frozen linear readout, where the Schur diagonal remains an upper bound but the range matters. Under distribution shift the ordinary Schur identity does not transfer in general, and the invariance holds only through the evaluation-relative effective kernel of Proposition 3.

For the regularised case, if two surviving representations satisfy $G_B = FG_A$ then

$$\min_{\theta} L(G_B^{\top}\theta) + \lambda\|\theta\|_2^2 = \min_{\eta} L(G_A^{\top}\eta) + \lambda\,\eta^{\top}(F^{\top}F)^{-1}\eta,$$

so an isotropic penalty in one range is anisotropic in the other unless F is conformal. The regularised optimum can therefore differ in direction once the surviving predictive space has dimension at least two, and AUROC need not be range-invariant, and the direction of the change is not determined in general. For a frozen linear readout θ, range invariance fails unless $(P_A - P_B)^\top \theta = 0$, though the Schur diagonal remains an upper bound on the squared correlation any linear score can achieve after guarding under the same distribution.

A.10 Transported accessibility under distribution shift

The identity of Theorem 1 is a one-distribution statement. A map fitted under distribution A and evaluated under B obeys a transported formula instead, which we record because it governs how our fitting-split checks relate to held-out evaluation.

Proposition 3 (Transported accessibility)**.** *Let P be a fixed linear map fitted under A, with kernel $K_A = \ker(P)$, and evaluate under B. Put $K_{A\to B} = K_A \cap \mathrm{colsp}(\Sigma_B)$ and let G_B have columns spanning $\Sigma_B^{\dagger/2} K_{A\to B}$. Then*

$$A_0^{(B)}(PZ, Y_T) = B_{0,T}^{(B)\top}\big(\Pi_B - P_{\mathrm{colsp}(G_B)}\big) B_{0,T}^{(B)} = A_{TT}^{(B)} - M_{TS} M_{SS}^{\dagger} M_{TS}^{\top},$$

with $M_{SS} = G_B^\top G_B$ and $M_{TS} = B_{0,T}^{(B)\top} G_B$. The map guards under B if and only if $\mathrm{colsp}(C_S^{(B)}) \subseteq K_A$.

Proof. Apply Theorem 4 with Σ_B in place of Σ_Z. The effective kernel of P relative to B is $K_{A\to B} = K_A \cap \mathrm{colsp}(\Sigma_B)$, whose whitened image under $\Sigma_B^{\dagger/2}$ is $\mathrm{colsp}(G_B)$, so the surviving accessibility is $B_{0,T}^{(B)\top}(\Pi_B - P_{\mathrm{colsp}(G_B)})B_{0,T}^{(B)}$. Writing $P_{\mathrm{colsp}(G_B)} = G_B(G_B^\top G_B)^\dagger G_B^\top$ and substituting $M_{SS} = G_B^\top G_B$, $M_{TS} = B_{0,T}^{(B)\top} G_B$ gives the second form. The guarding condition is Corollary 4 read under B: P annihilates $C_S^{(B)}$ exactly when $\mathrm{colsp}(C_S^{(B)}) \subseteq K_A$. □

The ordinary B-Schur complement therefore does not apply to a train-fitted eraser evaluated on a different sample. Separating genuine out-of-sample recoverability from the finite-sample bias of plug-in A_0 would need permutation and split-block controls that we have not run.

B Experimental protocol and provenance

Dataset	Train	Valid	Test	m	$p_{\min}$	p_{med}	$p_{\max}$	Repr.
CelebA	162,770	19,867	19,962	15	0.0228	0.3843	0.8342	45
Bias-in-Bios	257,478	39,642	99,069	29	0.0035	0.0177	0.4610	24
EUR-Lex	45,000	6,000	6,000	50	0.0176	0.0245	0.0810	24
GoEmotions	43,410	5,426	5,427	28	0.0018	0.0299	0.3276	24

Table 2: Datasets. m is the number of binary concepts and $p_\bullet$ the fraction of training examples labelled positive for a concept, reported as the minimum, median and maximum over the m concepts. "Repr." is the number of cached representations.

Extraction. Each cached pair holds the pooled embedding matrix ($n \times d$, float32), the binary label matrix ($n \times m$, int8), sample identifiers and task names. Vision backbones are loaded through `timm`, each under its own preprocessing configuration, with input resolutions from 224 to 1024px: token outputs are mean-pooled over the sequence and feature maps are global-average-pooled. Two entries load an external checkpoint into a bare architecture rather than a pretrained tag, `mocov3-vitb` and `simclr-rn50`, the latter with its projection head discarded. Text encoders use one uniform protocol: final layer, mean pooling under the attention mask, and a maximum sequence length of 128 tokens. The limit is set by extraction cost, which grows with sequence length over the 117 representation × dataset pairs, so one window is used for every encoder. EUR-Lex documents are much longer than this window, and its estimates are correspondingly the least reliable in the set (Section 5.4).

Probes and estimators. Probes are ℓ_2-regularised logistic regressions at fixed inverse regularisation strength $C = 1.0$ with a fixed seed. The fit counts are as follows. The nine erasure sets of Table 9 require 6,615 post-erasure fits, of which 5,880 are retained concepts and enter the Schur comparison. The expanded single-concept panel adds 1,220 runs and 42,280 retained-concept fits, the screening panel adds 100 runs, and the probe-family panel repeats a reduced design under four probe settings. Baseline probes add 3,243 fits, reused across the erasure sets of a given representation.

The canonical logistic solver reaches its configured iteration limit on Bias-in-Bios without converging. We retain the fixed configuration so that every representation is scored identically rather than tuning it per corpus, and the reduced-panel linear support vector machine of Table 16, which does not share that solver, gives the same qualitative ordering.

The panels of Sections 5.2 and 5.3 use subsets of the full grid, fixed before any of their erasures were run. For the expanded single-concept panel, ten representations per corpus were selected by a deterministic, outcome-independent rule: cached models were grouped by family, the registry family for vision and the Hugging Face organisation prefix for text, the groups and their members were ordered by name, and the first model from each of the first ten groups was retained. The panel then erases every concept of each corpus in turn. The reduced probe-family panel uses the first three entries of each corpus's stored ten-representation subset and eight erased concepts at indices $\mathrm{round}(\mathrm{linspace}(0, m-1, 8))$ in the corpus's canonical concept order, so the first and last concepts are always included and the intermediate ones are spaced as evenly as integer indices allow. Both reduced-panel choices depend only on pre-existing list order and concept count, not on any erasure or probe outcome. The screening panel enumerates all $\binom{15}{3} = 455$ CelebA triples, scores each by the predicted retained loss, writes the ranking to disk, and only then erases the five lowest-scoring and five highest-scoring sets. The ten CelebA representations of the expanded and screening panels are `aimv2_1b`, `clip-rn50`, `dino-vitb16`, `dinov2-vitb14`, `dinov3-convnext_base`, `hiera_mae`, `mae-base`, `mocov3-vitb`, `resnet-resnet50` and `simclr-rn50`. The ten text encoders, the same on all three text datasets, are `gte-base-en-v1.5`, `bge-base-en-v1.5`, `snowflake-arctic-embed-m-v1.5`, `bert-base-uncased`, `distilbert-base-uncased`, `electra-base-discriminator`, `instructor-base`, `bekko-embedding-v1-a8m`, `e5-base-v2` and `deberta-v3-base`. The reduced probe-family panel uses the first three of each list. All results for these representations are in Tables 23 to 26.

Accessibility is estimated by the plug-in $\widehat{A_\lambda} = D_\Gamma^{-1/2}\hat{C}^\top(\hat{\Sigma}_Z + \lambda I)^{-1}\hat{\Sigma}_Z(\hat{\Sigma}_Z + \lambda I)^{-1}\hat{C}D_\Gamma^{-1/2}$, with $\lambda = \lambda_{\mathrm{rel}}\ \mathrm{tr}(\hat{\Sigma}_Z)/d$, so that the relative ridge λ_{rel} is specified relative to the mean eigenvalue of $\hat{\Sigma}_Z$. The plug-in is used rather than a split-block estimator for the reason given in Appendix C: because $\hat{\Sigma}_Z$, $\hat{C}$ and $\hat{\Gamma}$ are blocks of one sample joint covariance, the plug-in inherits the bound of Theorem 2 exactly, whereas a split-block estimator does not. Throughout the footprint analysis d means $\mathrm{rank}\,\hat{\Sigma}_Z$ and not the raw embedding width. It ranges from 255 to 2048 across pairs, with dataset means of 781 to 946, and every real-dataset pair sits at $n/d \geq 42$.

Erasure sets. Nine erasure sets were run, giving 321 runs in total across the four datasets. On CelebA: the single attributes `Male` and `Smiling`, and three semantically coupled triples: `gender_makeup` (`Male`, `Heavy_Makeup`, `Wearing_Lipstick`), `facial_hair` (`Mustache`, `Goatee`, `No_Beard`) and `age_appearance` (`Young`, `Attractive`, `Chubby`). On Bias-in-Bios: the single protected attribute `gender`. On EUR-Lex: `top_mi_cluster`, the four concepts with highest mutual information (`eurlex_693`, `eurlex_1605`, `eurlex_1118`, `eurlex_2635`). On GoEmotions: the single emotion `anger` and the four-member `joy_cluster` (`joy`, `amusement`, `excitement`, `admiration`). Erasers are fitted on the training split. Algebraic identities are evaluated on that fitting split, while the fixed train-fitted eraser is applied to validation for downstream probe evaluation: baseline probes are fit on train and scored on validation, and post-erasure probes are refit on erased train representations and scored on erased validation representations.

C CONTROLLED EXPERIMENTS AND ESTIMATOR BEHAVIOUR

This appendix examines three limitations of the accessibility measure: nonlinear encoding, finite-sample inflation of the estimator, and prevalence confounding. Each concerns one factor of $(A_0)_{jj} = c_j^\top \Sigma_Z^\dagger c_j / \gamma_j$. The second is a property of the estimator. The first and third are properties of the population quantity.

C.1 NONLINEAR-ONLY ENCODING

Both datasets have $d = 2$ features, one binary concept and $n = 4{,}000$ samples, split evenly into a fitting half and an evaluation half. For XOR the features are $X \sim \mathcal{N}(0, I_2)$ and the label is $Y = \mathbb{1}[X_1 > 0] \oplus \mathbb{1}[X_2 > 0]$, a deterministic function of X. For Rings the angle is uniform on $[0, 2\pi)$, the radius is drawn from $\mathcal{N}(0.5, 0.05^2)$ for the inner class and $\mathcal{N}(1.5, 0.05^2)$ for the outer class with equal probability, and Y indicates the outer ring. In both cases the classes are balanced and, by symmetry, $\mathrm{Cov}(X, Y) = 0$, so the concept has no linear signal at all. Table 3 compares accessibility with a logistic probe and with a two-hidden-layer MLP (64 units each) fitted on the same split.

Data	$(A_0)_{jj}$	linear probe AUROC	MLP probe AUROC
XOR	9.7×10^{-4}	0.510	1.000
Rings	7.0×10^{-4}	0.507	1.000

Table 3: Accessibility on data whose class information is carried nonlinearly, so that the population $C = \mathrm{Cov}(Z, Y)$ vanishes. The framework reports near-zero *linear* accessibility where a nonlinear probe achieves near-perfect discrimination.

The linear probe (0.510, 0.507) agrees with the near-zero accessibility estimate, while the MLP recovers the concept almost perfectly. A_0 therefore measures linear accessibility correctly but does not detect a nonlinear encoding. We do not run an eraser on these data, but the consequence for erasure follows. In these examples the population linearly accessible target subspace is empty, so LEACE has no nonzero target direction to remove, and A_0 reports the concept as already inaccessible to a linear readout.

C.2 ESTIMATOR INFLATION NEAR $n \approx d$

When $C = 0$ the population accessibility vanishes regardless of how ill-conditioned Σ_Z is. The plug-in estimator $\widehat{\mathrm{tr}\, A_0} = \mathrm{tr}(D_\Gamma^{-1/2} \hat{C}^\top \hat{\Sigma}_Z^\dagger \hat{C} D_\Gamma^{-1/2})$ does not share that property: a finite-sample $\hat{C}$ that is pure noise is amplified by an unstable $\hat{\Sigma}_Z^\dagger$.

We study this on planted data, meaning features and labels drawn jointly Gaussian from a covariance whose cross-covariance C is set by hand, so that the population A_0 is known in closed form. The generator takes $\Sigma = V \,\mathrm{diag}(s) V^\top$ for a random orthogonal V and a power-law spectrum $s_k = k^{-\alpha}$, and sets $C = \Sigma F$ with $F = \sigma \mathcal{N}(0, 1)^{d \times m}$, where σ controls the strength of the linear signal. We sweep n/d, the spectrum exponent $\alpha \in \{0.5, 1, 2\}$ and the relative ridge, at $d = 60$ and $m = 6$, in two settings with six seeds per configuration: an independent one with $\sigma = 0$, so that $C = 0$ and the population $\mathrm{tr}(A_0)/m$ is 0, and a weakly coupled one with $\sigma = 0.08$, which gives a small positive population value and so tests whether the inflation also occurs around a nonzero target. The numerical ratios below are all taken on the $\alpha = 2$ slice, which is also the slice shown in panel (a). For that slice each setting has an empirical reference value of $\mathrm{tr}(A_0)/m$, measured on a single draw of 200,000 samples from the same generator: 0.0003 in the independent setting, whose exact population value is 0, and 0.0115 in the weakly coupled one, whose exact population value is 0.0113. Panel (a) of Table 4 gives the independent setting, Figure 2 the sweep itself, and Figure 3 the same draws against the held-out probe.

For $n \leq d$ the unregularised estimate equals 1.0000, the maximal accessibility, for labels statistically independent of the features. This is a factor of 3200 above the reference, while the held-out probe R^2 over the same draws is negative throughout. The value is 1 because for $n \leq d$ the sample covariance is rank deficient and $\hat{\Sigma}^\dagger$ inverts the subspace the data span. In that subspace the regression interpolates, so $\hat{C}^\top \hat{\Sigma}^\dagger \hat{C} = \hat{\Gamma}$ and the estimate reaches its own Schur bound. The reported accessibility

(a) Plug-in inflation under independent labels ($C = 0$), spectrum exponent $\alpha = 2$

n/d	$\lambda = 0$	10^{-3}	10^{-2}	0.1	1	held-out R^2	$\hat{\kappa}$
0.25	1.0000	0.9930	0.9374	0.6780	0.3120	-1.400	5.2×10^2
0.50	1.0000	0.9713	0.8158	0.4515	0.1496	-2.282	4.2×10^3
0.75	1.0000	0.9095	0.6521	0.3171	0.1168	-3.635	2.7×10^4
1.00	1.0000	0.7793	0.4987	0.2117	0.0652	-59.898	2.1×10^6
1.50	0.6800	0.5804	0.3583	0.1451	0.0443	-2.444	3.8×10^4
2.00	0.5159	0.4590	0.2944	0.1225	0.0407	-1.113	1.7×10^4
5.00	0.1987	0.1839	0.1218	0.0472	0.0142	-0.266	6.3×10^3
10.00	0.1027	0.0958	0.0646	0.0250	0.0078	-0.120	4.6×10^3
25.00	0.0414	0.0388	0.0266	0.0099	0.0028	-0.046	3.9×10^3
100.00	0.0100	0.0094	0.0065	0.0025	0.0008	-0.011	3.7×10^3

(b) Estimator comparison on planted data ($d = 40$, $m = 6$, three seeds)

		plug-in		split-block		Ledoit–Wolf	
n/d	λ_{rel}	tr mean	tr sd	tr mean	tr sd	tr mean	tr sd
0.50	10^{-3}	0.998	3.6×10^{-4}	1.668	0.373	1.254	0.095
	10^{-2}	0.979	0.003	1.599	0.353	1.241	0.094
	10^{-1}	0.845	0.017	1.162	0.255	1.126	0.091
1.00	10^{-3}	0.959	0.015	13.307	2.022	0.901	0.029
	10^{-2}	0.884	0.010	3.940	0.256	0.890	0.029
	10^{-1}	0.682	0.007	1.217	0.152	0.791	0.034
2.00	10^{-3}	0.703	0.034	1.663	0.210	0.666	0.044
	10^{-2}	0.676	0.032	1.496	0.174	0.654	0.044
	10^{-1}	0.533	0.032	0.860	0.090	0.564	0.046
5.00	10^{-3}	0.556	0.028	0.791	0.046	0.550	0.026
	10^{-2}	0.543	0.026	0.758	0.039	0.540	0.026
	10^{-1}	0.452	0.023	0.562	0.022	0.464	0.024
10.00	10^{-3}	0.506	0.014	0.604	0.023	0.504	0.014
	10^{-2}	0.495	0.014	0.586	0.021	0.495	0.014
	10^{-1}	0.418	0.012	0.469	0.017	0.423	0.012
25.00	10^{-3}	0.475	3.0×10^{-4}	0.531	0.040	0.475	2.2×10^{-4}
	10^{-2}	0.466	2.2×10^{-4}	0.520	0.039	0.466	2.4×10^{-4}
	10^{-1}	0.395	9.0×10^{-4}	0.434	0.035	0.397	9.3×10^{-4}
50.00	10^{-3}	0.457	0.008	0.483	0.022	0.457	0.008
	10^{-2}	0.449	0.008	0.473	0.022	0.449	0.008
	10^{-1}	0.385	0.007	0.402	0.021	0.386	0.007
100.00	10^{-3}	0.453	0.003	0.461	0.006	0.453	0.003
	10^{-2}	0.445	0.003	0.452	0.006	0.445	0.003
	10^{-1}	0.381	0.003	0.385	0.006	0.382	0.003

Table 4: Finite-sample estimator behaviour near $n \approx d$. **(a)** $\widehat{\mathrm{tr}\, A_\lambda}/m$ for the $\alpha = 2$ slice of the independent-label Gaussian synthetic model ($C = 0$), with the intermediate ridge columns (10^{-4}, 0.3) omitted. Held-out R^2 and the condition-number estimate $\hat{\kappa}$ are at $\lambda = 0$. The exact population accessibility is 0 and the 200,000-sample empirical reference is 0.0003, so every entry of the $\lambda = 0$ column is inflation above that reference. **(b)** plug-in, split-block and Ledoit–Wolf estimates in a separate Gaussian synthetic sweep with a nonzero planted cross-covariance, where the trace mean and sd are over three seeds. The recovered rank is 6.00 in every cell, so that column is not printed.

is then the in-sample R^2 of an interpolating probe. Conditioning amplifies the effect but does not cause it: $\hat{\kappa}$ peaks at 2.1×10^6 at $n/d = 1$. Ridge reduces the inflation without removing it, and the estimate is strictly decreasing over the seven swept ridge values in every one of the 360 individual draws: at $n/d = 1$ raising the relative ridge from 0 to 1 cuts the estimate from 1.0000 to 0.0652, a 15-fold reduction that is still 200 times the reference. The weakly coupled setting behaves in the same way around a nonzero target: with reference value $\mathrm{tr}(A_0)/m = 0.0115$, the unregularised estimate is again 1.0000 for all $n/d \leq 1$, an 87-fold overstatement, with held-out $R^2 = -60.3$ at $n/d = 1$, while the estimate falls to 0.0206 at $n/d = 100$.

The split-block excursion above the Schur bound. The sweep of panel (b) reports a split-block trace of 13.31 at $n/d = 1$ and relative ridge 10^{-3}, against 1.67 at $n/d = 0.5$ and 0.79 at $n/d = 5$. A value of 13.31 is impossible for the population quantity, since Theorem 2 forces every diagonal entry to at most 1. The two estimators differ in whether they respect that bound. The Schur bound is

a statement about a joint covariance: it holds because $\left[\begin{smallmatrix}\Sigma & C\\ C^\top & \Gamma\end{smallmatrix}\right]$ is positive semidefinite, which forces $C^\top \Sigma^\dagger C \preceq \Gamma$. The plug-in estimator satisfies it, because $\hat{\Sigma}$, $\hat{C}$ and $\hat{\Gamma}$ are blocks of one sample joint covariance, which is positive semidefinite as a Gram matrix of the same rows. The split-block estimator takes $\hat{\Sigma}$ from split A and $\hat{C}$ from split B, and that pair is not the block structure of any single joint covariance, so the inequality does not apply and $\hat{C}_B^\top \hat{\Sigma}_A^\dagger \hat{C}_B$ is unbounded above, since $\hat{\Sigma}_A^\dagger$ is large along directions where $\hat{\Sigma}_A$ has small eigenvalues and $\hat{C}_B$, independent of split A, has no reason to be small in those directions. Both statements hold empirically: across the 2,520 swept draws (420 configurations of two settings, three spectrum exponents, ten sample ratios and seven ridge values, each with six seeds) the plug-in estimate exceeds 1 in 0.0% of cases while the split-block exceeds it in 25.3%. Near $n \approx d$, the split-block estimator no longer inherits the finite-sample upper bound of the plug-in estimator and can produce arbitrarily large estimates. We therefore report held-out probe performance alongside both estimates.

Panel (b) of Table 4 reports a separate sweep on planted data with $d = 40$, comparing the plug-in, split-block and Ledoit–Wolf estimators at eight values of n/d and three relative ridges. The recovered rank equals the true rank of six in every cell, whereas the trace depends on both the sample ratio and the estimator.

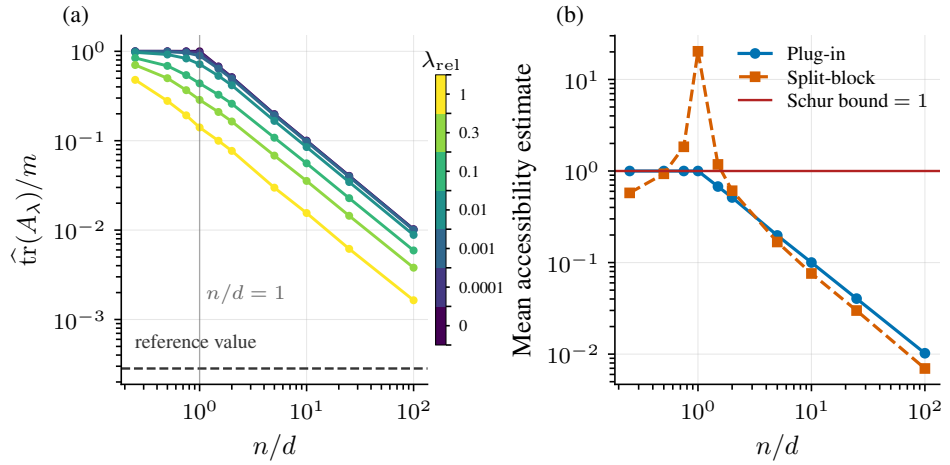

Figure 2: Finite-sample inflation of accessibility estimates, independent setting. **(a)**: estimated $\mathrm{tr}(A_\lambda)/m$ against n/d, one curve per relative ridge strength, averaged over the three spectrum exponents. The dashed horizontal line is the empirical reference value 0.0003 on the same average, and the exact population value is 0. The vertical line marks the interpolation threshold $n = d$. The unregularised estimate stays at 1 for $n/d \leq 1$, and each increase in λ reduces the inflation. **(b)**: mean plug-in and split-block estimates across n/d. The plug-in estimator reaches but does not exceed its finite-sample Schur bound of 1, whereas the split-block estimator need not satisfy that bound and becomes unstable near the interpolation regime, before the two converge as the sample ratio grows.

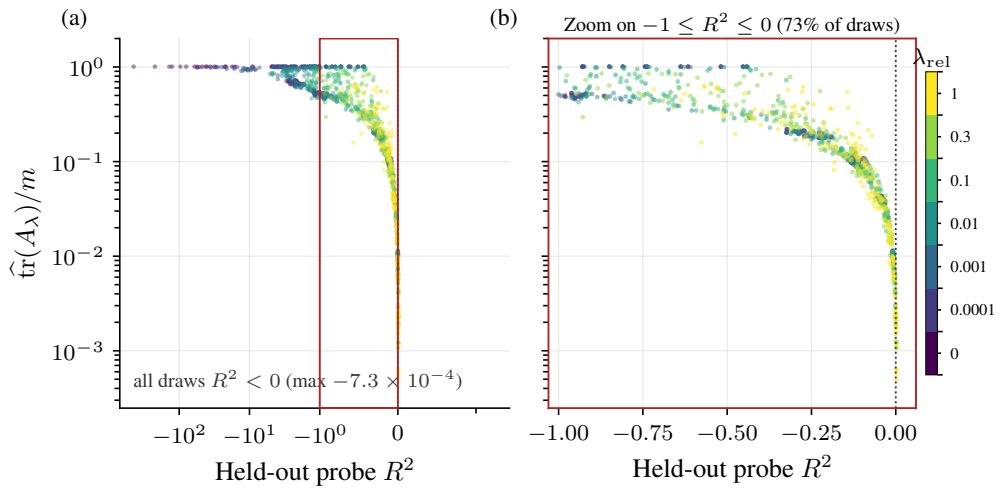

Figure 3: Held-out diagnostic for the independent-label sweep of Figure 2, plotting the accessibility estimate against the held-out probe R^2. **(a)**: the full range on a symmetric-log horizontal axis. **(b)**: a linear zoom on the marked interval $-1 \le R^2 \le 0$, which holds 73% of the draws. Every draw in this setting has negative held-out R^2, from -7.3×10^{-4} down to -436.9, while every estimate is positive and the unregularised ones at $n/d \le 1$ sit at the ceiling: the inflation is not visible to the estimate itself, only to an independent measurement.

Calibration on planted data. On a third planted configuration ($d = 20$, $m = 6$), the diagonal of A_0 approaches the population R^2 as the sample grows, and the Schur residual error stays at floating-point level, near 10^{-15}, at every sample size from 10^3 to 10^5. On real datasets the identities are compared against this floor.

C.3 Prevalence confounding

Table 5 sweeps prevalence at fixed class separation, and Figure 4 places that sweep beside the nonlinear-encoding result.

p	γ	$(A_0)_{jj}$	measured AUROC	calibrated AUROC
0.500	0.2500	0.1905	0.7563	0.7537
0.300	0.2100	0.1626	0.7466	0.7517
0.200	0.1600	0.1307	0.7482	0.7534
0.100	0.0900	0.0782	0.7613	0.7537
0.050	0.0475	0.0405	0.7527	0.7474
0.020	0.0196	0.0192	0.7568	0.7602
0.010	0.0099	0.0094	0.7368	0.7556

Table 5: Prevalence confounding of R^2 at fixed class separation. Accessibility falls by a factor of 20 while measured AUROC is flat, and the calibration of Equation (5) tracks the flat line.

Under a Gaussian model in which both classes share one covariance, the separation of concept j is summarised by a signal-to-noise ratio S_j, which determines the AUROC:

$$S_j = \frac{R_j^2}{\gamma_j(1 - R_j^2)}, \qquad \mathrm{AUROC}_j = \Phi\big(\sqrt{S_j/2}\big), \tag{5}$$

At fixed class separation S_j is fixed, so as prevalence falls the label variance $\gamma_j = p_j(1-p_j)$ shrinks and R_j^2 falls with it, while the AUROC does not change. Raw R^2 should therefore not be compared across concepts of different prevalence. Three of the four datasets have median prevalence below 0.03, so ranking their concepts by raw R^2 would largely rank them by rarity.

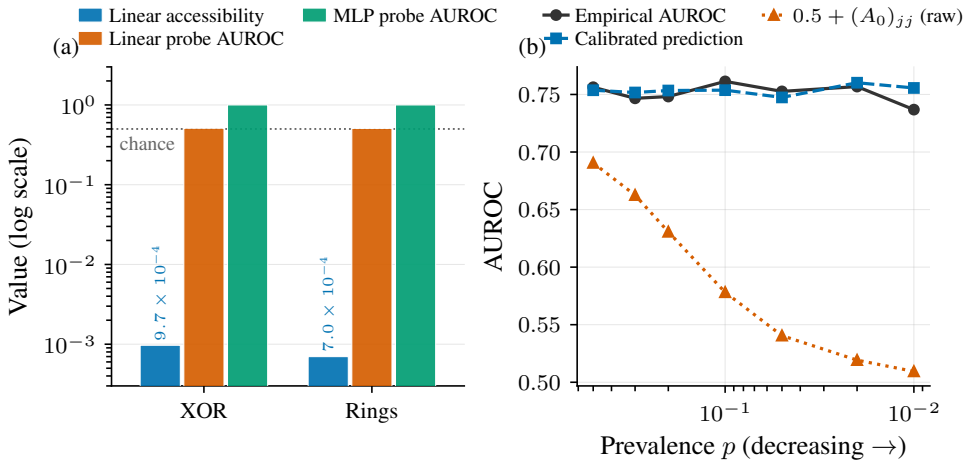

Figure 4: **(a)**: on XOR and concentric-ring data the linear accessibility is near zero and the linear probe sits at chance, while the nonlinear MLP probe achieves near-perfect discrimination. The vertical axis is logarithmic. **(b)**: at fixed class separation the measured AUROC stays flat near 0.75 as prevalence falls, while the uncalibrated accessibility-based heuristic $0.5 + (A_0)_{jj}$ decays, and the Gaussian calibration of Equation (5) tracks the measured AUROC. Prevalence is on a logarithmic axis and decreases from left to right.

D ADDITIONAL DIAGNOSTICS

This appendix reports how the accessible task covariance compares with label-free spectral scores, how it behaves as a model-selection statistic, the cross-concept coupling identity, the shortcut-concept diagnostic, and descriptive observations about the representations themselves.

D.1 COMPARISON AGAINST LABEL-FREE SPECTRAL SCORES

Label-free spectral measures describe how variance is spread over eigendirections, without reference to where the concept signal sits inside that spectrum. The alignment sweep tests this: it holds the embedding eigenvalue spectrum fixed while sliding the concept signal from the top eigendirections ($t = 0$) to the bottom ($t = 1$), so any metric computed from the population spectrum alone is constant by design, while a metric that uses the labels need not be.

Over the sweep RankMe varies by 2.8% and NESum by 9.6%, while the ridge probe R^2 falls from 0.793 to -0.005 (Figure 5). The regularised trace $\mathrm{tr}(A_\lambda)/m$ at $\lambda_{\mathrm{rel}} = 10^{-2}$ tracks that fall at Spearman 0.964 and Pearson 0.927, whereas RankMe correlates with it at Pearson -0.236. The unregularised quantities behave in the complementary way: $\mathrm{tr}(A_0)/m$ varies by 1.1% across the sweep and the ordinary-least-squares probe R^2 by 1.0%, both near-constant, because full whitening is invariant to reparameterisation and an unregularised probe can reach any direction of nonzero variance. A_0 predicts what an unregularised probe achieves and A_λ what a ridge probe achieves, and the gap between them measures the alignment.

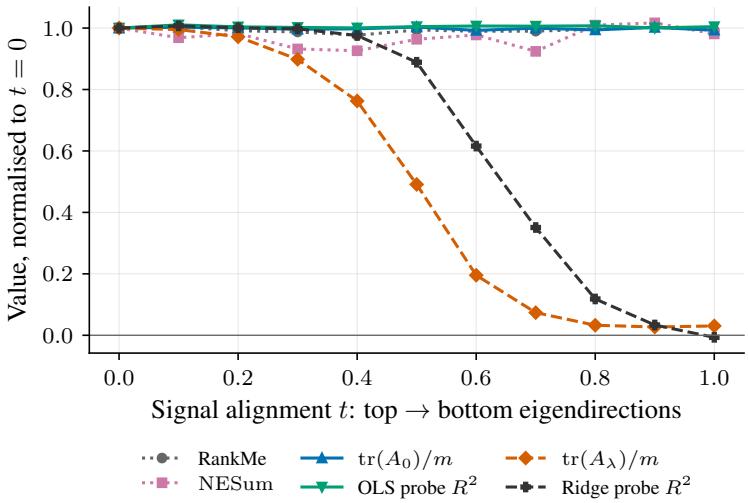

Figure 5: Alignment sweep at fixed population embedding spectrum, with all displayed quantities normalised to their value at $t = 0$. The concept signal is moved continuously from the top eigendirections at $t = 0$ to the bottom eigendirections at $t = 1$. RankMe, NESum, $\mathrm{tr}(A_0)/m$ and the OLS probe R^2 remain approximately unchanged, whereas the ridge probe R^2 and $\mathrm{tr}(A_\lambda)/m$ both decrease substantially. The label-free spectral summaries therefore do not respond to a change in the alignment of the concept signal with the covariance spectrum.

D.2 Cross-concept coupling: the Fisher identity

The heteroskedastic Fisher coupling identity, derived in Appendix A, predicts the correlation between the logits of Fisher discriminants computed from the class-conditional means and covariances. Compared against the logits of the trained probes, the prediction agrees at Spearman 0.970 over $4{,}725$ (concept pair, representation) combinations on CelebA and the agreement degrades with label sparsity, to 0.749 over $9{,}744$ on Bias-in-Bios, 0.722 over $9{,}072$ on GoEmotions, and 0.469 over $29{,}400$ on EUR-Lex, which has the lowest mean prevalence and the least reliable class-conditional covariance estimates. The ordering follows mean prevalence, which is consistent with the identity needing the within-class covariances to be well estimated, since rare positive classes leave them estimated from few samples.

D.3 Accessible coupling against label correlation

We call $|(\tilde{A}_0)_{ij}| - |R^Y_{ij}|$ the coupling gap: it is positive when a pair of concepts is more strongly coupled in the representation than in the labels. Table 6 summarises it by dataset. Figure 6 shows the CelebA coupling matrices beside the observed collateral, and Figure 7 gives the distribution over CelebA pairs.

Dataset	μ_{task}	μ_{label}	mean	median	max	frac. > 0
CelebA	0.9720	0.8021	0.1374	0.1152	0.5214	98.7%
EUR-Lex	0.9821	0.9142	0.0414	0.0244	0.5939	93.4%
GoEmotions	0.6564	0.2106	0.0699	0.0513	0.6260	88.2%
Bias-in-Bios	0.5579	0.2214	0.0391	0.0213	0.7026	82.1%

Table 6: Accessible-task and label mutual coherence, each averaged over the dataset's representations, and the distribution of the coupling gap $|(\tilde{A}_0)_{ij}| - |R^Y_{ij}|$ over all (concept pair, representation) entries.

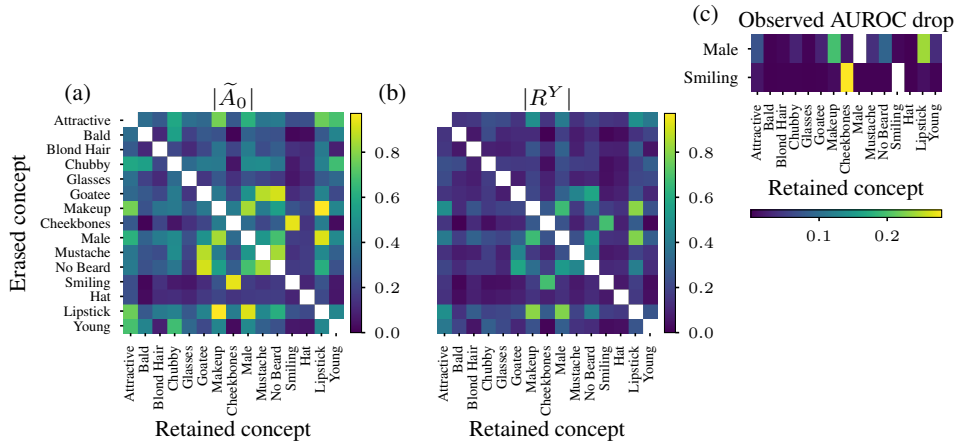

Figure 6: CelebA, averaged over the 45 backbones, with the same retained-concept ordering in all three panels. **(a)**: $|\tilde{A}_0|$, the accessible coupling between attributes. **(b)**: $|R^Y|$, the raw label correlation. **(c)**: the observed held-out AUROC drop for the two available single-concept erasures, `Male` and `Smiling`. Multi-concept erasures cannot be assigned to a unique erased-concept row and are therefore not shown in this pairwise panel. Panels (a) and (b) are normalised couplings on $[0, 1]$ and share a colour scale. Panel (c) is measured in AUROC-drop units and uses its own scale, so its colours are not quantitatively comparable with panels (a) and (b). Theorem 2 bounds A_0 by R^Y in the Loewner order rather than entrywise. The gender-presentation region is prominent in the accessible-coupling matrix, and the `Male` row of panel (c) shows large drops for the corresponding retained attributes.

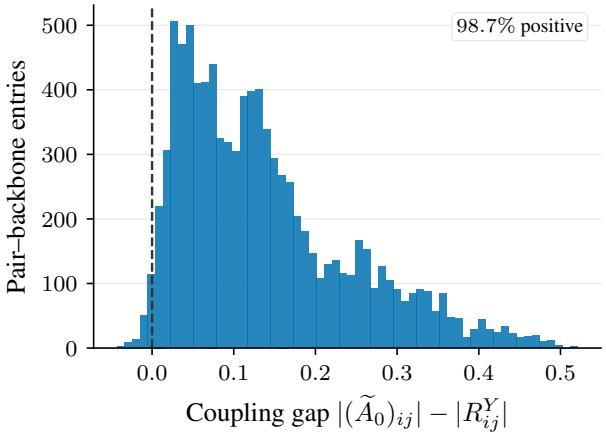

Figure 7: Coupling gap on CelebA. Distribution over all (concept pair, backbone) entries of $|(\tilde{A}_0)_{ij}| - |R^Y_{ij}|$. Most entries lie above zero (98.7%), so the representation-specific accessible coupling is typically larger in magnitude than the corresponding raw label correlation. Entries farther to the right are pairs and backbones for which the accessible-coupling surrogate implies a larger relative collateral-loss signal than the squared-label-correlation surrogate. This is a statement about entries: Theorem 2 orders A_0 and R^Y in the Loewner sense only, and does not predict entrywise dominance in either direction.

D.4 Shortcut-concept indices

Table 7 gives, for the five highest-scoring attributes, the full and conditional accessibilities from which the shortcut index is computed. Both columns are ridge accessibilities, unlike the unregularised $(A_0)_{jj}$ of Table 20. The remaining three scored attributes, as R^2_λ / $R^2_{\lambda|S}$ / index, are `Goatee` (controls `Male`) 0.4097 / 0.3453 / 0.1804, `Bald` (`Male`, `Young`) 0.3561 / 0.3117 / 0.1406 and `Blond_Hair` (`Male`) 0.6018 / 0.5531 / 0.0826.

Attribute	Controls	R^2_λ	$R^2_{\lambda\|S}$	shortcut index
`Wearing_Lipstick`	`Male`	0.7257	0.3002	0.5927
`Heavy_Makeup`	`Male`	0.6486	0.3465	0.4750
`Attractive`	`Male,Young,Smiling`	0.4626	0.2583	0.4457
`No_Beard`	`Male`	0.5883	0.4357	0.2864
`Mustache`	`Male`	0.2426	0.1942	0.2354

Table 7: Shortcut indices $1 - R^2_{j|S}/R^2_j$ on CelebA, computed from ridge accessibility A_λ at relative ridge $\lambda/\bar{\lambda}_\Sigma = 10^{-2}$. Rows are the top five of the eight scored attributes by that index, and the other three are given in the text above. Each column is averaged over the 45 backbones, the index being formed per backbone before averaging, so it need not equal one minus the ratio of the two displayed means. Two of the three attributes the theory singles out as shortcut candidates (`Wearing_Lipstick` and `Attractive`) are in the top three. `Mustache`, the third, is fifth, and `Heavy_Makeup`, which the theory does not single out, is second.

D.5 Ordinal model selection

Table 8 gives the ranking correlations with their bootstrap intervals.

Dataset	n	$\rho(\mathrm{tr}(A)/m, \mathrm{balacc})$	$\rho(\mathrm{erank}(A), \mathrm{balacc})$	$\rho(\mathrm{erank}(\text{logit cov}), \mathrm{balacc})$ [95% CI]
CelebA	45	$+0.946$	$+0.920$	$+0.208$ $[-0.110, +0.494]$
Bias-in-Bios	24	$+0.502$	$+0.357$	$+0.457$ $[+0.037, +0.745]$
EUR-Lex	24	$+0.212$	$+0.163$	-0.257 $[-0.665, +0.206]$
GoEmotions	24	$+0.278$	$+0.199$	-0.242 $[-0.646, +0.221]$

Table 8: Ordinal model selection. Spearman over representations. Here A is the ridge accessibility A_λ at $\lambda_{\mathrm{rel}} = 10^{-2}$. Bootstrap 95% intervals are given for the logit-covariance comparator, and cross zero on three of the four datasets.

Since $\mathrm{tr}(A)/m$ is by definition the mean of the per-concept R^2 on the diagonal of A (Equation (1)), ranking representations by it is ranking them by mean per-concept R^2, and the two stored columns are identical. This is a consistency check rather than a finding. The rank correlation between the calibrated AUROC predicted through Equation (5) and measured AUROC is 0.951 (CelebA), 0.969 (Bias-in-Bios), 0.890 (EUR-Lex) and 0.800 (GoEmotions), which is the prevalence correction working on real concepts rather than planted ones. The ordering of backbones induced by $\mathrm{tr}(A_\lambda)/m$ on CelebA is stable across five orders of magnitude of the relative ridge: the worst pairwise rank correlation over $\lambda_{\mathrm{rel}} \in [10^{-5}, 1]$ is 0.971, between the two endpoints.

D.6 Descriptive model-level results

Independently of the erasure results, the sweep compares representations under one fixed protocol. The statements below describe these four datasets under this extraction and probing protocol and are not general claims about pretraining. On CelebA, language-supervised pretraining leads: the top five backbones are `siglip_so400m_gap` (0.9262 balanced accuracy), `siglip_base_gap` (0.9257), `vitamin_clip` (0.9251), `clip-vitb16` (0.9246) and `eva02_base` (0.9243). By family, `clip` averages 0.9222 and `simclr` 0.7229. Dense-segmentation pretraining sits near the bottom (`samvit_base` 0.8360, `sam2_hiera` 0.8711) despite the largest input resolutions in the set (896–1024px). Embedding width shows little association with balanced accuracy here: Spearman $+0.178$ over the 45 backbones, and the weakest backbone, `simclr-rn50`, carries the set's largest width (2048, shared with four others). Among text encoders, purpose-built embedding models beat the raw masked-language-model controls on most datasets but not all, and by small margins. Width again does not track performance, with `bge-small` (384) in the top three on all three datasets, ahead of `bge-large` (1024) from the same family on all three, and `all-MiniLM-L6-v2` (384) leading EUR-Lex. `electra-base-discriminator` is the weakest raw encoder on every dataset.

E ERASURE ANALYSES

This appendix extends the erasure results of Section 5. It covers the nine erasure sets, rank correlations for the identities, the set-level coupling comparison across the three collateral measures, the comparison over every single-concept erasure, the probe-family and ridge checks, the screening experiment, the normalisation of the R^2 comparison, and the edit footprints.

E.1 ERASURE SETS

Dataset	**Erasure set**	$\lvert S\rvert$	rank	depth	ΔR^2 ret.	Δbal. ret.
CelebA	`Male`	1	1.0	0.8672	0.1523	0.0542
	`Smiling`	1	1.0	0.6386	0.0467	0.0202
	`age_appearance`	3	3.0	0.4105	0.1653	0.0523
	`facial_hair`	3	3.0	0.4289	0.1280	0.0457
	`gender_makeup`	3	3.0	0.7533	0.1062	0.0324
Bias-in-Bios	`gender`	1	1.0	0.8212	0.0046	0.0013
EUR-Lex	`top_mi_cluster`	4	4.0	0.8578	0.0274	0.0070
GoEmotions	`anger`	1	1.0	0.1838	0.0038	0.0051
	`joy_cluster`	4	4.0	0.2419	0.0086	0.0060

Table 9: All nine erasure sets, averaged over the representations of each dataset (321 runs). "rank" is the measured Mahalanobis edit rank of Corollary 3, "depth" the mean ΔR^2 on the erased concepts, and the last two columns the collateral damage on the retained concepts. For every single-concept set the depth equals that concept's pre-erasure accessibility to four decimal places, because the post-erasure accessibility is zero. Table 20 reports those values for the CelebA attributes.

Dataset	**Target**	erasure depth	retained ΔR^2
CelebA	`Male`	0.8672	0.1523
Bias-in-Bios	`gender`	0.8212	0.0046

Table 10: The two gender-erasure settings we evaluate, at comparable depth on their own target. Means over the 45 and 24 representations respectively. The two settings differ in dataset, concept inventory, retained targets and representation family as well as in modality, so the comparison does not isolate a modality effect. Table 9 lists every erasure set.

E.2 RANK CORRELATIONS FOR THE IDENTITIES

LEACE has the minimal guarding kernel, so the Schur comparison is the equality case of Theorem 1. Any other guarding map leaves retained accessibility no greater than this value in the Loewner order. The equality holds to near-machine precision on three datasets, where the worst case over every row is below 5×10^{-13}. EUR-Lex is looser, with maximum absolute error 5.9×10^{-4}, and the same near-degeneracy keeps its post-erasure target R^2 at 1.5×10^{-4} rather than at floating-point zero, against 7.1×10^{-26} on CelebA, 2.9×10^{-25} on Bias-in-Bios and 6.3×10^{-27} on GoEmotions. For the footprint identity on held-out data, where the whitening is no longer exact, the median rank correlation between predicted and measured profiles is 1.000 on every dataset and the minimum over all footprints is 0.980. The edit cost is 1.0 for the four single-concept sets, 3.0 for the three CelebA triples and 4.0 for the EUR-Lex and GoEmotions clusters.

Table 11 reports the identities by absolute error, which is the appropriate test for an equality. For completeness, the Spearman and Pearson correlations between Schur-predicted and measured post-erasure R^2 are both 1.0000 on all four datasets, over the same 2,880, 672, 1,104 and 1,224 rows, which are the 5,880 retained-concept rows of Section 5.4. The deviations are below the resolution of any scatter plot, so the error columns of Table 11 are the informative form.

For the footprint, the invariants $0 \le \ell_i \le 1$, $\sum_i \ell_i = r$ and $1 \le \mu_{\mathrm{er}} \le d/r$ hold in 100% of the 3,564 footprints, one for every representation paired with every concept and every erasure set. The rank of the edit matches the rank the eraser itself reports in 100%, and the median absolute deviation between predicted leverage and measured per-coordinate mean squared edit is 1.1×10^{-16}.

	Collateral damage (Schur)			Edit footprint	
Dataset	rows	MAE	max abs. error	footprints	max abs. error
CelebA	2,880	3.8×10^{-15}	4.8×10^{-13}	900	1.2×10^{-13}
Bias-in-Bios	672	9.4×10^{-16}	3.3×10^{-14}	720	4.1×10^{-14}
EUR-Lex	1,104	9.9×10^{-6}	5.9×10^{-4}	1,224	4.8×10^{-13}
GoEmotions	1,224	5.4×10^{-16}	3.7×10^{-14}	720	4.7×10^{-14}

Table 11: Numerical verification of the two closed-form identities by absolute error. Left: Schur-predicted against measured post-erasure R^2 on retained concepts, over concept $\times$ representation $\times$ erasure-set rows. Right: leverage-predicted against measured per-coordinate mean squared edit. The larger EUR-Lex residuals are associated with the near-degenerate covariance discussed in the text.

E.3 Set-level coupling comparison across collateral measures

Section 5.2 restricts the comparison between accessible coupling and label correlation to single-concept erasures, because that is the setting in which the predictor is exactly the relative loss of Corollary 2. For a multi-concept erasure set the pairwise form is not identified, since there is no unique erased concept i in $(\tilde{A}_0)_{ij}$. To score the multi-concept sets at all, we summarise the set by

$$\max_{i \in S} |(\tilde{A}_0)_{ij}|^2 \qquad \text{against} \qquad \max_{i \in S} |R^Y_{ij}|^2.$$

The maximum is the simplest summary that reduces to the single-concept predictor when $|S| = 1$. It is a heuristic, not a consequence of Corollary 2, and a different summary, such as the mean or the full Schur form $A_{jS}A_{SS}^{\dagger}A_{Sj}$, could change the margins.

Panel (a) of Table 12 gives the resulting set-level comparison over all four datasets and all three collateral measures.

The set-level analysis is qualitatively consistent with the single-concept analysis in the main text. On the AUROC drop, accessible coupling is ahead on all four datasets and all four representation-cluster intervals exclude zero. On the balanced-accuracy drop it is ahead on CelebA and GoEmotions, while the Bias-in-Bios and EUR-Lex intervals include zero. On the R^2 drop the pattern differs, with label correlation ahead on Bias-in-Bios and EUR-Lex, but that block scores a relative predictor against an absolute loss. The difference in normalisation alone does not explain that reversal: on the expanded single-concept panel the comparison with absolute R^2 loss also favours accessible coupling on Bias-in-Bios and EUR-Lex (Table 19). On the AUROC drop the margin is largest on GoEmotions and EUR-Lex, the two datasets with the lowest label-correlation ρ, and smallest on CelebA and Bias-in-Bios. Restricting the set-level analysis to multi-concept sets, as a point estimate without an interval, widens the CelebA AUROC margin from $+0.079$ to $+0.141$ but narrows the GoEmotions margin from $+0.286$ to $+0.216$, so the effect of set size is not uniform across datasets. Panel (b) of Table 12 repeats the single-concept comparison one representation at a time.

(a) Set level, all configured erasure sets, all three collateral measures

Dataset	rows	ρ coupling	ρ label corr.	$\Delta\rho$	95% CI
AUROC drop (probe metric)					
CelebA	2880	0.855	0.775	$+0.079$	$[+0.042, +0.103]$
Bias-in-Bios	672	0.662	0.589	$+0.073$	$[+0.040, +0.107]$
EUR-Lex	1104	0.552	0.280	$+0.272$	$[+0.214, +0.324]$
GoEmotions	1224	0.705	0.419	$+0.286$	$[+0.247, +0.322]$
Balanced-accuracy drop (probe metric)					
CelebA	2880	0.819	0.727	$+0.093$	$[+0.050, +0.120]$
Bias-in-Bios	672	0.343	0.313	$+0.030^{\dagger}$	$[-0.024, +0.086]$
EUR-Lex	1104	0.270	0.244	$+0.026^{\dagger}$	$[-0.034, +0.089]$
GoEmotions	1224	0.435	0.320	$+0.115$	$[+0.069, +0.161]$
R^2 drop (consistency comparison, algebraically linked to the Schur formula)					
CelebA	2880	0.899	0.900	$-0.001^{\dagger}$	$[-0.021, +0.011]$
Bias-in-Bios	672	0.848	0.956	$-0.108^{\ddagger}$	$[-0.132, -0.081]$
EUR-Lex	1104	0.483	0.677	$-0.194^{\ddagger}$	$[-0.273, -0.133]$
GoEmotions	1224	0.883	0.661	$+0.222$	$[+0.199, +0.250]$

(b) Per representation, original four-target single-concept panel

Dataset	reps	wins	ties	losses	win frac.	median $\Delta\rho$	sign-test p
AUROC drop							
CelebA	45	44	0	1	0.978	$+0.022$	2.61×10^{-12}
Bias-in-Bios	24	21	0	3	0.875	$+0.052$	2.77×10^{-4}
GoEmotions	24	24	0	0	1.000	$+0.328$	1.19×10^{-7}
Balanced-accuracy drop							
CelebA	45	40	0	5	0.889	$+0.027$	7.88×10^{-8}
Bias-in-Bios	24	13	0	11	0.542	$+0.010$	0.839
GoEmotions	24	20	0	4	0.833	$+0.136$	0.00154

(c) Per erased target, expanded single-concept panel

Scope	targets	wins	ties	losses	win frac.	median $\Delta\rho$
AUROC drop						
Pooled	122	115	0	7	0.943	$+0.192$
CelebA	15	11	0	4	0.733	$+0.036$
Bias-in-Bios	29	29	0	0	1.000	$+0.332$
EUR-Lex	50	48	0	2	0.960	$+0.096$
GoEmotions	28	27	0	1	0.964	$+0.449$
Balanced-accuracy drop						
Pooled	122	96	0	26	0.787	$+0.052$
CelebA	15	9	0	6	0.600	$+0.021$
Bias-in-Bios	29	27	0	2	0.931	$+0.147$
EUR-Lex	50	33	0	17	0.660	$+0.012$
GoEmotions	28	27	0	1	0.964	$+0.210$

Table 12: Additional robustness breakdowns for accessible coupling against label correlation. A win is $\Delta\rho > 0$, and ties are excluded from the counts. **(a)** set level: intervals are paired representation-cluster bootstraps with $10{,}000$ resamples, and for multi-concept sets the pairwise predictor is aggregated as $\max_{i \in S}$ over erased concepts. The R^2 block scores the *absolute* drop while Corollary 2 makes the pairwise predictor a *relative* quantity, so it is a comparison with absolute R^2 loss (Appendix E.7). Here † marks intervals containing zero and ‡ intervals lying entirely below zero. **(b)** one Spearman comparison per representation on the original four-target panel, which uses all 45, 24 and 24 representations rather than the ten per corpus of the expanded panel. The last column is an exact two-sided binomial sign test. Representations within a dataset share erased and retained concepts, so the test treats as independent units that are not, and is indicative only. EUR-Lex is absent because its only erasure set in this grid is multi-concept. **(c)** one comparison per erased concept on the expanded panel. Targets within a corpus share representations, so no target-level significance test is reported.

E.4 Comparison over every single-concept erasure

Tables 13 and 14 give the comparison of Section 5.2 on the AUROC and balanced-accuracy drops under three resampling schemes, Table 15 lists the per-corpus intervals, and panel (c) of Table 12 scores each erased concept separately.

Within a corpus every concept is erased in every representation, so erased targets and representations form a crossed design. The three bootstrap schemes differ in which factor they resample. The target-cluster scheme resamples erased concepts with replacement and so propagates uncertainty over the

Scope / unit	rows	clusters	ρ coupling	ρ label	$\Delta\rho$	95% CI
Pooled (all four corpora), three resampling units, stratified by corpus						
target-cluster	42280	122	0.459	0.212	$+0.247$	$[+0.213, +0.281]$
representation-cluster	42280	40	0.459	0.212	$+0.247$	$[+0.216, +0.278]$
two-way-cluster	42280	1220	0.459	0.212	$+0.247$	$[+0.201, +0.294]$
Per corpus (target-cluster bootstrap)						
CelebA	2100	15	0.770	0.743	$+0.027$	$[+0.001, +0.056]$
Bias-in-Bios	8120	29	0.435	0.106	$+0.329$	$[+0.241, +0.419]$
EUR-Lex	24500	50	0.325	0.219	$+0.106$	$[+0.085, +0.127]$
GoEmotions	7560	28	0.566	0.100	$+0.466$	$[+0.379, +0.556]$

Table 13: Expanded single-concept panel, held-out AUROC drop (canonical logistic probe $C{=}1$). Every available concept of every corpus is erased in turn, and $\Delta\rho > 0$ favours normalised accessible coupling $(\tilde{A}_0)^2_{ij}$ over squared label correlation $(R^Y_{ij})^2$. Bootstrap scheme as described above.

Scope / unit	rows	clusters	ρ coupling	ρ label	$\Delta\rho$	95% CI
Pooled (all four corpora), three resampling units, stratified by corpus						
target-cluster	42280	122	0.303	0.206	$+0.098$	$[+0.081, +0.114]$
representation-cluster	42280	40	0.303	0.206	$+0.098$	$[+0.070, +0.127]$
two-way-cluster	42280	1220	0.303	0.206	$+0.098$	$[+0.065, +0.133]$
Per corpus (target-cluster bootstrap)						
CelebA	2100	15	0.720	0.700	$+0.020$	$[-0.015, +0.056]$
Bias-in-Bios	8120	29	0.326	0.136	$+0.190$	$[+0.140, +0.238]$
EUR-Lex	24500	50	0.186	0.172	$+0.013$	$[+0.003, +0.024]$
GoEmotions	7560	28	0.310	0.093	$+0.217$	$[+0.174, +0.262]$

Table 14: Expanded single-concept panel, held-out balanced-accuracy drop (canonical logistic probe $C{=}1$). Columns as in Table 13.

Corpus	$\Delta\rho$	CI (target)	CI (representation)	CI (two-way)
AUROC drop				
CelebA	$+0.027$	$[+0.001, +0.056]$	$[-0.068, +0.092]$	$[-0.077, +0.109]$
Bias-in-Bios	$+0.329$	$[+0.241, +0.419]$	$[+0.191, +0.441]$	$[+0.167, +0.481]$
EUR-Lex	$+0.106$	$[+0.085, +0.127]$	$[+0.072, +0.138]$	$[+0.064, +0.146]$
GoEmotions	$+0.466$	$[+0.379, +0.556]$	$[+0.426, +0.504]$	$[+0.370, +0.568]$
Balanced-accuracy drop				
CelebA	$+0.020$	$[-0.015, +0.056]$	$[-0.094, +0.096]$	$[-0.106, +0.116]$
Bias-in-Bios	$+0.190$	$[+0.140, +0.238]$	$[+0.078, +0.291]$	$[+0.066, +0.309]$
EUR-Lex	$+0.013$	$[+0.003, +0.024]$	$[-0.002, +0.028]$	$[-0.007, +0.034]$
GoEmotions	$+0.217$	$[+0.174, +0.262]$	$[+0.161, +0.277]$	$[+0.144, +0.299]$

Table 15: Per-corpus $\Delta\rho$ under all three resampling units. The target-cluster column repeats the per-corpus intervals of Tables 13 and 14.

choice of concept, and the representation-cluster scheme does the same over representations. The two-way scheme resamples both independently and weights the row for concept t and representation r by the product of their two multiplicities, so it propagates uncertainty over both factors at once. The main-text table reports the two-way interval for every row because it propagates variation over both erased targets and representations. The per-corpus rows of Tables 13 and 14 use the target-cluster interval, and Table 15 lists all three. All three tables use paired bootstraps with $10{,}000$ resamples, and in every pooled row all three schemes are stratified by corpus: erased targets and representations are both nested in a corpus, so each corpus's own targets and representations are resampled within that corpus, and the two-way scheme forms only the 1220 observed target $\times$ representation cells. Interval widths are reported as measured and no ordering between the schemes is assumed. The per-target table counts wins rather than reporting a significance test, since targets within a corpus share representations and are not independent trials.

		AUROC drop				Balanced-accuracy drop			
Scope	rows	ρ coup.	ρ label	$\Delta\rho$	95% CI (target)	ρ coup.	ρ label	$\Delta\rho$	95% CI (target)
Logistic C=0.1									
Pooled	2832	0.508	0.278	+0.230	[+0.153, +0.301]	0.393	0.253	+0.140	[+0.089, +0.193]
CelebA	336	0.856	0.778	+0.079	[+0.045, +0.130]	0.824	0.738	+0.087	[+0.046, +0.151]
Bias-in-Bios	672	0.355	−0.017	+0.371	[+0.189, +0.540]	0.288	0.062	+0.226	[+0.044, +0.418]
EUR-Lex	1176	0.325	0.235	+0.090	[+0.040, +0.139]	0.197	0.138	+0.060	[+0.018, +0.094]
GoEmotions	648	0.471	0.081	+0.391	[+0.213, +0.523]	0.324	0.084	+0.240	[+0.140, +0.318]
Logistic C=1.0 *(canonical)*									
Pooled	2832	0.499	0.269	+0.230	[+0.151, +0.303]	0.375	0.246	+0.129	[+0.079, +0.175]
CelebA	336	0.856	0.777	+0.078	[+0.045, +0.129]	0.803	0.724	+0.080	[+0.037, +0.144]
Bias-in-Bios	672	0.346	−0.025	+0.371	[+0.175, +0.559]	0.270	0.069	+0.201	[+0.112, +0.272]
EUR-Lex	1176	0.294	0.216	+0.078	[+0.028, +0.123]	0.195	0.198	−0.004	[−0.029, +0.025]
GoEmotions	648	0.484	0.081	+0.403	[+0.221, +0.546]	0.307	0.041	+0.266	[+0.133, +0.360]
Logistic C=10									
Pooled	2832	0.481	0.244	+0.237	[+0.163, +0.303]	0.374	0.242	+0.132	[+0.085, +0.175]
CelebA	336	0.856	0.779	+0.078	[+0.045, +0.128]	0.801	0.721	+0.080	[+0.036, +0.144]
Bias-in-Bios	672	0.359	−0.002	+0.361	[+0.175, +0.534]	0.274	0.091	+0.183	[+0.104, +0.259]
EUR-Lex	1176	0.286	0.185	+0.102	[+0.050, +0.147]	0.202	0.168	+0.034	[−0.017, +0.075]
GoEmotions	648	0.415	0.029	+0.386	[+0.215, +0.512]	0.314	0.049	+0.265	[+0.141, +0.342]
Linear SVM C=1.0									
Pooled	2832	0.487	0.245	+0.242	[+0.163, +0.314]	0.349	0.242	+0.107	[+0.063, +0.151]
CelebA	336	0.856	0.782	+0.074	[+0.040, +0.124]	0.771	0.687	+0.084	[+0.040, +0.147]
Bias-in-Bios	672	0.356	−0.022	+0.377	[+0.188, +0.559]	0.272	0.134	+0.138	[+0.005, +0.277]
EUR-Lex	1176	0.274	0.157	+0.117	[+0.046, +0.179]	0.150	0.144	+0.006	[−0.026, +0.046]
GoEmotions	648	0.467	0.061	+0.406	[+0.222, +0.542]	0.328	0.074	+0.254	[+0.170, +0.313]

Table 16: Probe-family robustness for held-out AUROC and balanced-accuracy degradation. The reduced panel contains 8 approximately evenly spaced erased concepts and 3 representations per corpus (32 erased targets pooled), and is distinct from the expanded all-concept panel of Tables 13 and 14. It is smaller so that all four probe configurations could be fitted on the same erased embeddings, and it resamples fewer target clusters, so it does not carry the statistical power of the expanded panel. The bootstrap unit is the erased target (8 per corpus, 32 pooled), 10,000 paired resamples. The pooled target-cluster bootstrap is stratified by corpus: each corpus's own 8 erased targets are resampled within that corpus and the corpora are concatenated, with the same estimator and seed (42) as the expanded panel, and per-corpus rows resample that corpus's targets.

E.5 Probe family and ridge accessibility

We test the sensitivity of the comparison to the probe family and to regularisation of the accessibility estimator.

Table 16 repeats the comparison on a reduced panel under four probe conditions, logistic regression at $C \in \{0.1, 1, 10\}$ and a linear support vector machine, all fitted on the same erased embeddings. The pooled AUROC margin ranges from $+0.230$ to $+0.242$ across the four, a spread smaller than the width of any one of its intervals.

Table 17 replaces A_0 by the ridge accessibility A_λ in the predictor. The Schur identity does not hold for $\lambda > 0$, so these rows are heuristic and are reported empirically. On the AUROC drop the margin grows monotonically with λ, from $+0.247$ at $\lambda = 0$ to $+0.284$ at $\lambda_{\text{rel}} = 1$, whereas on the balanced-accuracy drop it does not grow. To test whether the dependence on the predictor ridge mirrors the regularisation of the downstream probe, we scanned the pooled Spearman ρ over the seven evaluated values of λ_{rel} under each probe configuration. On the AUROC drop the highest correlation is at the largest evaluated value, $\lambda_{\text{rel}} = 1$, under all four configurations, so the improvement does not come from matching the two regularisations. On the balanced-accuracy drop the correlation is nearly flat for small λ_{rel}, with its maximum at or near $\lambda = 0$ under all four conditions. The scan is descriptive and carries no uncertainty estimate.

	AUROC drop				Balanced-accuracy drop			
λ_{rel}	ρ coup.	ρ label	$\Delta\rho$	95% CI (target)	ρ coup.	ρ label	$\Delta\rho$	95% CI (target)
Pooled, 42,280 rows								
0 (A_0)	0.459	0.212	+0.247	[+0.213, +0.281]	0.303	0.206	+0.098	[+0.081, +0.114]
10^{-5}	0.459	0.212	+0.247	[+0.213, +0.281]	0.303	0.206	+0.098	[+0.081, +0.114]
10^{-4}	0.459	0.212	+0.247	[+0.213, +0.281]	0.303	0.206	+0.098	[+0.081, +0.114]
10^{-3}	0.460	0.212	+0.248	[+0.214, +0.282]	0.303	0.206	+0.097	[+0.081, +0.114]
10^{-2}	0.466	0.212	+0.254	[+0.220, +0.288]	0.303	0.206	+0.097	[+0.081, +0.114]
10^{-1}	0.482	0.212	+0.270	[+0.234, +0.305]	0.304	0.206	+0.099	[+0.082, +0.116]
10^{0}	0.496	0.212	+0.284	[+0.245, +0.322]	0.299	0.206	+0.093	[+0.075, +0.112]
CelebA, 2,100 rows								
0 (A_0)	0.770	0.743	+0.027	[+0.001, +0.056]	0.720	0.700	+0.020	[−0.015, +0.056]
10^{-5}	0.770	0.743	+0.027	[+0.001, +0.056]	0.720	0.700	+0.020	[−0.015, +0.056]
10^{-4}	0.769	0.743	+0.027	[+0.000, +0.056]	0.720	0.700	+0.020	[−0.015, +0.056]
10^{-3}	0.769	0.743	+0.026	[−0.0001, +0.055]	0.720	0.700	+0.019	[−0.016, +0.056]
10^{-2}	0.770	0.743	+0.027	[+0.000, +0.058]	0.721	0.700	+0.021	[−0.016, +0.059]
10^{-1}	0.774	0.743	+0.031	[+0.000, +0.065]	0.725	0.700	+0.025	[−0.016, +0.067]
10^{0}	0.775	0.743	+0.033	[−0.005, +0.072]	0.730	0.700	+0.029	[−0.021, +0.080]
Bias-in-Bios, 8,120 rows								
0 (A_0)	0.435	0.106	+0.329	[+0.241, +0.419]	0.326	0.136	+0.190	[+0.140, +0.238]
10^{-5}	0.435	0.106	+0.329	[+0.241, +0.419]	0.326	0.136	+0.190	[+0.140, +0.238]
10^{-4}	0.435	0.106	+0.330	[+0.242, +0.419]	0.326	0.136	+0.190	[+0.140, +0.238]
10^{-3}	0.438	0.106	+0.332	[+0.244, +0.422]	0.327	0.136	+0.191	[+0.141, +0.239]
10^{-2}	0.452	0.106	+0.346	[+0.258, +0.437]	0.331	0.136	+0.195	[+0.145, +0.244]
10^{-1}	0.492	0.106	+0.386	[+0.298, +0.476]	0.340	0.136	+0.203	[+0.154, +0.252]
10^{0}	0.562	0.106	+0.456	[+0.363, +0.549]	0.342	0.136	+0.205	[+0.153, +0.256]
EUR-Lex, 24,500 rows								
0 (A_0)	0.325	0.219	+0.106	[+0.085, +0.127]	0.186	0.172	+0.013	[+0.003, +0.024]
10^{-5}	0.325	0.219	+0.106	[+0.085, +0.128]	0.186	0.172	+0.013	[+0.003, +0.024]
10^{-4}	0.325	0.219	+0.106	[+0.085, +0.128]	0.185	0.172	+0.013	[+0.003, +0.024]
10^{-3}	0.329	0.219	+0.110	[+0.088, +0.132]	0.185	0.172	+0.013	[+0.002, +0.023]
10^{-2}	0.340	0.219	+0.121	[+0.098, +0.145]	0.185	0.172	+0.013	[+0.002, +0.024]
10^{-1}	0.362	0.219	+0.143	[+0.115, +0.170]	0.188	0.172	+0.015	[+0.002, +0.029]
10^{0}	0.375	0.219	+0.156	[+0.121, +0.189]	0.186	0.172	+0.013	[−0.002, +0.030]
GoEmotions, 7,560 rows								
0 (A_0)	0.566	0.100	+0.466	[+0.379, +0.556]	0.310	0.093	+0.217	[+0.174, +0.262]
10^{-5}	0.566	0.100	+0.466	[+0.379, +0.556]	0.310	0.093	+0.217	[+0.174, +0.262]
10^{-4}	0.566	0.100	+0.466	[+0.379, +0.556]	0.310	0.093	+0.217	[+0.174, +0.262]
10^{-3}	0.566	0.100	+0.466	[+0.380, +0.556]	0.310	0.093	+0.217	[+0.174, +0.262]
10^{-2}	0.569	0.100	+0.469	[+0.382, +0.559]	0.311	0.093	+0.218	[+0.175, +0.264]
10^{-1}	0.574	0.100	+0.474	[+0.385, +0.567]	0.311	0.093	+0.218	[+0.175, +0.263]
10^{0}	0.568	0.100	+0.468	[+0.373, +0.567]	0.303	0.093	+0.210	[+0.164, +0.256]

Table 17: Ridge accessibility as a predictor, against both held-out metrics. The predictor is $\left(A_\lambda[i,j]/\sqrt{A_\lambda[i,i]A_\lambda[j,j]}\right)^2$, computed from the pre-erasure training data, and $\lambda_{\text{rel}} = 0$ recovers A_0. The Schur identity of A_0 does not extend to $\lambda > 0$, so for positive λ the predictor has no exactness guarantee and is evaluated empirically. The label-correlation column is the same baseline in every row. Paired target-cluster bootstrap with $10{,}000$ resamples, stratified by corpus in the pooled rows, so the pooled row at $\lambda_{\text{rel}} = 0$ reproduces the pooled target-cluster interval of Table 13. Row counts are given in the group headings.

E.6 Screening of candidate erasure sets

Table 18 lists the ten CelebA triples selected from the 455 by the A_0-only score, together with the held-out validation results, including the representation-by-set correlations.

We fix $|S| = 3$ to match the cardinality of the three multi-concept CelebA erasure sets in the original evaluation grid, so that the experiment varies set composition while holding intervention size fixed.

The ten candidate sets were fixed from the pre-erasure A_0-based ranking before any validation erasure was evaluated, so they were not chosen with knowledge of the outcome. The experiment evaluates the ordering of candidate sets rather than the calibration of predicted magnitudes, since the sample is five sets against five. The scoring time covers only the Schur computation over the 455 triples and excludes the one-time construction of A_0, which any other analysis of the same representation also requires. Most of the erasure time, 96%, is spent fitting the twelve retained-concept probes rather than the eraser itself.

(a) The 10 pre-specified candidate sets

Rank	Erasure set S (3 concepts)	pred. abs. loss	pred. rel. loss
5 lowest predicted collateral			
1	`Eyeglasses, Wearing_Hat, Bald`	0.0701	0.1564
2	`Eyeglasses, Wearing_Hat, Blond_Hair`	0.0710	0.1384
3	`Smiling, Wearing_Hat, High_Cheekbones`	0.0809	0.1382
4	`Eyeglasses, Young, Bald`	0.0846	0.1791
5	`Eyeglasses, Smiling, Wearing_Hat`	0.0870	0.1766
5 highest predicted collateral			
451	`Goatee, Heavy_Makeup, High_Cheekbones`	0.2644	0.4832
452	`Smiling, Mustache, Wearing_Lipstick`	0.2661	0.4863
453	`Goatee, Wearing_Lipstick, High_Cheekbones`	0.2681	0.4950
454	`Mustache, Heavy_Makeup, High_Cheekbones`	0.2718	0.4792
455	`Mustache, Wearing_Lipstick, High_Cheekbones`	0.2751	0.4900

(b) Held-out validation after a real LEACE erasure

Held-out outcome	low group	high group	high/low
mean retained AUROC drop	0.0245	0.1289	5.26×
max retained AUROC drop	0.1089	0.3616	3.32×
mean retained bal. acc. drop	0.0263	0.1237	4.70×
max retained bal. acc. drop	0.1032	0.3305	3.20×
Spearman(pred., held-out mean AUROC drop), 10 candidate sets			$\rho = 0.794$
Spearman(pred., held-out mean bal. acc. drop), 10 candidate sets			$\rho = 0.806$
Descriptive: Spearman over 100 rep×set pairs (not independent)			$\rho = 0.746$ / 0.715

Table 18: Candidate-set screening on CelebA. **(a)** ten candidate triples selected from all $\binom{15}{3} = 455$ sets using only the pre-erasure A_0 Schur screening score $\mathrm{mean}_{j \in T}\left[(A_0)_{jj} - (A_{\mathrm{pred}})_{jj}\right]$. No erasure and no downstream probe is used to construct this ranking. The two loss columns are the predicted absolute and relative losses, averaged over the retained concepts and the ten representations. The candidate sets were fixed before any validation erasure was evaluated. **(b)** held-out outcomes after LEACE erasure for the five lowest- and five highest-scoring groups of panel (a), averaged over the same ten CelebA representations. Group means and ratios are descriptive. Set-level Spearman correlations use the ten candidate sets, and correlations over representation-by-set entries are reported descriptively because those observations are not independent.

E.7 Normalisation of the R^2 comparison

Corollary 2 makes $(\tilde{A}_0)^2_{ij}$ the relative loss, the fraction of concept j's own accessibility that erasing i removes. Comparing it with an absolute R^2 drop omits the factor $(A_0)_{jj}$, so the two quantities are on different scales. The R^2 block of Table 12(a) makes that comparison, although Table 19 shows that on the single-concept panel the difference in scale alone does not reverse the result.

With matched scales the agreement is algebraic rather than empirical. Over the same $42{,}280$ rows the relative drop equals $(\tilde{A}_0)^2_{ij}$ to a mean absolute error of 1.7×10^{-13} for the observed eraser (maximum 1.0×10^{-10}, with a single row above 10^{-10}) and 1.6×10^{-16} for the Schur prediction (maximum 3.4×10^{-15}, no row above 10^{-10}). Per corpus the observed mean absolute error ranges from 3.2×10^{-15} on CelebA to 8.8×10^{-13} on GoEmotions. The two algebraic rearrangements $\Delta R^2 = (\tilde{A}_0)^2_{ij}(A_0)_{jj}$ and $\Delta R^2 = (A_0)^2_{ij}/(A_0)_{ii}$ hold to floating-point zero, at mean absolute errors of 3.8×10^{-17} and 1.5×10^{-17}.

Table 19 compares the two normalisations. With matched scaling the predictor reproduces the absolute R^2 drop, as Corollary 2 requires. With unmatched scaling it still has the higher pooled Spearman correlation, 0.908 against 0.721 for label correlation.

Scope	rows	vs *relative* R^2 drop			vs *absolute* R^2 drop		
		ρ coupling	ρ label	$\Delta\rho$	ρ coupling	ρ label	$\Delta\rho$
Pooled	42280	1.0000	0.6007	+0.3993	0.9078	0.7210	+0.1868
CelebA	2100	1.0000	0.9047	+0.0953	0.9171	0.9440	−0.0269
Bias-in-Bios	8120	1.0000	0.4195	+0.5805	0.9422	0.5361	+0.4061
EUR-Lex	24500	1.0000	0.7359	+0.2641	0.9219	0.8097	+0.1122
GoEmotions	7560	1.0000	0.2643	+0.7357	0.9303	0.3691	+0.5612

Table 19: Relative and absolute R^2-loss comparisons. Squared normalised accessible coupling $(\tilde{A}_0)^2_{ij}$ corresponds to the relative accessibility loss of Corollary 2. The left block compares predictors with the relative R^2 loss, and the right block reports the corresponding comparison with the absolute R^2 loss, which differs from the left by the factor $(A_0)_{jj}$.

E.8 Edit footprints

Figure 8 plots the per-coordinate leverage ℓ_i of Proposition 1, sorted in decreasing order, for two single-concept erasures at the concentrated and distributed ends of the range defined in Section 3.3. The identity also holds for multi-concept sets. For the three-concept set `gender_makeup` (`Male`, `Heavy_Makeup` and `Wearing_Lipstick`) on `clip-vitb16`, where the erased subspace has rank three, predicted leverage matches the measured per-coordinate mean squared edit to a maximum deviation of 7.8×10^{-16} on the fitting split. On the held-out split, where the whitening is no longer exact, the equality holds only approximately: the maximum deviation is 0.020 for this set, with a median of 2.7×10^{-3} and a maximum of 0.13 over all $3{,}564$ footprints, while the ordering of the coordinates is preserved (rank correlation 0.9995 here, with a median of 1.000 over all footprints).

The erasure coherence μ_{er} spans 9.8 to 1696 across the $3{,}564$ footprints, with a median of 77.0, indicating substantial variation in footprint concentration. Concentration is associated with prevalence: frequent, well-encoded concepts occupy a tight subspace and rare ones are spread across the representation. The full per-attribute ranking is Table 27.

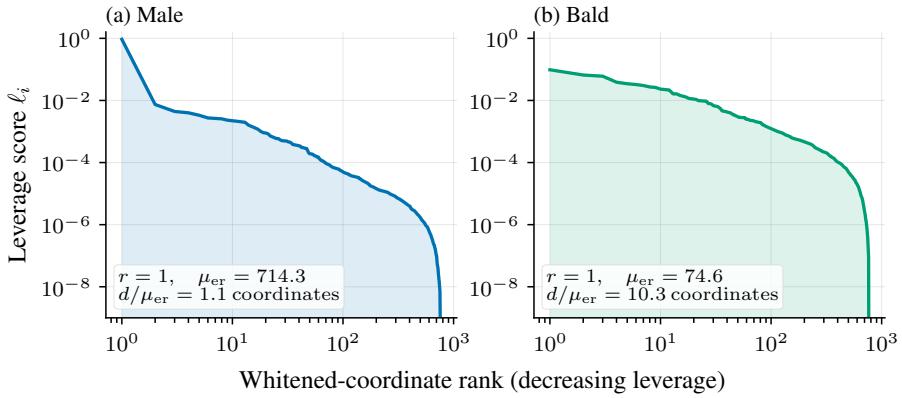

Figure 8: LEACE edit footprints range from concentrated to distributed. Sorted leverage profiles ℓ_i for two CelebA erasures on `clip-vitb16`, with whitened coordinates ordered by decreasing leverage and both axes logarithmic. **(a)**: `Male` concentrates essentially all leverage in the first few whitened coordinates, with $r = 1$, $\mu_{\text{er}} = 714.3$ and $d/\mu_{\text{er}} = 1.1$ effective whitened coordinates. **(b)**: `Bald` spreads the same rank-one edit over $d/\mu_{\text{er}} = 10.3$ effective whitened coordinates, with $\mu_{\text{er}} = 74.6$.

F DATASET, MODEL, AND RESULT INVENTORIES

This appendix lists the reference inventories: the CelebA attribute set (Table 20), the model sets (Tables 21 and 22), the per-representation results (Tables 23 to 26), and the extremes of the erasure-coherence ranking (Table 27). Model names are given as they appear in the loader registries.

F.1 THE CELEBA ATTRIBUTE SET

Table 20 lists the 15 CelebA attributes used throughout, with the cluster each belongs to, its training prevalence, and its accessibility R^2 and probe balanced accuracy averaged over the 45 backbones.

Attribute	**Cluster**	prevalence	R^2	bal. acc.
`Male`	gender presentation	0.4194	0.8672	0.9733
`Heavy_Makeup`	gender presentation	0.3843	0.6586	0.9033
`Wearing_Lipstick`	gender presentation	0.4696	0.7342	0.9150
`Mustache`	facial hair	0.0408	0.2544	0.8957
`Goatee`	facial hair	0.0635	0.4271	0.9204
`No_Beard`	facial hair	0.8342	0.6052	0.9172
`Young`	age and appearance	0.7789	0.4759	0.8239
`Attractive`	age and appearance	0.5136	0.4706	0.7976
`High_Cheekbones`	age and appearance	0.4524	0.5304	0.8459
`Chubby`	age and appearance	0.0577	0.2850	0.8655
`Eyeglasses`	control	0.0646	0.7699	0.9693
`Wearing_Hat`	control	0.0494	0.7187	0.9688
`Bald`	control	0.0228	0.3769	0.9424
`Blond_Hair`	control	0.1491	0.6145	0.9255
`Smiling`	expression	0.4797	0.6386	0.8949

Table 20: The 15 CelebA attributes, their cluster, their training prevalence, and their unregularised accessibility $R^2 = (A_0)_{jj}$ and probe balanced accuracy averaged over the 45 backbones. The control attributes are correlated with `Male` in the labels but have an independent visual signature, which makes them a useful check: a diagnostic that merely reproduced label correlation would score them like the gender-presentation cluster.

F.2 THE MODEL SETS

Table 21: The 45-backbone vision model set used on CelebA. "Loader" is `timm` for a pretrained tag and `timm_ckpt` where an external checkpoint is loaded into a bare architecture.

Backbone	Family	timm architecture	Loader
clip-vitb32	clip	vit_base_patch32_clip_224.openai	timm
clip-vitb16	clip	vit_base_patch16_clip_224.openai	timm
clip-vitl14	clip	vit_large_patch14_clip_224.openai	timm
clip-rn50	clip	resnet50_clip.openai	timm
resnet-resnet50	resnet	resnet50.tv_in1k	timm
dinov2-vits14	dinov2	vit_small_patch14_dinov2.lvd142m	timm
dinov2-vitb14	dinov2	vit_base_patch14_dinov2.lvd142m	timm
dinov2-vitl14	dinov2	vit_large_patch14_dinov2.lvd142m	timm
dinov2-vits14_reg	dinov2	vit_small_patch14_reg4_dinov2.lvd142m	timm
dinov2-vitb14_reg	dinov2	vit_base_patch14_reg4_dinov2.lvd142m	timm
dinov2-vitl14_reg	dinov2	vit_large_patch14_reg4_dinov2.lvd142m	timm
dino-vits16	dino	vit_small_patch16_224.dino	timm
dino-vits8	dino	vit_small_patch8_224.dino	timm
dino-vitb16	dino	vit_base_patch16_224.dino	timm
dino-vitb8	dino	vit_base_patch8_224.dino	timm
mae-base	mae	vit_base_patch16_224.mae	timm
mae-large	mae	vit_large_patch16_224.mae	timm
mae-huge	mae	vit_huge_patch14_224.mae	timm
vit-s16	vit_sup	vit_small_patch16_224.augreg_in1k	timm
vit-b16	vit_sup	vit_base_patch16_224.augreg_in1k	timm
vit-l16	vit_sup	vit_large_patch16_224.augreg_in21k	timm
beit_base	timm_ssl	beit_base_patch16_224.in22k_ft_in22k	timm
beitv2_base	timm_ssl	beitv2_base_patch16_224.in1k_ft_in1k	timm
eva02_base	timm_ssl	eva02_base_patch14_224.mim_in22k	timm
eva02_large	timm_ssl	eva02_large_patch14_224.mim_in22k	timm
resnext_ssl	timm_ssl	resnext101_32x16d.fb_ssl_yfcc100m_ft_in1k	timm

continued on next page

continued from previous page

Backbone	Family	timm architecture	Loader
resmlp_dino	timm_ssl	resmlp_12_224.fb_dino	timm
samvit_base	timm_ssl	samvit_base_patch16.sa1b	timm
aimv2_1b	timm_ssl	aimv2_1b_patch14_224.apple_pt	timm
vitamin_clip	timm_ssl	vitamin_base_224.datacomp1b_clip	timm
siglip_base_gap	timm_ssl	vit_base_patch16_siglip_gap_224	timm
siglip_so400m_gap	timm_ssl	vit_so400m_patch14_siglip_gap_224	timm
convnextv2_base_fcmae	timm_ssl	convnextv2_base.fcmae_ft_in22k_in1k	timm
convnextv2_large_fcmae	timm_ssl	convnextv2_large.fcmae_ft_in22k_in1k	timm
sam2_hiera_tiny	dense_seg	sam2_hiera_tiny.fb_r896	timm
sam2_hiera_small	dense_seg	sam2_hiera_small.fb_r896	timm
sam2_hiera	dense_seg	sam2_hiera_base_plus.fb_r896	timm
sam2_hiera_large	dense_seg	sam2_hiera_large.fb_r1024	timm
hiera_mae	dense_seg	hiera_base_224.mae	timm
dinov3-vitb	dinov3	vit_base_patch16_dinov3.lvd1689m	timm
dinov3-vitl	dinov3	vit_large_patch16_dinov3.lvd1689m	timm
dinov3-convnext_base	dinov3	convnext_base.dinov3_lvd1689m	timm
dinov3-convnext_tiny	dinov3	convnext_tiny.dinov3_lvd1689m	timm
mocov3-vitb	mocov3	vit_base_patch16_224	timm_ckpt
simclr-rn50	simclr	resnet50	timm_ckpt

Encoder	d	**Encoder**	d
Alibaba-NLP/gte-base-en-v1.5	768	intfloat/e5-base-v2	768
Alibaba-NLP/gte-large-en-v1.5	1024	intfloat/e5-large-v2	1024
BAAI/bge-base-en-v1.5	768	intfloat/multilingual-e5-base	768
BAAI/bge-large-en-v1.5	1024	intfloat/multilingual-e5-large	1024
BAAI/bge-m3	1024	microsoft/deberta-v3-base	768
BAAI/bge-small-en-v1.5	384	mixedbread-ai/mxbai-embed-large-v1	1024
Snowflake/snowflake-arctic-embed-m-v1.5	768	nomic-ai/nomic-embed-text-v1.5	768
bert-base-uncased	768	roberta-base	768
distilbert-base-uncased	768	sentence-transformers/all-MiniLM-L6-v2	384
google/electra-base-discriminator	768	sentence-transformers/all-mpnet-base-v2	768
hkunlp/instructor-base	768	sentence-transformers/gtr-t5-base	768
hotchpotch/bekko-embedding-v1-a8m	384	sentence-transformers/sentence-t5-base	768

Table 22: The 24-encoder text model set, cached identically on all three text datasets, listed in two columns. Names are the exact Hugging Face identifiers, which Tables 24 to 26 abbreviate by dropping the organisation prefix.

F.3 PER-REPRESENTATION RESULTS

Complete per-representation results, sorted by balanced accuracy. $\mathrm{erank}(L)$ is the effective rank of the probe-logit covariance and d is the embedding width. The columns bal, AUROC, F_1 and AP are the balanced accuracy, AUROC, macro F_1 and average precision of the baseline probes, averaged over concepts. $\mathrm{tr}(A)/m$ and $\mathrm{erank}\,A$ are the trace per concept and the effective rank of the ridge accessibility A_λ at $\lambda_{\mathrm{rel}} = 10^{-2}$, and $\mathrm{AUR}_{\mathrm{cal}}$ is the mean calibrated AUROC of Equation (5). Text encoders are named without their organisation prefix, which Table 22 supplies, while vision backbones carry the registry names of Table 21.

Table 23: CelebA: all 45 representations, sorted by balanced accuracy.

Model	d	bal	AUROC	F_1	AP	$\mathrm{tr}(A)/m$	$\mathrm{erank}\,A$	$\mathrm{AUR}_{\mathrm{cal}}$	$\mathrm{erank}\,L$
siglip_so400m_gap	1152	0.9262	0.9741	0.8746	0.8958	0.6548	7.56	0.9884	4.98
siglip_base_gap	768	0.9257	0.9731	0.8698	0.8902	0.6420	7.38	0.9866	4.70
vitamin_clip	768	0.9251	0.9734	0.8698	0.8938	0.6285	7.27	0.9864	4.65
clip-vitb16	768	0.9246	0.9726	0.8664	0.8881	0.6206	7.21	0.9864	4.42
eva02_base	768	0.9243	0.9724	0.8694	0.8894	0.6415	7.43	0.9868	4.43
clip-vitb32	768	0.9242	0.9725	0.8684	0.8878	0.6308	7.30	0.9867	4.40
dinov3-vitb	768	0.9237	0.9724	0.8660	0.8865	0.6126	7.06	0.9849	4.89
dinov3-vitl	1024	0.9230	0.9722	0.8689	0.8825	0.6229	7.21	0.9863	4.96
clip-vitl14	1024	0.9206	0.9708	0.8649	0.8796	0.6167	7.21	0.9866	4.50
vit-l16	1024	0.9204	0.9710	0.8629	0.8835	0.6131	7.10	0.9848	4.92
beit_base	768	0.9202	0.9706	0.8614	0.8733	0.6156	7.16	0.9848	4.78
eva02_large	1024	0.9197	0.9710	0.8663	0.8874	0.6474	7.49	0.9873	4.68
clip-rn50	2048	0.9193	0.9723	0.8738	0.8886	0.6663	7.68	0.9892	5.02
aimv2_1b	2048	0.9175	0.9712	0.8715	0.8852	0.6431	7.53	0.9881	5.14
mae-large	1024	0.9171	0.9690	0.8587	0.8746	0.5770	6.62	0.9815	5.08

continued on next page

continued from previous page

Model	d	bal	AUROC	F_1	AP	$\mathrm{tr}(A)/m$	erank A	$\mathrm{AUR}_{\mathrm{cal}}$	erank L
mae-huge	1280	0.9165	0.9696	0.8624	0.8781	0.5884	6.74	0.9831	5.23
dinov2-vitl14_reg	1024	0.9165	0.9681	0.8531	0.8636	0.5789	6.66	0.9829	4.76
dinov2-vitb14_reg	768	0.9163	0.9681	0.8520	0.8631	0.5759	6.56	0.9819	4.42
dinov2-vitl14	1024	0.9160	0.9677	0.8513	0.8639	0.5877	6.75	0.9836	4.66
dinov2-vitb14	768	0.9156	0.9677	0.8504	0.8590	0.5785	6.61	0.9819	4.58
convnextv2_base_fcmae	1024	0.9149	0.9674	0.8540	0.8632	0.5972	6.99	0.9839	4.77
dinov2-vits14	384	0.9144	0.9664	0.8450	0.8647	0.5530	6.29	0.9776	4.14
dinov2-vits14_reg	384	0.9142	0.9664	0.8442	0.8574	0.5508	6.24	0.9771	4.19
convnextv2_large_fcmae	1536	0.9130	0.9676	0.8577	0.8669	0.6097	7.15	0.9848	5.08
dinov3-convnext_tiny	768	0.9125	0.9653	0.8458	0.8559	0.5598	6.66	0.9791	4.78
beitv2_base	768	0.9101	0.9638	0.8435	0.8524	0.5548	6.42	0.9778	4.18
dinov3-convnext_base	1024	0.9093	0.9628	0.8414	0.8481	0.5398	6.55	0.9778	4.73
dino-vitb8	768	0.9091	0.9638	0.8436	0.8573	0.5466	6.32	0.9746	4.41
dino-vitb16	768	0.9084	0.9634	0.8428	0.8594	0.5432	6.33	0.9740	4.66
mocov3-vitb	768	0.9084	0.9634	0.8426	0.8579	0.5427	6.29	0.9750	4.49
hiera_mae	768	0.9078	0.9627	0.8418	0.8606	0.5380	6.21	0.9754	4.73
mae-base	768	0.9056	0.9619	0.8379	0.8572	0.5162	6.12	0.9730	4.99
resmlp_dino	384	0.9040	0.9594	0.8345	0.8578	0.5117	6.17	0.9685	4.61
dino-vits16	384	0.8994	0.9570	0.8239	0.8426	0.5011	5.90	0.9646	4.22
dino-vits8	384	0.8993	0.9565	0.8255	0.8429	0.5009	5.89	0.9641	4.27
sam2_hiera_large	1152	0.8984	0.9570	0.8353	0.8493	0.4934	6.11	0.9693	5.14
resnext_ssl	2048	0.8961	0.9581	0.8353	0.8368	0.5488	6.46	0.9782	4.51
resnet-resnet50	2048	0.8877	0.9524	0.8281	0.8345	0.5209	6.37	0.9714	4.91
sam2_hiera_small	768	0.8842	0.9461	0.8143	0.8343	0.4402	5.76	0.9585	5.01
vit-b16	768	0.8820	0.9447	0.8051	0.8168	0.4676	5.74	0.9579	4.18
sam2_hiera_tiny	768	0.8800	0.9431	0.8066	0.8199	0.4206	5.71	0.9544	4.99
vit-s16	384	0.8747	0.9381	0.7902	0.8046	0.4464	5.48	0.9497	3.70
sam2_hiera	896	0.8711	0.9376	0.7950	0.8088	0.4017	5.46	0.9495	4.94
samvit_base	256	0.8360	0.9089	0.7454	0.7450	0.2592	4.17	0.9173	4.13
simclr-rn50	2048	0.7229	0.7884	0.6383	0.5499	0.1746	4.98	0.8280	3.39

Table 24: Bias-in-Bios: all 24 representations, sorted by balanced accuracy.

Model	d	bal	AUROC	F_1	AP	$\mathrm{tr}(A)/m$	erank A	$\mathrm{AUR}_{\mathrm{cal}}$	erank L
snowflake-arctic-embed-m-v1.5	768	0.9380	0.9853	0.7608	0.7682	0.4000	22.93	0.9990	12.69
bge-small-en-v1.5	384	0.9375	0.9817	0.7140	0.7236	0.3367	21.40	0.9974	12.60
e5-base-v2	768	0.9371	0.9858	0.7644	0.7656	0.4058	22.94	0.9990	11.95
mxbai-embed-large-v1	1024	0.9370	0.9862	0.7843	0.7834	0.4555	24.28	0.9994	12.55
bge-large-en-v1.5	1024	0.9369	0.9858	0.7813	0.7800	0.4309	23.81	0.9992	13.50
bekko-embedding-v1-a8m	384	0.9364	0.9784	0.7112	0.6696	0.3155	21.49	0.9960	9.70
nomic-embed-text-v1.5	768	0.9362	0.9843	0.7554	0.7568	0.4114	23.39	0.9989	12.91
e5-large-v2	1024	0.9359	0.9862	0.7820	0.7782	0.4232	23.43	0.9992	11.82
bge-base-en-v1.5	768	0.9358	0.9840	0.7561	0.7558	0.3924	22.96	0.9988	13.06
instructor-base	768	0.9351	0.9800	0.7563	0.7193	0.3804	23.21	0.9980	9.63
distilbert-base-uncased	768	0.9347	0.9833	0.7491	0.7453	0.4097	23.19	0.9989	10.74
multilingual-e5-base	768	0.9342	0.9832	0.7522	0.7393	0.3723	22.25	0.9985	11.50
sentence-t5-base	768	0.9342	0.9792	0.7590	0.7196	0.3953	23.74	0.9981	8.96
bert-base-uncased	768	0.9337	0.9839	0.7492	0.7404	0.4154	23.35	0.9990	10.87
roberta-base	768	0.9332	0.9837	0.7461	0.7473	0.4075	23.33	0.9989	11.02
multilingual-e5-large	1024	0.9331	0.9853	0.7759	0.7660	0.4061	23.03	0.9990	12.59
all-mpnet-base-v2	768	0.9328	0.9832	0.7475	0.7492	0.4158	23.84	0.9986	12.50
all-MiniLM-L6-v2	384	0.9325	0.9776	0.7003	0.6700	0.3253	21.75	0.9964	11.41
gtr-t5-base	768	0.9307	0.9785	0.7425	0.7076	0.3618	23.05	0.9973	9.28
bge-m3	1024	0.9304	0.9840	0.7635	0.7499	0.3989	23.03	0.9989	13.00
deberta-v3-base	768	0.8915	0.9604	0.6473	0.5087	0.2245	17.14	0.9878	11.53
gte-base-en-v1.5	768	0.8899	0.9555	0.6440	0.5336	0.2140	17.91	0.9883	10.98
electra-base-discriminator	768	0.8883	0.9573	0.6444	0.5066	0.2203	17.14	0.9868	10.67
gte-large-en-v1.5	1024	0.7463	0.8038	0.5139	0.1644	0.0317	11.59	0.8166	2.35

Table 25: EUR-Lex: all 24 representations, sorted by balanced accuracy.

Model	d	bal	AUROC	F_1	AP	$\mathrm{tr}(A)/m$	erank A	$\mathrm{AUR}_{\mathrm{cal}}$	erank L
all-MiniLM-L6-v2	384	0.9363	0.9761	0.7707	0.7084	0.4262	20.94	0.9960	16.07
nomic-embed-text-v1.5	768	0.9350	0.9786	0.8193	0.7521	0.4910	23.83	0.9988	20.20
bge-small-en-v1.5	384	0.9345	0.9759	0.7727	0.7082	0.4233	21.04	0.9960	17.87
e5-base-v2	768	0.9337	0.9807	0.8284	0.7564	0.4983	24.16	0.9988	19.92
all-mpnet-base-v2	768	0.9336	0.9760	0.8119	0.7477	0.4667	23.16	0.9983	15.63
distilbert-base-uncased	768	0.9318	0.9772	0.8104	0.7369	0.4349	22.16	0.9984	18.13
bge-base-en-v1.5	768	0.9317	0.9805	0.8259	0.7518	0.4906	23.71	0.9986	20.53
sentence-t5-base	768	0.9314	0.9772	0.8153	0.7466	0.4871	23.90	0.9986	19.13
snowflake-arctic-embed-m-v1.5	768	0.9310	0.9775	0.8136	0.7493	0.4629	23.28	0.9986	17.37
multilingual-e5-base	768	0.9304	0.9794	0.8252	0.7523	0.4923	24.10	0.9986	21.04
bekko-embedding-v1-a8m	384	0.9297	0.9736	0.7626	0.6988	0.4140	21.24	0.9950	19.06
bert-base-uncased	768	0.9290	0.9766	0.8110	0.7317	0.4416	22.39	0.9982	19.87

continued on next page

continued from previous page

Model	d	bal	AUROC	F_1	AP	$\mathrm{tr}(A)/m$	erank A	$\mathrm{AUR}_{\mathrm{cal}}$	erank L
bge-large-en-v1.5	1024	0.9289	0.9814	0.8409	0.7725	0.5109	24.61	0.9992	21.62
instructor-base	768	0.9282	0.9775	0.8151	0.7408	0.4849	23.99	0.9986	20.15
gtr-t5-base	768	0.9281	0.9762	0.8124	0.7397	0.4919	24.09	0.9986	18.05
multilingual-e5-large	1024	0.9266	0.9785	0.8338	0.7628	0.5073	24.76	0.9992	22.73
mxbai-embed-large-v1	1024	0.9262	0.9782	0.8304	0.7597	0.4983	24.27	0.9991	21.33
roberta-base	768	0.9251	0.9743	0.8050	0.7244	0.4318	21.70	0.9982	19.89
e5-large-v2	1024	0.9251	0.9812	0.8442	0.7639	0.5201	25.20	0.9992	24.91
bge-m3	1024	0.9219	0.9772	0.8220	0.7524	0.4745	23.65	0.9988	20.99
deberta-v3-base	768	0.9090	0.9598	0.7554	0.6396	0.3332	17.38	0.9941	19.42
gte-base-en-v1.5	768	0.9089	0.9582	0.7464	0.6365	0.3019	16.71	0.9932	18.31
electra-base-discriminator	768	0.9081	0.9598	0.7531	0.6287	0.3368	17.25	0.9938	20.51
gte-large-en-v1.5	1024	0.8506	0.9063	0.6436	0.4547	0.1552	8.93	0.9654	2.68

Table 26: GoEmotions: all 24 representations, sorted by balanced accuracy.

Model	d	bal	AUROC	F_1	AP	$\mathrm{tr}(A)/m$	erank A	$\mathrm{AUR}_{\mathrm{cal}}$	erank L
bge-small-en-v1.5	384	0.7933	0.8886	0.5978	0.3491	0.1571	17.60	0.9485	9.24
e5-base-v2	768	0.7878	0.8988	0.6264	0.3566	0.1876	18.87	0.9632	10.35
sentence-t5-base	768	0.7873	0.8888	0.6329	0.3534	0.1882	19.45	0.9628	6.25
bekko-embedding-v1-a8m	384	0.7848	0.8682	0.5894	0.3087	0.1428	17.30	0.9339	5.78
multilingual-e5-base	768	0.7837	0.8918	0.6263	0.3550	0.1891	18.64	0.9642	10.58
instructor-base	768	0.7829	0.8839	0.6277	0.3545	0.1812	19.32	0.9607	6.14
all-MiniLM-L6-v2	384	0.7819	0.8646	0.5783	0.2922	0.1255	16.69	0.9265	8.35
roberta-base	768	0.7795	0.8816	0.6112	0.3352	0.1702	18.33	0.9579	10.39
gtr-t5-base	768	0.7785	0.8653	0.6159	0.3332	0.1596	19.10	0.9505	5.32
nomic-embed-text-v1.5	768	0.7785	0.8850	0.6183	0.3562	0.1843	18.91	0.9617	11.27
bge-m3	1024	0.7781	0.8970	0.6379	0.3614	0.2037	18.92	0.9707	13.13
bge-base-en-v1.5	768	0.7767	0.8871	0.6193	0.3536	0.1849	18.91	0.9637	11.54
distilbert-base-uncased	768	0.7755	0.8733	0.6127	0.3345	0.1698	17.68	0.9577	11.94
bert-base-uncased	768	0.7695	0.8710	0.6056	0.3236	0.1632	17.81	0.9548	11.18
bge-large-en-v1.5	1024	0.7685	0.8855	0.6268	0.3528	0.1920	19.34	0.9677	13.09
mxbai-embed-large-v1	1024	0.7674	0.8849	0.6247	0.3551	0.1895	19.27	0.9671	13.56
multilingual-e5-large	1024	0.7666	0.8830	0.6310	0.3365	0.1939	19.22	0.9661	12.12
snowflake-arctic-embed-m-v1.5	768	0.7647	0.8635	0.6029	0.3116	0.1580	18.79	0.9540	10.76
all-mpnet-base-v2	768	0.7592	0.8606	0.5912	0.2755	0.1476	18.38	0.9458	10.72
e5-large-v2	1024	0.7564	0.8747	0.6187	0.3285	0.1865	19.42	0.9640	12.11
deberta-v3-base	768	0.7557	0.8625	0.5895	0.2863	0.1446	16.95	0.9459	10.53
electra-base-discriminator	768	0.7501	0.8490	0.5877	0.2881	0.1432	16.71	0.9439	10.14
gte-base-en-v1.5	768	0.7065	0.7881	0.5421	0.2099	0.0937	15.96	0.9034	7.11
gte-large-en-v1.5	1024	0.5875	0.6258	0.4594	0.0892	0.0311	17.34	0.8199	2.51

F.4 Erasure-coherence extremes across single-concept erasures

Ranked from most concentrated to most distributed. Both μ_{er} and the effective coordinate count are means over the representations of the dataset, and d is the dataset-mean whitened dimension. The effective coordinate count is the mean of d/μ_{er}, not the ratio of the means, so it is not d divided by μ_{er}. Table 27 prints the two extremes of the 122-concept ranking.

Rank	**Dataset**	Attribute	μ_{er}	eff. coords	d
10 most concentrated					
1	CelebA	Wearing_Lipstick	628.1	2.26	946
2	CelebA	Male	623.7	2.37	946
3	CelebA	Heavy_Makeup	558.8	2.43	946
4	Bias-in-Bios	gender	479.8	4.17	783
5	CelebA	Attractive	407.9	2.91	946
6	Bias-in-Bios	prof_21	348.4	2.90	783
7	Bias-in-Bios	prof_2	308.7	6.14	783
8	Bias-in-Bios	prof_19	285.6	3.24	783
9	CelebA	No_Beard	264.1	4.49	946
10	CelebA	Young	263.2	4.47	946
Discussed example outside the extrema					
52	CelebA	Bald	87.4	11.12	946
10 most distributed					
113	EUR-Lex	eurlex_5283	29.6	29.35	781
114	Bias-in-Bios	prof_23	28.7	28.53	783
115	Bias-in-Bios	prof_3	28.0	30.35	783
116	EUR-Lex	eurlex_2734	28.0	28.98	781
117	Bias-in-Bios	prof_10	28.0	30.24	783
118	GoEmotions	nervousness	28.0	32.54	783
119	EUR-Lex	eurlex_3774	25.5	32.14	781
120	GoEmotions	grief	25.4	34.04	783
121	GoEmotions	embarrassment	24.0	35.70	783
122	GoEmotions	pride	24.0	36.08	783

Table 27: Erasure coherence for single-concept erasures: the 10 most concentrated and the 10 most distributed of the 122 ranked concepts, together with `Bald`, the distributed example of Figure 8. Rank is the position in the full 122-concept ranking.