When Does Concept Erasure Change Behavior?

Gagan Aryan
Independent Researcher
gaganaryan19@gmail.com

Abstract

Concept erasure can remove linearly recoverable information from a representation without showing that this information influenced the model output being evaluated. For linear edits followed by explicit linear output maps, we characterize the resulting logit change using the output map and representation covariance. We then compare downstream effects with control edits matched for validation-side damage to test whether any benefit depends on the targeted concept. On Bios, Least-squares Concept Erasure (LEACE) reduces gender-probe accuracy to chance but changes only about 0.4% of occupation-logit energy and yields no detectable improvement in true positive rate (TPR) gaps beyond shuffled-label controls. Iterative Nullspace Projection (INLP) changes more occupation-logit energy at higher ranks, but its small, uncertain fairness gains are accompanied by occupation-accuracy loss. A controlled text shortcut yields a large target-specific gain when the classifier relies on the erased marker. In final-state, next-token audits of TinyLlama and Qwen2.5-7B, gender erasure substantially reduces *he/she* margins but leaves occupation-token choices intact. These results show that successful erasure, output sensitivity, and downstream benefit are distinct and should be evaluated separately.

1 Introduction

Representation editing methods remove or steer concepts in hidden states, including Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020), amnesic probing (Elazar et al., 2021), linear adversarial concept erasure (Ravfogel et al., 2022a), Least-squares Concept Erasure (LEACE) (Belrose et al., 2023), and affine steering (Singh et al., 2024). Related methods include demographic-free debiasing (Orgad and Belinkov, 2023), spectral and multilingual attribute removal (Shao et al., 2023b,a, 2025), nonlinear kernel erasure (Ravfogel et al., 2022b), and fair representation objectives with explicit fairness–utility tradeoffs (Kleindessner et al., 2023). These tools are attractive for debiasing, shortcut removal, and interpretability. But a common evaluation step is underspecified: after a concept becomes unrecoverable from a representation, what has been shown about the evaluated model behavior?

Our answer is that recoverability and operational use are different. Here, operational use means that the output map for the evaluated behavior is sensitive to the edited directions; it is not a claim of full causal mediation through the model. For a classifier, the output map is the task head; for a decoder language model (LM) constrained to answer tokens, it can consist of selected unembedding rows or a binary contrast such as *he* versus *she*. A concept can be linearly recoverable in a direction that one output map ignores but another uses. We therefore ask whether the evaluated logits can change and whether any downstream benefit exceeds that of similarly damaging control edits.

Prior work shows that intrinsic bias or probe measures can disagree with downstream behavior (Gonen and Goldberg, 2019; Goldfarb-Tarrant et al., 2021; Cao et al., 2022; Orgad et al., 2022; Shen et al., 2022), and that guardedness or probe-based removal has formal limits (Ravfogel et al., 2020, 2023; Kumar et al., 2022). Control-task and disentanglement evaluations stress selectivity (Hewitt and Liang, 2019; Huang et al., 2024). Causal and interchange interventions instead analyze mechanisms using counterfactuals or activation changes (Geiger et al., 2021, 2022; Canby et al., 2025). Our narrower contribution is an exact logit-change diagnostic for declared linear edits and output maps, paired with damage-matched controls that test whether observed gains are target-specific. We apply them to Bios, a controlled text shortcut, and final-state decoder-LM

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 262–280
November 6-10, 2026 ©2026 Association for Computational Linguistics

answer-token predictions.

2 Operational Use Is Readout-Local

Scope. The analytic diagnostic below assumes a linear edit at the audited representation and an explicit linear map producing the evaluated logits. It does not characterize arbitrary nonlinear downstream computation, earlier-layer edits followed by further model blocks, or open-ended generation. Outside this setting, the diagnostic is at most a local screen; claims must rest on empirical behavior and matched controls.

Consider a behavior with linear readout $R \in \mathbb{R}^{k\times d}$ applied to a centered representation $h \in \mathbb{R}^d$ with covariance Σ. For a classifier, R is the task head; for a decoder LM, R can be selected unembedding rows or a token-contrast row. For a centered linear edit $h' = Mh$, let $D = I - M$. Then

$$\Delta z = RDh,$$
$$\frac{\mathbb{E}\|\Delta z\|_2^2}{\mathbb{E}\|Rh\|_2^2} = \frac{\operatorname{tr}(RD\Sigma D^\top R^\top)}{\operatorname{tr}(R\Sigma R^\top)}.$$

For orthogonal projection edits, $D = P$ and this recovers the readout-energy expression below. For implemented linear edits, including the affine LEACE map on centered representations, we use the actual edit-operator ratio or its empirical logit-delta estimate. Under the stated linear-edit, linear-output-map scope, this ratio is exact.

Let P be the orthogonal projector onto the concept-predictive subspace removed by a full-strength linear erasure, so the edited representation is $(I - P)h$. Define the readout-energy alignment of the erased subspace as

$$\mathrm{RE}(P; R) = \frac{\|RP\|_F^2}{\|R\|_F^2}.$$

Proposition. If representations are whitened, $\mathbb{E}[hh^\top] = I$, then

$$\frac{\mathbb{E}\|\Delta z\|_2^2}{\mathbb{E}\|Rh\|_2^2} = \frac{\|RP\|_F^2}{\|R\|_F^2} = \mathrm{RE}(P; R).$$

For a random rank-r subspace independent of R, $\mathbb{E}[\mathrm{RE}(P; R)] = r/d$. Appendix A gives the proof. For centered non-whitened representations with covariance Σ, the exact ratio is $\operatorname{tr}(RP\Sigma PR^\top)/\operatorname{tr}(R\Sigma R^\top)$; the Frobenius diagnostic above is the special case $\Sigma = I$.

The implication is direct. Low-rank erasure has small average effect unless the evaluated output map places nontrivial weight energy in the erased subspace. In BERT-base (Devlin et al., 2019), a rank-one random subspace captures only $1/768 \approx 0.0013$ of readout energy in expectation. Target-specific behavioral improvement requires changing directions to which that output map is sensitive, not merely probe-predictive directions.

This diagnostic also compares erasure algorithms through the edits they choose. LEACE solves a least-change linear guardedness problem: it removes the cross-covariance signal that makes the concept linearly recoverable while minimizing representation change. INLP instead constructs a sequence of protected-attribute classifier directions and projects onto the intersection of their nullspaces. As INLP rank grows, the edited subspace expands, so its potential overlap with task- and readout-aligned structure can grow as well.

Nonlinear probes, mutual-information estimates, and predictive usable information strengthen the erasure test by asking what target information remains recoverable under a richer decoder class (Voita and Titov, 2020; Pimentel et al., 2020; Moyer et al., 2018; Gupta et al., 2021; Xu et al., 2020). They do not establish whether a fixed task output map used that information. Causal probing, tracing, and interchange interventions instead seek mechanism-level effects using counterfactuals, alignments, or activation patching (Geiger et al., 2021, 2022; Guerner et al., 2023; Meng et al., 2022; Dobrzeniecka et al., 2025; Marks et al., 2025). Our narrower diagnostic uses the declared edit and output map to screen whether the evaluated logits can move; the matched-placebo protocol then tests whether an observed benefit is target-specific. These approaches are complementary.

3 Audit Protocol

We separate three claims. *Erasure* means a probe can no longer recover the target attribute. *Benefit* means a downstream metric improves, such as fairness or out-of-distribution (OOD) accuracy. *Specificity* means the benefit exceeds what placebo edits obtain at the same damage budget. Throughout, positive fairness benefit means a gap decreased or OOD accuracy improved; larger leakage means more residual probe recoverability, with 0.5 chance for binary targets.

Let A be an edit, $B(A)$ its held-out benefit, and $C(A)$ a predeclared validation-side damage cost

such as logit perturbation or in-distribution accuracy drop. For a declared placebo family $\mathcal{C}$, we select the control strength whose validation cost is closest to $C(A)$ and compute

$$\rho(A) = \Pr_{A_0 \sim \mathcal{C}} \left[B(A_0^{\text{match}}) \leq B(A) \right].$$

Held-out benefit is never used for matching or strength selection. Intuitively, $\rho \approx 0.5$ means the real edit is ordinary among matched controls, while $\rho \approx 1$ means unusually high benefit at the same matched damage. Because ρ is a matched-control percentile, not a p-value, we interpret it only alongside the benefit interval and utility cost.

The placebo family defines the scientific claim and is chosen before held-out evaluation. A control should (i) preserve the edit algorithm, nominal rank, and strength-matching procedure; (ii) break the relation between the edited direction and target labels or markers; (iii) match on a predeclared validation cost, never held-out benefit; and (iv) be reported separately from controls testing different claims. For example, fitting LEACE to shuffled gender labels on Bios preserves the erasure algorithm and label balance while breaking the gender–representation relation. We compare it with the real edit at similar validation logit damage. INLP controls similarly shuffle target labels, and synthetic-shortcut controls shuffle markers. Random-subspace controls test only whether a benefit exceeds generic low-rank perturbation. Appendix C gives pseudocode, a worked example, and the full taxonomy (Table 3); Appendix Table 4 reports control families separately.

We also require an admissible utility region before interpreting specificity. In Bios, occupation-accuracy drops above 0.10 are treated as high-damage diagnostics rather than candidate improvements. Appendix Table 6 shows that the qualitative conclusion is stable across thresholds from 0.02 to 0.15.

Unless stated otherwise, results are means over five seeds. Percentiles are computed within seed against matched controls and then averaged; intervals are seed-level bootstrap or standard-error intervals. Appendix E reports the experimental settings, and Appendix Table 8 maps each result to its reproduction script and artifact.

4 Experiments

4.1 Bios: Erasure Without Fairness Benefit

Bios is our main fairness case. We use BERT-base classification-token (CLS) representations on a stratified 10000/2000/2000 occupation split (De-Arteaga et al., 2019). The evaluated output map is a frozen linear occupation head; before editing, occupation accuracy is 0.6949, root-mean-square (RMS) true positive rate (TPR) gap is 0.1292, and gender leakage is 0.9889. The benefit is reduction in RMS over per-occupation gender TPR gaps, so positive values mean the gap decreased; occupation accuracy drop is the utility cost.

Table 1 gives the core result. LEACE reduces gender leakage to chance but changes only 0.0042 of centered occupation-logit energy. The corresponding covariance-weighted operator ratio is 0.0044; the whitened Frobenius diagnostic is 0.0001 (Appendix Table 18). Despite high pre-edit recoverability, LEACE has no detectable fairness gain and does not exceed matched shuffled-label or random-subspace controls (Figure 2).

Section 4.4 reports the post-edit refit, nonlinear-head, placebo-family, and model-scope checks in the main text rather than relying on the appendix for their interpretation.

INLP traces a damage–benefit frontier: from rank 20 to 50, the occupation-logit ratio rises from 0.0527 to 0.1809, but occupation damage rises and all five-seed fairness intervals cross zero. Ten-seed real-edit estimates for INLP-30/50 are $+0.0045\,[-0.0028, +0.0118]$ and $+0.0060\,[-0.0037, +0.0157]$; they do not enter the five-seed matched percentiles (Appendix Table 25).

4.2 Synthetic Shortcut Positive Control

The synthetic shortcut is a falsifiability check, not an ecological-validity claim. We construct a binary text task with a semantic content token and a synthetic style marker. During training and validation, the marker matches the label with probability 0.95; at out-of-distribution (OOD) test time, the correlation reverses to 0.05. A frozen BERT CLS classifier learns the marker shortcut. We fit LEACE on a marker-balanced calibration split and compare against shuffled-marker LEACE controls.

Here the classifier uses the erased marker. Across five seeds, LEACE improves OOD accuracy by 0.6020 (interval $[0.3952, 0.8088]$). Val-

Edit	Leakage	Logit ratio	Occ. drop	RMS TPR benefit	ρ
LEACE	0.5022	0.0042	≈ 0	$-0.0041\,[-0.0095, +0.0014]$	0.30
INLP-20	0.6308	0.0527	0.0310	$+0.0046\,[-0.0016, +0.0093]$	0.77
INLP-30	0.5755	0.0941	0.0567	$+0.0077\,[-0.0032, +0.0178]$	0.83
INLP-50	0.5368	0.1809	0.1295	$+0.0079\,[-0.0050, +0.0208]$	0.94

Table 1: Focused Bios audit. Leakage is post-edit gender-probe accuracy; chance is 0.5. Logit ratio is the centered covariance-weighted logit-delta ratio for the implemented edit, using the occupation readout R. ρ is the logit-matched placebo percentile; 0.5 is ordinary and 1.0 is unusually high benefit at matched damage. All rows use the same five controlled seeds for benefits and percentiles.

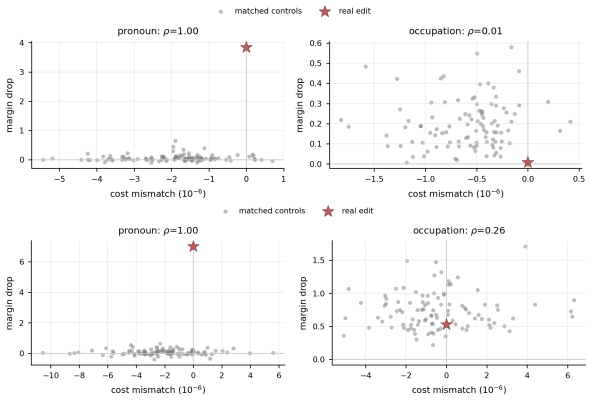


Figure 1: TinyLlama (top) and Qwen2.5-7B (bottom) answer-token audits. Gray points are 100 controls matched to the real edit's validation logit ratio; red stars are real gender edits. Margin drop is the held-out decrease in the gold-vs-alternative logit margin. Pronoun effects exceed all controls, while occupation effects do not.

idation accuracy changes by -0.0428 (interval $[-0.0523, -0.0333]$), and marker leakage falls from 1.0000 to 0.5476. The logit- and utility-matched percentiles are 0.996 and 1.000. The audit thus finds a target-specific gain when the classifier relies on the marker.

4.3 Decoder LM Answer-Token Audit

We next audit constrained answer-token behavior in TinyLlama-1.1B (Zhang et al., 2024) and Qwen2.5-7B (Qwen Team, 2025). We fit the same rank-4 gender eraser at the final normalized answer-position state immediately before the unembedding and compare it with 100 shuffled-label controls matched on centered full-vocabulary logit ratio (Figure 1). The full-vocabulary unembedding energy of the erased subspace is tiny in all four runs (0.0009–0.0020), but the behavior-local output map separates outcomes. For pronouns, the *he/she* contrast has much larger readout energy (0.216 in TinyLlama, 0.106 in Qwen); pronoun accuracy falls from 1.00 to 0.56 and 0.58, respectively, and held-out margins shrink beyond all matched controls. For occupation extraction, candidate-token energy is much smaller (0.0032 and 0.0053); both models retain 1.00 accuracy and show no corresponding improvement over controls. These are two separate one-seed, rank-4 next-token audits, not tests of nearby ranks or multi-step generation. They show that the same erasure can change pronoun choices while leaving occupation-token choices intact in this controlled setting.

The held-out pronoun-margin drops are $3.843\,[3.777, 3.911]$ for TinyLlama and $6.997\,[6.627, 7.352]$ for Qwen; both exceed all 100 matched controls. The corresponding occupation-margin drops are $0.008\,[-0.033, 0.050]$ and $0.529\,[-0.014, 1.226]$, with control percentiles 0.01 and 0.26. These intervals resample held-out examples, not training or erasure seeds. The contrast is consistent across these two models, but does not establish stability under different random seeds or nearby erasure ranks (Appendix Table 11).

4.4 Robustness and Secondary Shift Checks

Table 2 brings three Bios checks into the main paper. Re-training a linear occupation head after LEACE does not reveal a fairness improvement: across ten real-edit seeds, the benefit is $-0.0006\,[-0.0075, +0.0063]$. A frozen one-hidden-layer MLP head likewise shows no detectable gain after gender leakage falls to 0.5019. This MLP check uses actual model outputs for the behavioral audit; the linear logit-change identity does not apply to its hidden layer. On RoBERTa-large, the estimated fairness effect is negative rather than merely indistinguishable from zero. These checks limit the claim to the evaluated heads and representations; they do not establish that all nonlinear predictors or encoders respond alike.

Control choice also matters. On BERT

Bios LEACE check	TPR-gap benefit	Logit ρ
BERT, linear refit	$-0.0006\,[-0.0075, +0.0063]$	0.57
BERT, frozen MLP	$-0.0073\,[-0.0176, +0.0030]$	0.30
RoBERTa-large, linear	$-0.0133\,[-0.0181, -0.0086]$	0.03

Table 2: Bios LEACE robustness checks. Benefit is the reduction in RMS gender TPR gaps. The linear refit uses ten real-edit seeds; its matched percentile uses ten controlled seeds. The MLP and RoBERTa-large checks use five seeds. All intervals are over seeds; see Appendix Tables 20, 21, and 15 for utility costs and controls.

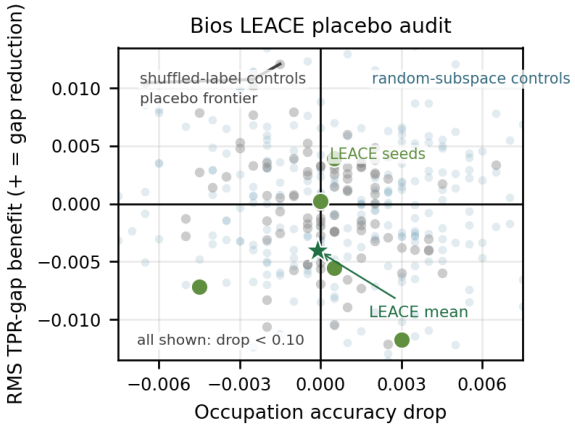


Figure 2: Bios LEACE against matched shuffled-label LEACE and random-subspace edits. The real edit lies inside both control clouds: erasing gender leakage does not yield an unusual fairness gain at similar occupation-accuracy damage.

Bios, LEACE ranks at 0.30 among logit-matched shuffled-label LEACE edits and 0.32 among rank-matched random-subspace edits; validation-accuracy matching yields 0.39 and 0.40, respectively (Appendix Table 4). Neither control family suggests a target-specific fairness improvement. The shuffled-label control preserves the erasure procedure and tests the stronger claim; the random-subspace control tests only whether the edit beats generic low-rank damage. We do not pool these percentiles across control families.

A secondary natural-text shift gives a separate, non-fairness check. A linear entailment head trained on frozen BERT representations from Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2018) is evaluated on HANS lexical-overlap examples (McCoy et al., 2019). The experiment uses 10000 training and 2000 validation MNLI examples, 6000 HANS examples, and 20 shuffled-overlap controls per seed across five seeds. LEACE reduces overlap-label probe accuracy to 0.4898, but HANS accuracy changes by $-0.0064\,[-0.0200, +0.0071]$; validation accuracy changes by $-0.0572\,[-0.0656, -0.0488]$. Its matched-control percentiles are 0.40 for both logit and utility damage (Appendix Tables 12–13). This is a second negative result about shortcut removal, not another protected-attribute fairness test.

The controlled positive result is not confined to BERT CLS features. With frozen mean-pooled RoBERTa-base and distilgpt2 representations, LEACE improves synthetic-shift accuracy by $0.6976\,[0.6766, 0.7186]$ and $0.6860\,[0.6617, 0.7103]$, respectively (Appendix Table 10). Both edits beat their shuffled-marker controls, with logit-matched percentiles 0.996 and 1.000. These gains incur substantially larger validation-accuracy drops (0.2090 and 0.1990) than the BERT positive control (0.0428). Thus this check supports the audit's ability to recognize a known shortcut across several frozen feature extractors; it does not show that the same intervention helps on natural biographies or open-ended LM outputs.

5 Discussion and Conclusion

The LEACE–INLP contrast reflects their objectives together with data geometry. LEACE solves a least-change linear guardedness problem, a property motivated as preserving downstream utility (Belrose et al., 2023). In Bios, that edit scarcely changes occupation logits, so it can erase gender recoverability while preserving task utility and leaving TPR gaps essentially unchanged. RoBERTa-large is an important qualification: the analogous edit worsens the measured fairness gap rather than merely leaving it unchanged. Nor is the Bios result a general limitation of LEACE: in the controlled shortcut, where the erased marker drives the classifier output, the audit finds a large target-specific gain. With INLP, increasing rank raises both occupation-logit change and occupation-accuracy loss. The decoder-LM contrast applies only to final-state answer-token predictions: alignment with the relevant candidate tokens, not global unembedding overlap, distinguishes the observed outcomes. The analytic identity is exact only for linear edits and explicit linear output maps; for nonlinear heads or downstream pathways, behavioral effects must be measured directly and compared with matched controls. Erasure, output sensitivity, utility cost, and target-specific benefit are distinct measurements.

Limitations

Our main fairness results are frozen-representation classifier audits. The decoder-LM experiments cover constrained next-token logits under final-hidden-state linear edits, not earlier-layer edits, adapters, open-ended generation, or multi-step generation interventions; their intervals bootstrap examples rather than independent model seeds. The Frobenius readout-energy proposition is exact for whitened representations, linear edits, and explicit linear output maps; for centered non-whitened representations, the covariance-weighted logit ratio applies. Dataset-geometry, nonlinear erasure, and fundamental-limit results address different recoverability regimes (Johansson, 2024; Shao et al., 2023b; Ravfogel et al., 2022b; Iskander et al., 2023; Chowdhury et al., 2025); task-preserving erasers such as SPLICE are complementary method-side responses (Holstege et al., 2025). Nonlinear heads require empirical actual-output diagnostics; the analytic identity does not extend to them. The audit also depends on the placebo family: a weak control can overstate specificity, while a strong control tests a narrower claim. Bios uses binary gender annotations and incomplete word swaps; we have not audited race, age, or intersectional attributes, and the diagnostic should be re-instantiated per task and representation. Our negative results are no-detectable-improvement claims under reported intervals, not formal equivalence tests. The experiments demonstrate the audit protocol and the recoverability/use distinction, not universal claims about all editing methods.

Ethics Statement

This work studies fairness-related representation editing on Bios, which contains occupation biographies and binary gender annotations. These annotations are incomplete proxies for gender and should not be interpreted as a comprehensive treatment of gender identity. Biographical text may contain identifying information, so no raw biographies are included in this paper or its source archive. We use the data only to audit whether representation edits support the specific downstream fairness claim they are often used to motivate. Our results should not be read as recommending occupation prediction from biographies, nor as validating RMS TPR gap as a complete fairness criterion. The central recommendation is methodological: claims of debiasing or shortcut improvement should report utility costs, uncertainty, and matched placebo comparisons rather than relying on recoverability alone.

References

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

Marc E. Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. 2025. How reliable are causal probing interventions? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 857–878.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570.

Somnath Basu Roy Chowdhury, Kumar Avinava Dubey, Ahmad Beirami, Rahul Kidambi, Nicholas Monath, Amr Ahmed, and Snigdha Chaturvedi. 2025. Fundamental limits of perfect concept erasure. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 901–909. PMLR.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*, pages 4171–4186.

Alicja Dobrzeniecka, Antske Fokkens, and Pia Sommerauer. 2025. Improving causal interventions in amnesic probing with mean projection or LEACE. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12981–12993.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. 2022. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1926–1940.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 609–614.

Clément Guerner, Tianyu Liu, Anej Svete, Alexander Warstadt, and Ryan Cotterell. 2023. A geometric notion of causal probing. *Preprint*, arXiv:2307.15054.

Umang Gupta, Aaron M. Ferber, Bistra Dilkina, and Greg Ver Steeg. 2021. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7610–7619.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2733–2743.

Floris Holstege, Shauli Ravfogel, and Bram Wouters. 2025. Preserving task-relevant information under linear concept removal. *Preprint*, arXiv:2506.10703.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. 2024. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 8669–8687.

Shadi Iskander, Kira Radinsky, and Yonatan Belinkov. 2023. Shielded representations: Protecting sensitive attributes through iterative gradient-based projection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5961–5977.

Richard Johansson. 2024. What happens to a dataset transformed by a projection-based concept removal method? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17486–17492.

Matthäus Kleindessner, Michele Donini, Chris Russell, and Muhammad Bilal Zafar. 2023. Efficient fair PCA for fair representation learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5250–5270. PMLR.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. 2022. Probing classifiers are unreliable for concept removal and detection. In *Advances in Neural Information Processing Systems*.

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

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *Proceedings of the International Conference on Learning Representations*, pages 23888–23923.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. 2018. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, volume 31.

Hadas Orgad and Yonatan Belinkov. 2023. BLIND: Bias removal with no demographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8801–8821.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2602–2628.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622.

Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI technical report.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256.

Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. 2023. Log-linear guardedness and its implications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 9413–9431.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D. Cotterell. 2022a. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18400–18421. PMLR.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. 2022b. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6034–6055.

Shun Shao, Yftah Ziser, and Shay B. Cohen. 2023a. Erasure of unaligned attributes from neural representations. *Transactions of the Association for Computational Linguistics*, 11:488–510.

Shun Shao, Yftah Ziser, and Shay B. Cohen. 2023b. Gold doesn't always glitter: Spectral removal of linear and nonlinear guarded attribute information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1603–1614.

Shun Shao, Yftah Ziser, Zheng Zhao, Yifu Qiu, Shay B. Cohen, and Anna Korhonen. 2025. Iterative multilingual spectral attribute erasure. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29230–29255.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Does representational fairness imply empirical fairness? In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 81–95.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. 2024. Representation surgery: Theory and practice of affine steering. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 45663–45680. PMLR.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 183–196.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1112–1122.

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

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. TinyLlama: An open-source small language model. *Preprint*, arXiv:2401.02385.

A Readout-Energy Proposition

Let $R \in \mathbb{R}^{k\times d}$ be a linear readout, $h \in \mathbb{R}^d$ a representation, and P an orthogonal projector onto the erased subspace. The full-strength linear erasure maps h to $(I - P)h$, so the logit change is $\Delta z = RPh$.

Expected logit perturbation. For a centered linear edit $h' = Mh$ with $D = I - M$,

$$\frac{\mathbb{E}\|RDh\|_2^2}{\mathbb{E}\|Rh\|_2^2} = \frac{\mathrm{tr}(RD\Sigma D^\top R^\top)}{\mathrm{tr}(R\Sigma R^\top)}.$$

For an orthogonal projection erasure, $D = P$, giving the expression below. For centered non-whitened representations with covariance $\Sigma = \mathbb{E}[hh^\top]$,

$$\frac{\mathbb{E}\|\Delta z\|_2^2}{\mathbb{E}\|Rh\|_2^2} = \frac{\mathrm{tr}(RP\Sigma PR^\top)}{\mathrm{tr}(R\Sigma R^\top)}.$$

If representations are whitened, $\Sigma = I$. Then

$$\mathbb{E}\|\Delta z\|_2^2 = \mathrm{tr}\Big(RP\mathbb{E}[hh^\top]PR^\top\Big).$$

Since P is an orthogonal projector, $P^2 = P = P^\top$, so

$$\begin{aligned}\mathbb{E}\|\Delta z\|_2^2 &= \mathrm{tr}(RPPR^\top)\\ &= \mathrm{tr}(RPR^\top) = \|RP\|_F^2.\end{aligned}$$

Similarly,

$$\mathbb{E}\|Rh\|_2^2 = \mathrm{tr}(R\mathbb{E}[hh^\top]R^\top) = \|R\|_F^2.$$

Therefore

$$\frac{\mathbb{E}\|\Delta z\|_2^2}{\mathbb{E}\|Rh\|_2^2} = \frac{\|RP\|_F^2}{\|R\|_F^2}.$$

Random-subspace baseline. If P is a uniformly random rank-r orthogonal projector independent of R, then $\mathbb{E}[P] = (r/d)I$ by rotational symmetry. Thus

$$\begin{aligned}\mathbb{E}\|RP\|_F^2 &= \mathrm{tr}(R\mathbb{E}[P]R^\top)\\ &= \frac{r}{d}\|R\|_F^2,\end{aligned}$$

and $\mathbb{E}[\|RP\|_F^2/\|R\|_F^2] = r/d$.

B Minimal Constructions

Fairness benefit by damage. Let a representation contain a protected coordinate G and a classifier-score coordinate T, with fixed head $h(r) = \mathbb{1}\{T > 0\}$. Suppose group 0 has higher true positive rate (TPR) than group 1. Editing T to flip an ϵ-mass subset of correctly classified positive examples in group 0 reduces the TPR gap by ϵ, while preserving G and decreasing task accuracy. Thus a raw TPR-gap improvement need not identify a target-specific fairness improvement.

Erasure without behavioral change. Let $G, U \in \{0, 1\}$ be independent fair bits and let $r = (G, U)$. Let the original classifier be $h(r) = U$. The edit $A(G, U) = (0, U)$ removes all linear information about G but leaves $h(A(r)) = h(r)$. If group TPRs differed before the edit because of the relation between U, G, and labels, they remain different after perfect erasure of G.

C Audit Protocol

For each structured edit, we predefine a benefit $B(A)$ and one or more costs $C(A)$. For robustness experiments, benefits are out-of-distribution (OOD) or worst-group accuracy gains. For Bios, the benefit is reduction in root-mean-square (RMS) TPR gap under the original occupation head. Costs include logit-delta variance, validation or in-distribution (ID) accuracy drop, and task accuracy drop. When validation accuracy, logit perturbation, and ID accuracy disagree as costs, declare a primary cost in advance and report secondary costs alongside. Do not select the matching cost post hoc.

For each placebo control A_0, we sweep the same strength grid as the structured edit within that operator family and choose the point whose validation or calibration cost is closest to the structured edit's cost. If two strengths are tied, we choose the lower-damage strength; this conservative tie-break prevents controls from receiving extra perturbation relative to the audited edit. We do not interpolate, discard imperfect matches, or use held-out test benefit for matching. In multi-seed audits, we compute matched percentiles per seed and average across seeds; final Bios table intervals bootstrap seed-level means and percentiles.

Within each audit, the structured edit and its placebo controls use the same predeclared strength grid. The grid may differ across operator families. Minimum Description Length (MDL), principal component analysis (PCA), and random-subspace diagnostics use $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$; LEACE, affine, INLP, Pair-PC, and paired-interpolation audits use denser low-

damage grids reported with the artifacts.

Pseudocode. For a new setting, the audit is:

1. Define the behavioral benefit B and admissible utility bound before held-out evaluation.
2. Declare the placebo family $\mathcal{C}$ and the validation-side matching cost C.
3. Fit the structured edit and each placebo edit on train/calibration data only.
4. Sweep the same strength grid for the structured edit and controls.
5. Select the structured edit strength using validation-only criteria, then match each control to the nearest validation cost.
6. Evaluate held-out benefit once, and report benefit, cost, ρ, and an interval over seeds or bootstrap resamples.

Worked placebo example. For Bios LEACE, the structured edit is a supervised closed-form eraser fit to gender labels. The closest machinery-matched placebo is therefore LEACE fit to shuffled gender labels: it preserves the erasure algorithm, representation distribution, label balance, strength grid, and matching procedure, but breaks the target relation between examples and gender. Beating this control licenses the claim that benefit depends on the true gender relation rather than on applying LEACE-like covariance surgery. A rank-matched random-subspace control would be easier to beat and would license only a weaker claim, namely that the benefit is not explained by generic low-rank perturbation. PCA controls answer a different question again: whether benefits arise from removing high-variance structure rather than the declared target. Here and below, CLS denotes classification-token representations. When several placebo families are scientifically plausible, the audit should report them separately rather than collapse them into one number.

D Minimum Description Length (MDL)-Guided Edit

The MDL-guided edit is one audited candidate, not the paper's main contribution. Given in-distribution calibration data and a shifted or counterfactual validation set, we train linear probes and score their brittleness by

$$\mathrm{Gap}(h) = \frac{\mathrm{Acc}_{\mathrm{ID}}(h) - \mathrm{Acc}_{\mathrm{shift}}(h)}{\log(1 + \overline{\mathrm{MDL}}_{\mathrm{ID}}(h))}.$$

The selected brittle row map is B; a stable task row map S is selected by shifted or group-balanced performance and low codelength (Voita and Titov, 2020; Pimentel et al., 2020).

With centered representations $X = r - \mu$ and covariance $\Sigma = \mathbb{E}[XX^\top]$, define the stable-conditioned residual

$$R = B - B\Sigma S^\top (S\Sigma S^\top)^+ S.$$

The full-strength covariance edit is

$$A_R = I - \Sigma R^\top (R\Sigma R^\top)^+ R,$$

and the applied edit is

$$r' = \mu + \left[I - \alpha\Sigma R^\top (R\Sigma R^\top)^+ R\right](r - \mu).$$

Conditional on B and S, A_R is the minimum-reconstruction-loss linear map satisfying covariance erasure of R, with unique action on the covariance support. The proposition does not imply that MDL finds the causal nuisance handle.

E Experimental Details

Bios. We use BERT-base-uncased classification-token (CLS) representations on a stratified 10000/2000/2000 occupation split. Gender-word swaps define the shift for label-free MDL selection. Individual gender labels are used only for evaluation and supervised LEACE/affine/INLP baselines. LEACE uses `concept-erasure==0.2.4` with 5 seeds and 20 shuffled-label controls per seed. Affine audits use 5 seeds and 50 shuffled-partition controls per seed. INLP audits use the public `nullspace_projection` implementation with by-class autoregressive LinearSVC classifiers, 5 seeds, and 20 shuffled-label controls per seed. MDL/PCA diagnostics use 5 seeds and 20 random-subspace controls per seed.

The nonlinear-head Bios audit uses the same frozen representations and LEACE controls, but trains a one-hidden-layer multilayer perceptron (MLP) occupation head with 256 hidden units and ReLU activation for 80 epochs with AdamW. The head-subspace diagnostic fits the standard linear head and computes both the Frobenius-energy

Placebo family	Construction	Claim licensed if beaten
Random subspace	Remove or perturb a rank-matched random subspace with the same strength grid.	Benefit is not explained by generic low-rank damage at the same budget.
Variance/PCA	Remove high-variance directions, or compare against rank-matched PCA variants.	Benefit is not merely due to deleting large-variance representation components.
Shuffled target labels	Run the same erasure method after permuting protected, marker, or shortcut labels.	Benefit depends on the true target, not just the erasure machinery.
Shuffled partitions	For affine edits, construct source/target partitions after shuffling group assignment.	Distribution alignment matters beyond arbitrary matched affine shifts.
Shuffled counterfactual deltas	Apply paired-interpolation controls after shuffling example-level swap deltas.	Pairwise counterfactual direction matters beyond generic swap-like movement.
Oracle/known shortcut	Use a known controlled shortcut as a positive-control edit.	The audit can identify target-specific improvement when the targeted handle is real.

Table 3: Placebo-family taxonomy. Each control licenses a different scientific claim.

Representation	Placebo family	Logit pct.	Val-drop pct.
BERT-base CLS	Shuffled-label LEACE	$0.300\ [0.040, 0.660]$	$0.390\ [0.020, 0.770]$
BERT-base CLS	Rank-matched random subspace	$0.320\ [0.000, 0.664]$	$0.400\ [0.000, 0.811]$
distilgpt2 mean	Shuffled-label LEACE	$0.410\ [0.167, 0.653]$	$0.350\ [0.118, 0.582]$
distilgpt2 mean	Rank-matched random subspace	$0.324\ [0.153, 0.495]$	$0.356\ [0.169, 0.543]$

Table 4: Placebo-family sensitivity for Bios LEACE. The weaker random-subspace placebo does not make LEACE appear to yield a target-specific fairness improvement; percentiles remain ordinary and original-head fairness benefit remains near zero.

fraction of the head weight matrix inside the edited subspace and the centered covariance-weighted logit-energy ratio on the realized Bios features. The INLP head-energy sweep applies the same diagnostics to the subspace removed by the public INLP projection at ranks 10, 20, 30, and 50.

The RoBERTa-large Bios check (Liu et al., 2019) repeats the LEACE audit on frozen RoBERTa-large CLS representations with the same stratified 10000/2000/2000 split, five seeds, and 20 shuffled-label controls per seed. The distilgpt2 Bios checks use frozen mean-pooled decoder hidden states (Radford et al., 2019) on a stratified 5000/1000/1000 split, five seeds, and 20 controls per seed. They test the same original-head fairness-improvement claim on natural Bios text with decoder representations; they are not generative-model interventions.

Counterfactual-pair diagnostics use original and gender-word-swapped biographies. Direct paired interpolation applies $r + \alpha(r_{\text{swap}} - r)$ at $\alpha = 0.5$ over 5 seeds with 50 shuffled-delta controls per seed. Pair-PC rows use principal directions of $r - r_{\text{swap}}$ as candidate edit directions over 5 seeds with at least 20 random-subspace controls per seed.

Synthetic text shortcut. We generate binary text examples with a semantic content token and a synthetic style marker. During training and validation, the marker matches the label with probability 0.95; at OOD test time, it matches with probability 0.05. A linear classifier is trained on frozen BERT-base-uncased CLS features from 4000 examples. LEACE is fit on a marker-balanced calibration split to erase the marker from the representation. We audit against 50 shuffled-marker LEACE controls per seed over 5 seeds.

Model-scope controlled checks. To probe whether the positive-control code path is tied only to BERT CLS features, we also run appendix-only synthetic shortcut audits on frozen RoBERTa-base mean-pooled states and frozen distilgpt2 mean-pooled states. RoBERTa uses the same 4000/1000/1000 split and 50 shuffled-marker controls per seed as BERT. distilgpt2 uses 500 training examples, 200 validation examples, 200 OOD examples, five seeds, and 20 shuffled-marker controls per seed. These controlled runs are not evidence that the Bios or lexical-overlap negative results transfer to natural decoder large language models (LLMs).

Decoder LM answer-token audit. The decoder-LM audit uses TinyLlama/TinyLlama-1.1B-Chat-v1.0 and Qwen/Qwen2.5-7B-Instruct. Each run uses seed 13, a rank-4 linear projection fit to

gender labels, and 100 shuffled-label controls. Prompts have the form *Biography: John works as a doctor. Question: Which pronoun best refers to John? Answer:* for the pronoun task and *Question: What job is named in the sentence? Answer:* for the occupation task. We edit the final normalized answer-position hidden state before the language-model head (`model.norm`) and evaluate next-token logits. Pronoun candidates are one-token *he/she* continuations; occupation candidates are one-token occupation continuations supported by the tokenizer. Controls are scaled by solving for the α that matches the real edit's validation centered full-vocabulary logit ratio; no reported control is alpha-clipped. Intervals are bootstrap intervals over held-out examples, not independent model seeds.

Artifact terms and intended use. The audit uses the existing Bios distribution for research evaluation, not to build a system for predicting individuals' occupations or gender. The LabHC Bios dataset card[1] lists an MIT license for that distribution; the terms of any upstream source of biographies should be verified independently. The TinyLlama-1.1B-Chat-v1.0[2] and Qwen2.5-7B-Instruct[3] model cards list Apache 2.0 licenses. Our synthetic shortcut consists of generated examples rather than biographies. The source package for this paper contains no raw Bios text or model weights.

Lexical-overlap audit. We train a binary entailment-vs-non-entailment linear head on frozen BERT-base-uncased CLS features from 10000 Multi-Genre Natural Language Inference (MNLI) examples (Williams et al., 2018), validate on 2000 MNLI matched examples, and evaluate OOD robustness on 6000 Heuristic Analysis for Natural Language Inference Systems (HANS) examples (McCoy et al., 2019). The shortcut attribute is median-thresholded lexical coverage of hypothesis tokens by premise tokens, computed on the MNLI training subset. LEACE is audited against 20 shuffled-overlap controls per seed over 5 seeds.

[1] https://huggingface.co/datasets/LabHC/bias_in_bios

[2] https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0

[3] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

Damage threshold sensitivity. Table 6 varies the admissible occupation-accuracy drop used to decide which Bios edits count as low-damage candidates. Table 7 summarizes the resulting cross-case audit readout.

Reproducibility map. Table 8 maps each reported audit component to the script and artifact basename used to reproduce it.

Name in tables	Structured edit	Primary controls	Role in paper
Selected MDL	Label-free MDL-selected stable-conditional row edit.	Random rank-matched subspaces.	Target-selection diagnostic.
Raw MDL	Full MDL-selected residual without stable conditioning.	Random rank-matched subspaces.	High-damage diagnostic, not a candidate edit.
PCA	Rank-matched high-variance projection.	Random rank-matched subspaces.	Erasure/damage diagnostic.
Pair-PC	Principal components of gender-swap deltas.	Random rank-matched subspaces.	Counterfactual-pair geometry diagnostic.
Pair interp.	Direct interpolation toward gender-swapped representations.	Shuffled swap deltas.	Counterfactual-pair handle diagnostic.
LEACE	Closed-form supervised linear erasure.	Shuffled-label LEACE.	Strong erasure baseline.
Affine	Supervised mean or mean+covariance alignment.	Shuffled affine partitions.	Supervised erasure/steering baseline.
INLP	Iterative supervised nullspace projection.	Shuffled-label INLP.	Published-method rank frontier.

Table 5: Bios method and control map. Raw MDL and PCA are retained to demonstrate damage-driven metrics, not as acceptable edits.

Max occ. drop	Best admissible benefit	Best pct.	Conclusion
0.02	+0.0052	0.69	No strong low-damage improvement.
0.05	+0.0052	0.77	INLP-20 enters; CI crosses zero.
0.10	+0.0077	0.83	INLP-30 enters; CI crosses zero.
0.15	+0.0079	0.94	INLP-50 enters; steep utility cost.

Table 6: Bios admissible-damage sensitivity. INLP forms a smooth utility–benefit frontier: larger ranks improve the mean fairness benefit and specificity, but only by spending more occupation accuracy. Raw MDL and PCA enter only at much larger drops (0.69 and 0.40).

Setting	Target erased?	Benefit?	Damage	Specificity	Conclusion
Bios LEACE/affine	yes	no	low	low	erasure without occupation-readout benefit
Bios distilgpt2 LEACE/affine	yes	no	low	low	decoder-representation edit-family check
Bios Raw MDL/PCA	partial/yes	apparent	high	medium/high	damage-driven benefit diagnostics
Bios INLP-50	partial	small	high	suggestive	high-damage fairness trend
Synthetic shortcut LEACE	yes	large	modest	high	target-specific controlled improvement
Decoder LM pronouns	partial	large	matched	high	answer-token readout uses erased signal
Decoder LM occupations	partial	no	matched	low	answer-token readout mostly does not use erased signal
MNLI–HANS LEACE	yes	no	moderate	low	erasure without OOD benefit

Table 7: Auxiliary cross-case audit summary.

Component	Script basename	Main artifact basename
Bios finalization summary	summarize_bios_finalization_sweep.py	bios_finalization_sweep_summary.json
Bios claim-separation figure	make_bios_claim_separation_figure.py	bios_claim_separation.png
Bios LEACE alignment	audit_bios_leace_alignment.py	leace_head_alignment_5seed_summary.json
Bios INLP head-energy sweep	audit_bios_inlp_alignment.py	inlp_head_alignment_rank_sweep.json
Bios INLP 10-seed method check	run_bios_inlp_grid.py	inlp_rank30_50_method_10seed_summary.json
Bios covariance logit ratios	audit_bios_leace_alignment.py; audit_bios_inlp_alignment.py	inlp_sigma_head_energy_rank_sweep_5seed.json
Bios LEACE refit audit	audit_bios_leace_refit_matched.py	leace_refit_matched_10seed_draw20_combined.json
Bios MLP-head audit	audit_bios_leace_mlp.py	leace_mlp_head_audit_5seed_draw20_summary.json
Bios RoBERTa-large audit	audit_bios_leace.py	leace_audit_roberta_large_cls_strat10k_5seed_draw20_summary.json
Bios RoBERTa-large alignment	audit_bios_leace_alignment.py	leace_head_alignment_roberta_large_5seed_summary.json
Bios distilgpt2 audit	audit_bios_leace.py	leace_audit_distilgpt2_mean_strat5k_5seed_draw20_summary.json
Bios distilgpt2 alignment	audit_bios_leace_alignment.py	leace_head_alignment_distilgpt2_mean_strat5k_5seed_summary.json
Bios distilgpt2 affine audit	audit_bios_affine.py	affine_audit_distilgpt2_mean_strat5k_5seed_draw20_summary.json
Bios LEACE placebo sensitivity	audit_bios_leace_placebo_sensitivity.py	leace_placebo_sensitivity_*_summary.json
Synthetic text shortcut	synthetic_text_shortcut_bert.py	synthetic_text_shortcut/*summary.json
RoBERTa/distilgpt2 shortcut checks	synthetic_text_shortcut_bert.py	synthetic_text_shortcut/*_5seed_*summary.json
Decoder LM answer-token audits	notebooks/llm*_operational_use_audit_kaggle.ipynb	llm_operational_use/*controls100_results.json
MNLI–HANS shortcut audit	audit_nli_hans_overlap.py	leace_overlap_audit_5seed_draw20_summary.json
Specificity uncertainty	summarize_specificity_uncertainty.py	specificity_uncertainty_summary.md

Table 8: Reproducibility map for the audit pipeline. Scripts live under experiments/; artifacts live under artifacts/ or acl_short_paper/figures/ as appropriate.

F Full Results

Synthetic and HANS checks. Tables 9–13 give the full positive-control and MNLI–HANS rows.

Metric	Value
OOD Δ	$+0.6020\ [+0.3952, +0.8088]$
Validation accuracy Δ	$-0.0428\ [-0.0523, -0.0333]$
Leakage after	0.5476
Matched percentiles	$0.996/1.000$

Table 9: Synthetic text shortcut positive control using frozen BERT CLS features.

Model-scope checks. The RoBERTa and distilgpt2 runs in Table 10 keep the task controlled and the audit identical: LEACE erases a known marker from frozen hidden states and is compared against shuffled-marker LEACE controls at matched validation budgets. They show that the positive-control audit is not hard-coded to BERT CLS features. They do not establish that the Bios or HANS conclusions transfer to decoder LLM representation engineering.

Setting	Readout
RoBERTa-base mean	OOD $+0.6976\ [+0.6766, +0.7186]$; val. $-0.2090\ [-0.2232, -0.1948]$; leakage 0.5076; pct. $0.996/1.000$.
distilgpt2 mean	OOD $+0.6860\ [+0.6617, +0.7103]$; val. $-0.1990\ [-0.2095, -0.1885]$; leakage 0.5200; pct. $1.000/1.000$.

Table 10: Appendix-only model-scope checks on the controlled synthetic shortcut. These runs test audit portability, not natural LLM behavior change.

Metric	Value
HANS Δ	$-0.0064\ [-0.0200, +0.0071]$
Validation accuracy Δ	$-0.0572\ [-0.0656, -0.0488]$
Leakage after	0.4898
Matched percentiles	$0.400/0.400$

Table 12: MNLI–HANS lexical-overlap audit.

HANS heuristic	Baseline acc.	LEACE Δ
Constituent	0.5340	$-0.0136\ [-0.0422, +0.0150]$
Lexical overlap	0.5149	$-0.0054\ [-0.0149, +0.0041]$
Subsequence	0.5103	$-0.0007\ [-0.0199, +0.0186]$

Table 13: HANS heuristic-family breakdown. LEACE overlap erasure is flat across heuristic families.

Specificity uncertainty. Table 14 reports seed-level uncertainty for the matched specificity percentiles used in the main text. The central readout is that high-percentile rows either correspond to the controlled synthetic shortcut or to high-damage Bios diagnostics/frontiers; low-damage Bios LEACE and affine edits remain ordinary under both logit and utility matching.

Bios additional results. Tables 15–26 collect the additional Bios audits: architecture checks, head-subspace alignment, INLP rank behavior, refit and MLP-head diagnostics, and full supervised-edit rows.

Model	Readout	Leakage	Global RE	Local RE	Acc.	Margin drop	ρ
TinyLlama	he/she	0.542	0.0019	0.216	$1.00 \to 0.56$	$3.843\,[3.777, 3.911]$	1.00
TinyLlama	occupation	0.521	0.0020	0.0032	$1.00 \to 1.00$	$0.008\,[-0.033, 0.050]$	0.01
Qwen2.5-7B	he/she	0.708	0.0009	0.106	$1.00 \to 0.58$	$6.997\,[6.627, 7.352]$	1.00
Qwen2.5-7B	occupation	0.750	0.0010	0.0053	$1.00 \to 1.00$	$0.529\,[-0.014, 1.226]$	0.26

Table 11: Decoder-LM answer-token audit details. Global RE is Frobenius readout energy against the full unembedding. Local RE is the he/she contrast energy for pronouns and candidate-token row energy for occupations. Margin-drop intervals bootstrap held-out examples.

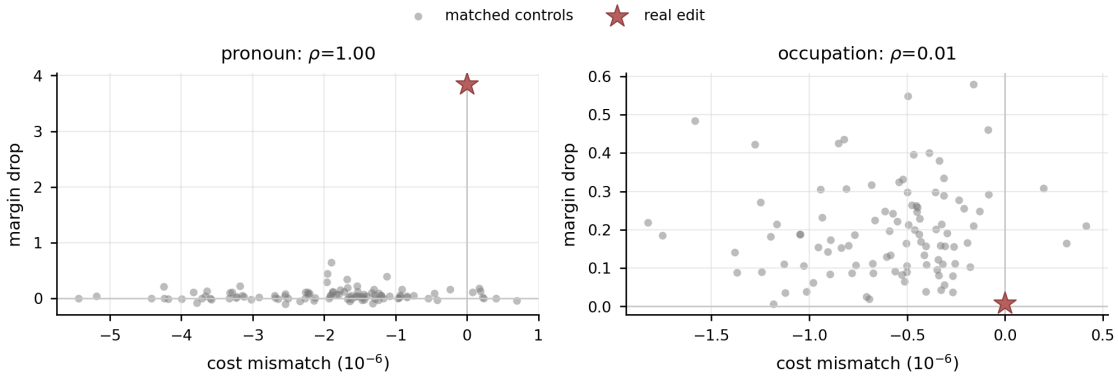


Figure 3: TinyLlama 100-control answer-token audit. Controls are shuffled-label erasers matched on validation centered full-vocabulary logit ratio; the x-axis shows cost mismatch relative to the real edit. Margin drop is the held-out decrease in the gold-vs-alternative answer-token logit margin.

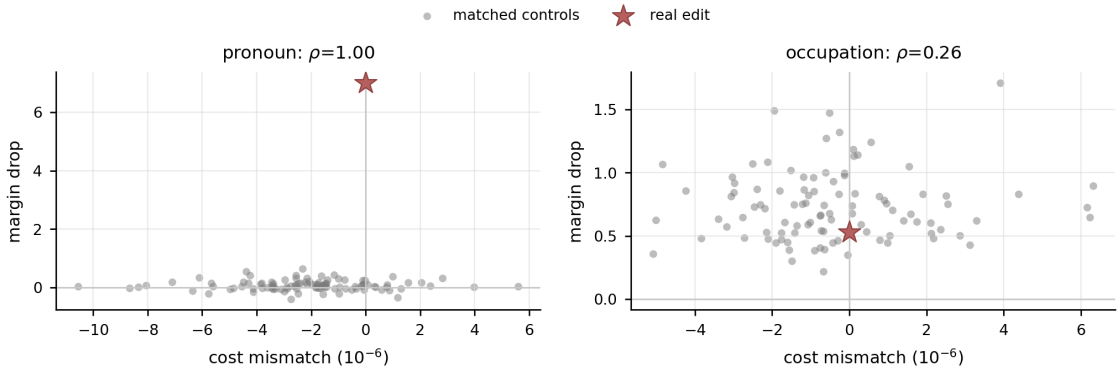


Figure 4: Qwen2.5-7B 100-control answer-token audit. Controls are shuffled-label erasers matched on validation centered full-vocabulary logit ratio; the x-axis shows cost mismatch relative to the real edit. Margin drop is the held-out decrease in the gold-vs-alternative answer-token logit margin.

Experiment	Method	Logit pct.	Utility pct.
Bios	Raw MDL	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]
Bios	PCA rank-28	0.760 [0.520, 0.960]	0.770 [0.540, 0.960]
Bios	Pair interpolation	0.688 [0.472, 0.904]	0.852 [0.584, 0.996]
Bios	LEACE	0.300 [0.040, 0.660]	0.390 [0.020, 0.770]
Bios distilgpt2	LEACE	0.410 [0.167, 0.653]	0.350 [0.118, 0.582]
Bios distilgpt2	Affine mean	0.320 [0.000, 0.649]	0.240 [0.000, 0.576]
Bios distilgpt2	Affine mean+cov	0.440 [0.164, 0.716]	0.420 [0.133, 0.707]
Bios	Affine mean	0.280 [0.004, 0.580]	0.256 [0.004, 0.636]
Bios	Affine mean+cov	0.584 [0.360, 0.812]	0.592 [0.352, 0.832]
Bios	INLP-20	0.770 [0.650, 0.880]	0.770 [0.650, 0.900]
Bios	INLP-30	0.830 [0.700, 0.950]	0.820 [0.700, 0.920]
Bios	INLP-50	0.940 [0.820, 1.000]	0.940 [0.820, 1.000]
Synthetic text	LEACE	0.996 [0.988, 1.000]	1.000 [1.000, 1.000]
RoBERTa synthetic	LEACE	0.996 [0.988, 1.000]	1.000 [1.000, 1.000]
distilgpt2 synthetic	LEACE	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]
MNLI–HANS	LEACE overlap	0.400 [0.000, 0.880]	0.400 [0.000, 0.880]

Table 14: Specificity percentile uncertainty. Utility means validation- or task-accuracy-drop matching.

Representation	Seeds	Leakage	Acc. Δ	Fair. benefit	Pct.
BERT-base CLS LEACE	5	0.5022	$+0.0001\,[-0.0023, +0.0025]$	$-0.0041\,[-0.0095, +0.0014]$	0.300/0.390
RoBERTa-large CLS LEACE	5	0.5001	$-0.0014\,[-0.0037, +0.0009]$	$-0.0133\,[-0.0181, -0.0086]$	0.030/0.000
distilgpt2 mean LEACE	5	0.5108	$-0.0012\,[-0.0064, +0.0040]$	$-0.0030\,[-0.0102, +0.0042]$	0.410/0.350
distilgpt2 mean affine	5	0.5020	$+0.0016\,[-0.0027, +0.0059]$	$-0.0025\,[-0.0093, +0.0044]$	0.320/0.240
distilgpt2 mean affine+cov	5	0.4958	$-0.0126\,[-0.0175, -0.0077]$	$-0.0024\,[-0.0182, +0.0135]$	0.440/0.420

Table 15: Bios architecture and edit-family checks. BERT-base is the main audit; RoBERTa-large is a five-seed encoder check; distilgpt2 rows are five-seed decoder-representation checks on a 5000/1000/1000 split.

Representation	Head energy	Delta rank	Fair. benefit	Leakage after
BERT-base CLS	0.0001 [0.0001, 0.0001]	1.0000	$-0.0041\,[-0.0095, +0.0014]$	0.4955
RoBERTa-large CLS	0.0005 [0.0004, 0.0006]	1.0000	$-0.0088\,[-0.0156, -0.0019]$	0.5081
distilgpt2 mean	0.0004 [0.0003, 0.0004]	1.0000	$-0.0030\,[-0.0102, +0.0042]$	0.5036

Table 16: Bios head-subspace alignment diagnostic. The frozen linear occupation head places almost none of its weight energy in the subspace changed by full-strength LEACE across encoder and decoder representations.

Edit	Frobenius HE	Covariance-weighted logit ratio
LEACE	0.000091	0.004400
INLP-10	0.013916	0.024594
INLP-20	0.028122	0.052710
INLP-30	0.043364	0.094107
INLP-50	0.072786	0.180907

Table 17: Bios Frobenius versus covariance-weighted logit diagnostics. Frobenius HE is the whitened-coordinate diagnostic; the covariance-weighted ratio uses centered BERT-base CLS features and the implemented edit or projection.

Diagnostic	Mean	95% interval
Frobenius HE	0.0000908	[0.0000769, 0.0001047]
Covariance-weighted ratio	0.00440	[0.00377, 0.00503]
Empirical projector ratio	0.00440	[0.00377, 0.00503]
Actual full-LEACE delta ratio, centered denom.	0.00420	[0.00360, 0.00480]
Actual full-LEACE delta ratio, uncentered denom.	0.00290	[0.00243, 0.00338]

Table 18: Bios LEACE logit-ratio diagnostics. The covariance-weighted projector ratio matches the empirical centered projector ratio; the actual full-LEACE delta ratio is the value used in the main Bios table.

INLP rank	Frobenius HE	Occ. damage	Orig. fair.	Pct.
10	$0.0139\,[0.0131, 0.0147]$	0.0156	-0.0035	0.45
20	$0.0281\,[0.0260, 0.0303]$	0.0310	$+0.0046$	0.77
30	$0.0434\,[0.0416, 0.0452]$	0.0567	$+0.0077$	0.83
50	$0.0728\,[0.0691, 0.0765]$	0.1295	$+0.0079$	0.94

Table 19: INLP Frobenius head-energy sweep on Bios. As rank increases, the removed subspace overlaps more with the frozen occupation head, and the fairness trend appears only alongside increasing occupation damage. Covariance-weighted ratios are in Table 17.

Edit	Refit fair.	Refit acc. Δ	Logit pct.	Val-drop pct.
LEACE	$-0.0006\,[-0.0075, +0.0063]$	$+0.0047\,[-0.0054, +0.0148]$	$0.570\,[0.332, 0.808]$	$0.525\,[0.268, 0.782]$

Table 20: Post-edit linear-head refit audit for Bios LEACE over 10 seeds. For each real or shuffled-LEACE edit, we train a new linear occupation head for 20 epochs on edited train representations and compare refit fairness benefit at matched frozen-head budgets.

Metric	Value
Head	MLP
Base accuracy	0.7099
Accuracy Δ	$+0.0054\,[+0.0017, +0.0091]$
Original-head fairness benefit	$-0.0073\,[-0.0176, +0.0030]$
Refit-head fairness benefit	$+0.0020\,[-0.0072, +0.0111]$
Leakage after edit	0.5019
Matched percentiles	$0.300/0.360$

Table 21: Bios LEACE audit with a one-hidden-layer MLP occupation head. Leakage erasure still does not yield an original-head fairness benefit.

Edit	Leakage	Orig. acc. Δ	Orig. fair. benefit
LEACE	0.5022	$+0.0001\,[-0.0023, +0.0025]$	$-0.0041\,[-0.0095, +0.0014]$

Table 22: Bios LEACE audit using `concept-erasure==0.2.4`.

Affine edit	Leakage	Orig. acc. Δ	Orig. fair. benefit
Mean	0.4984	$-0.0003\,[-0.0031, +0.0025]$	$-0.0038\,[-0.0084, +0.0009]$
Mean+cov	0.5021	$-0.0075\,[-0.0103, -0.0047]$	$+0.0017\,[-0.0019, +0.0054]$

Table 23: Bios supervised affine re-audit.

INLP rank	Leakage	Acc. Δ	Fair. benefit	Logit pct.	Val-drop pct.
10	0.7279	$-0.0156\,[-0.0224, -0.0094]$	$-0.0035\,[-0.0077, +0.0018]$	0.4500	0.4800
20	0.6308	$-0.0310\,[-0.0450, -0.0188]$	$+0.0046\,[-0.0016, +0.0093]$	0.7700	0.7700
30	0.5755	$-0.0567\,[-0.0742, -0.0445]$	$+0.0077\,[-0.0032, +0.0178]$	0.8300	0.8200
50	0.5368	$-0.1295\,[-0.1680, -0.0910]$	$+0.0079\,[-0.0050, +0.0208]$	0.9400	0.9400

Table 24: Bios exact-INLP rank sweep. Brackets show seed-bootstrap intervals. Higher rank gives a more specific positive trend, but the trend is accompanied by increasing occupation damage.

Metric	INLP-30	INLP-50
Leakage	$0.5770\,[0.5601, 0.5939]$	$0.5402\,[0.5218, 0.5586]$
Orig. acc. Δ	$-0.0514\,[-0.0622, -0.0407]$	$-0.1184\,[-0.1424, -0.0944]$
Orig. fair. benefit	$+0.0045\,[-0.0028, +0.0118]$	$+0.0060\,[-0.0037, +0.0157]$
Refit fair. benefit	$-0.0000\,[-0.0106, +0.0105]$	$+0.0064\,[-0.0066, +0.0195]$

Table 25: Ten-seed method-side INLP check for the borderline ranks. Rows combine five controlled seeds with five additional real-edit seeds; placebo percentiles remain from Table 24.

Pair diagnostic	Leakage after	Acc. Δ	Fair. benefit	Logit pct.	Val-drop pct.
Pair-PC linear edit	0.9888	+0.0026	−0.0006	0.4860	0.5260
Paired interpolation	0.6687	+0.0037	+0.0052	0.6880	0.8520

Table 26: Bios counterfactual-pair diagnostics over five seeds.

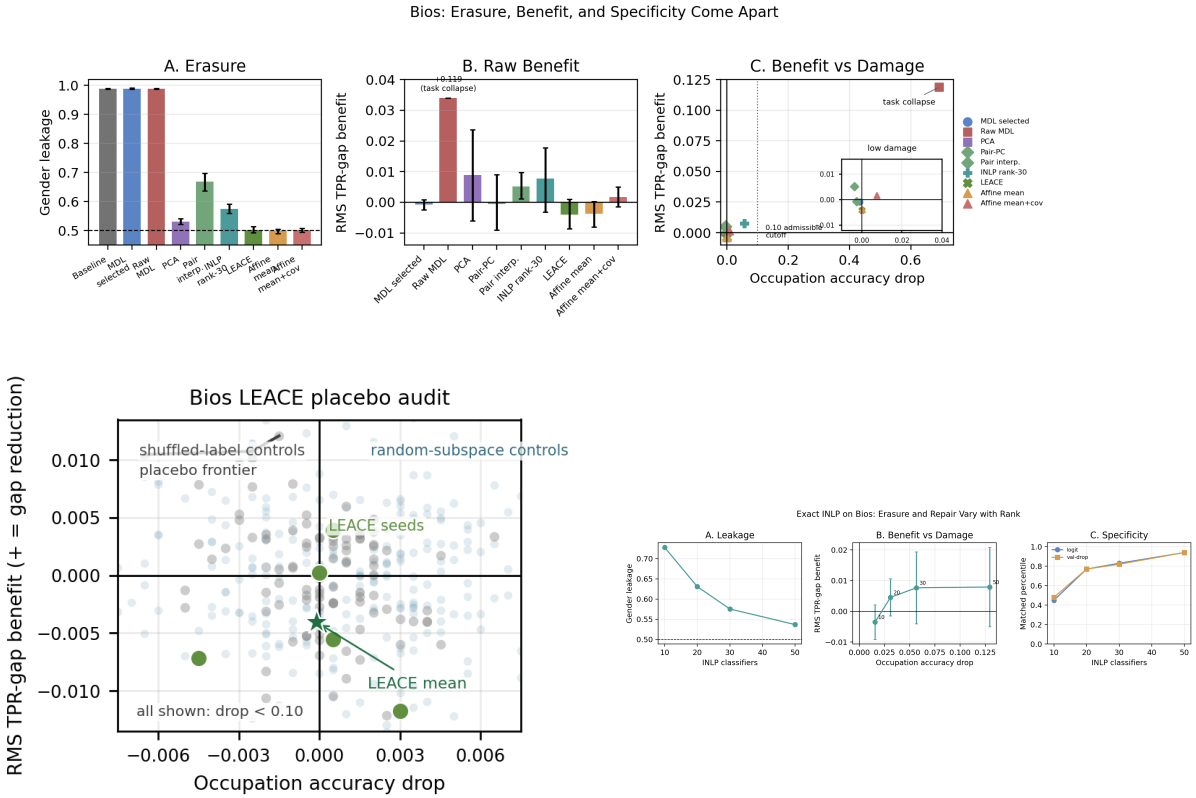


Figure 5: Appendix figure summary. Top: Bios claim-separation view. Bottom left: Bios LEACE against matched shuffled-LEACE controls. Bottom right: INLP rank sweep.