Separating Input-Carried from Model-Added Demographic Encoding in Language Models

Minh K. Quan, Pubudu N. Pathirana

School of Engineering, Deakin University, Waurn Ponds, VIC 3216, Australia
{m.quan, pubudu.pathirana}@deakin.edu.au

Abstract

When a language model encodes a sensitive attribute, the source matters: the signal may be carried by explicit or distributed input cues, or it may remain after controlling for a declared input representation. Standard probing conflates these cases, leaving auditors unsure whether to inspect the input pipeline or the learned representation. We propose Controlled Representation Attribution (CRA), a reference-conditional auditing protocol that residualizes each layer against an input reference and probes the residual. We use "model-added" conservatively to denote reference-residual sensitive information rather than a causal claim about model weights. CRA is supported by a recovery theorem—exact when the residual noise is equal-covariance Gaussian and the target direction is an eigenvector of that covariance, with an explicit finite-anisotropy bound otherwise—and by scope and reference-admissibility diagnostics that identify inadequate controls. In controlled simulations with known ground-truth directions, CRA wins the selectivity comparison in 60/60 runs; falsification tests further show that weak references can create spurious residual conclusions, while the admissibility gate withholds such claims. In a prospectively specified controlled-text audit, residual evidence survives name removal but not removal of broader lifestyle and stylistic cues, demonstrating dependence on the declared control. Archived pretrained-language-model analyses illustrate a lexical-gender/reference-residual-age contrast and a positive pooled medical-question-answering difference (22 of 27 cells, mean $+5.3$ percentage points), which is not powered at the dataset level; these analyses are exploratory rather than confirmatory. CRA is a source-attribution diagnostic that reports reference sensitivity and can abstain when controls are inadequate; it complements concept-erasure methods rather than replacing them.

Introduction

Responsible use of language models in sensitive domains requires understanding not only *whether* a representation predicts a sensitive attribute, but whether that prediction remains after a defensible, declared input control (Obermeyer et al. 2019; Zhang et al. 2020; Chen, Johansson, and Sontag 2021). A clinical representation may predict age because the text contains explicit age cues, because contextual computation combines symptoms into an age-correlated pattern, or because pretraining changes how those cues are represented. A standard probe on Z_l conflates these cases.

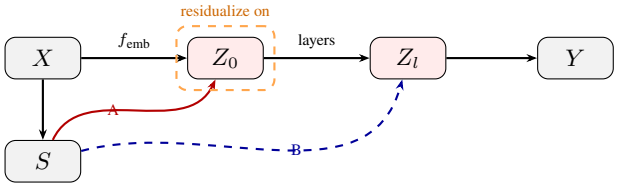

Figure 1: Two information paths for S into Z_l. Path A (red): variation explained by the declared Z_0 reference. Path B (blue, dashed): sensitive signal remaining after that control. CRA residualizes on Z_0 (orange); the decomposition is conditional on the audit specification.

We study a *reference-conditional* distinction. **Path A (reference-explained)** is the component of the layer representation predictable from a prespecified reference Z_0 under a prespecified residualizer. **Path B (reference-residual)** is sensitive information remaining after that control. We retain "model-added" as the motivating contrast in the registered title, but operationalize the identified quantity as "reference-residual": the residual may contain learned associations, nonlinear input functions, contextual interactions, pooling effects, or residualizer misspecification. A positive result is therefore a screen for follow-up analysis, not evidence that weights caused the signal or that an intervention is beneficial.

An audit result is indexed by $\mathcal{A} = (Z_0, g, k, \alpha, \mathcal{R}, \mathcal{P}, l)$: reference, pooling rule, rank, regularization, residualizer family, probe family, and layer. The result is a property of this audit specification, not an intrinsic label of a model–attribute pair. To limit researcher degrees of freedom, all primary and sensitivity specifications must be prospectively specified before inspecting sensitive-attribute outcomes; contradictory results are reported as reference-sensitive rather than collapsed to a preferred verdict.

A high reference-explained component suggests inspecting the input pipeline or reference choice; a high residual component motivates a deeper mechanistic or counterfactual analysis before any intervention. We present CRA (**C**ontrolled **R**epresentation **A**ttribution), a protocol that combines conditional probing with explicit diagnostics for when a residual direction adds information beyond ordinary

probing.

Contributions. **(1) Admissible-reference auditing.** We turn reference choice from an unreported degree of freedom into a falsifiable development-stage gate. A reference may support an origin-like claim only if it passes prespecified tests for sensitive-cue coverage, $Z_0 \to Z_l$ fit, Path-B preservation, and cross-fold direction stability; otherwise its result remains descriptive. The protocol adds a third verdict—*reference unresolved*—when no control is scientifically adequate. **(2) Reference-consensus direction.** For the surviving family, we aggregate sign-invariant rank-one projectors rather than signed coefficients and return their leading eigenvector together with an eigengap. Theorem 2 gives an approximate-condition recovery guarantee; simulations quantify the gain from multiple imperfect references. **(3) Claim-level finite-sample inference.** A fixed-score Monte Carlo permutation test supports cell-level detection; an intersection–union rule supports "all mandatory controls," while Holm or joint max-T calibration supports exploratory "any-cell" claims. Template-stratified permutation makes the exchangeability assumption explicit for generated text. **(4) Decision-changing falsification studies.** Reference-misspecification simulations show that a weak control can create a residual origin claim even when Path B is absent, and that the admissibility gate prevents it. A controlled natural-form lockbox identifies which distributed cue family is necessary rather than selecting the most favorable reference. **(5) Reliable negative intervention evidence.** We expand the intervention from three to 12 initializations, report paired uncertainty and every seed, and add an entangled-task Pareto study. Neither establishes superiority over the ordinary direction; the protocol therefore withholds that claim and treats the direction as actionable only when reference agreement and intervention evidence both support it.

What is and is not new. Residualization, conditional probes, logistic directions, randomization inference, and averaging projectors are individually familiar. The scientific novelty is the *decision system* formed by their coupling: references are screened without lockbox labels; a robust claim ranges only over admissible controls; direction signs are removed before aggregation; the eigengap determines whether a common action is defined; and the audit can return positive, negative, reference-sensitive, or reference-unresolved. A preregistered conditional probe plus sensitivity analysis can reproduce each scalar cell, but it cannot by itself justify that the selected control was adequate or return a reference-consensus feature-space action. When the consensus and ordinary directions coincide or a paired intervention is inconclusive, CRA explicitly reports that ordinary conditional probing is sufficient.

Related Work

Conditional probing and concept erasure. Linear probes reveal information in intermediate representations (Alain and Bengio 2017; Tenney, Das, and Pavlick 2019), but a probe on Z_l conflates signal already present in the input with signal that remains after a declared control (Belinkov 2022). Structural, information-theoretic, and amnesic probes already implement related controls (Hewitt and Manning 2019; Pimentel et al. 2020; Elazar et al. 2021); therefore CRA's scalar residual AUROC is a standard conditional-probing quantity, not a new estimator. INLP, LEACE, and SentDebias remove total linear signal rather than assess signal conditional on a reference (Ravfogel et al. 2020; Ravfogel, Goldberg, and Goldwasser 2022; Belrose et al. 2023; Liang et al. 2020). CRA is diagnostic: it asks whether a residual conclusion is stable across scientifically adequate references before any erasure is considered.

Directions, interventions, and causal limits. Work on linear concept directions and representation engineering motivates extracting a feature-space direction (Park, Choe, and Veitch 2023; Zou et al. 2023), while activation editing and causal mediation show how directions can seed follow-up interventions (Vig et al. 2020; Meng et al. 2022, 2023). A direction is not automatically causal or beneficial. Our contribution is to screen references without lockbox labels, aggregate signs through projectors, quantify cross-reference agreement through an eigengap, and withhold an actionable direction when adequate references disagree.

Fairness and clinical NLP. Bias audits in clinical and general language models document demographic dependence and downstream harm (Obermeyer et al. 2019; Zhang et al. 2020; Chen, Johansson, and Sontag 2021). CRA does not replace fairness, privacy, or utility evaluation. It supplies a narrower decision: whether sensitive predictability survives a prespecified family of admissible controls, and whether those controls support one stable direction for a subsequent mechanism study.

Framework

Reference-Conditional Two-Path Model

Let $X \in \mathcal{X}$ be an input text sequence, $S \in \{0,1\}$ a sensitive attribute, and Y the downstream target. Let $Z_0 = f_{\text{emb}}(X) \in \mathbb{R}^d$ be the declared input reference and $Z_l = f_l(X) \in \mathbb{R}^d$ the audited representation. Our experiments use mean-pooled token-plus-position embeddings as Z_0, but the estimand changes if the reference or pooling changes.

Definition 1 (Reference-conditional two-path model)**.** For a fixed audit specification $\mathcal{A}$, model the layer representation as

$$Z_l = A_l Z_0 + B_l S + \varepsilon_l, \qquad \mathbb{E}[\varepsilon_l \mid Z_0, S] = 0, \quad (1)$$

where $A_l Z_0$ is the component explained by the selected reference (Path A), and $B_l S$ is the S-predictive component not represented by that linear map (Path B). The names describe this statistical decomposition under $\mathcal{A}$; they do not identify a causal mechanism.

Remark 1 (Reference and readout dependence)**.** The title-level phrase "model-added" is operationalized as "reference-residual" and is not treated as a causal quantity. The residual can include learned weight-level associations, nonlinear functions of input information, contextual token interactions, pooling artifacts, or misspecification. Changing Z_0, pooling, residualizer, or probe changes the estimand. A positive result is therefore conditional statistical evidence relative to $\mathcal{A}$, not a Pearl-SCM arrow and not proof that weight editing is required.

Linear-control diagnostics. We report the ridge-PCA fit $R^2 = 1 - \|Z_l - \widehat{Z}_l(Z_0)\|_F^2 / \|Z_l - \bar{Z}_l\|_F^2$. Across the three distilGPT-2 configurations emphasized in the main analysis, the selected linear control explains $R^2 \in [0.67, 0.99]$ of layer variance (mean 0.85 across layers 1–5). This is an operational measure of how much total variation the control captures; it is *not* by itself a bound on sensitive Path-A contamination. We therefore also report (i) residual-AUROC stability as k increases, (ii) a direct Z_0 probe, (iii) held-out residual AUROC with uncertainty, and (iv) nonlinear and bag-of-words controls (Supplement §18–21).

The rank-one condition is well supported only for binary gender in the reported SVD check ($\sigma_1/\sigma_2 > 17{,}000$). Age group and hospital type are multi-directional ($\sigma_1/\sigma_2 \approx 1$–$2$), so their single-direction HSIC results are exploratory; for those attributes we emphasize the direction-free residual-AUROC diagnostic and report iterative rank-r deflation as an empirical extension rather than as part of the main theorem.

Outputs, Prespecification, and Reference Admissibility

CRA separates **residual detection** from **directional action**. Detection asks whether S remains predictable after a declared control and can apply to multidirectional attributes. Directional action additionally requires a stable rank-one direction, agreement across adequate references, and intervention evidence. For every single-reference direction we report $\delta_l = 1 - |\cos(\boldsymbol{v}_{\text{cra}}, \boldsymbol{v}_{\text{corr}})|$; near zero means no incremental directional value.

Before sensitive lockbox outcomes are inspected, the auditor records a reference ladder, residualizer, probe, layer, multiplicity family, and development-only admissibility thresholds. For reference j, let A_j be direct-reference AUROC, R_j^2 the held-out $Z_0^{(j)} \to Z_l$ fit, $\eta_j = \|(I - \Pi_j)s\|/\|s\|$ the preserved label amplitude, and K_j the median absolute cross-fold cosine of the residual direction. In the controlled studies we prospectively admit reference j only if

$$A_j \geq .65, \qquad R_j^2 \geq .08, \qquad \eta_j \geq .20, \qquad K_j \geq .50. \tag{2}$$

These are design-stage defaults for an origin-like cue audit, not universal constants. Applications must justify and freeze domain-specific equivalents. A rejected reference is still reported, but it cannot support the robust origin-like claim. If none is admitted, the verdict is *reference unresolved*; if admitted references disagree statistically, it is *reference-sensitive*. This gate directly operationalizes "strong enough but not over-aggressive": A_j and R_j^2 reject weak controls, while η_j rejects controls that remove nearly all candidate residual signal.

Reference-Consensus Direction

Suppose m admissible references yield unit directions v_j. Because a linear direction is defined only up to sign, signed coefficient averaging is ill-posed. We instead form

$$M = \sum_{j=1}^{m} w_j v_j v_j^\top, \quad w_j \geq 0, \quad \sum_j w_j = 1, \qquad \boldsymbol{v}_{\text{cons}} = \text{eig}_1(M). \tag{3}$$

Algorithm 1 Single-reference CRA cell estimator

Require: Representations $Z_l, Z_0 \in \mathbb{R}^{n \times d}$; labels $\boldsymbol{s} \in \{0,1\}^n$; ridge α; PCA rank k

$\widetilde{Z_0} \leftarrow \text{PCA}_k(Z_0)$

$Z_l^\perp \leftarrow Z_l - \widetilde{Z_0}(\widetilde{Z_0}^\top \widetilde{Z_0} + \alpha I)^{-1} \widetilde{Z_0}^\top Z_l$ *// control for selected reference*

$\boldsymbol{v}_{\text{cra}} \leftarrow \text{normalize}(\text{LogReg}(Z_l^\perp, \boldsymbol{s}).\text{coef\_})$ *// partial probe on residual*

return $\boldsymbol{v}_{\text{cra}}$ *// reference-residual direction*

The leading eigenvalue and eigengap $\gamma = \lambda_1(M) - \lambda_2(M)$ quantify reference agreement. Small γ means that no common one-dimensional action is identified, even when individual residual probes are significant. We use uniform weights in the completed experiments; weights based on frozen development-only stability are also permitted.

CRA Cell Estimator

Let $V \in \mathbb{R}^{d \times k}$ be the matrix whose columns are the top-k PCA basis vectors of Z_0 (feature space), and let $\widetilde{Z_0} = Z_0 V \in \mathbb{R}^{n \times k}$ be the projected sample matrix (we use $k = 50$; sensitivity in Supplement §5). The CRA procedure has two steps:

Step 1 ridge-regresses Z_l on $\widetilde{Z_0}$ and subtracts fitted values (Eq. 4), aiming to reduce the reference-explained component. How much sensitive signal is preserved depends on the sample-space overlap quantified below:

$$Z_l^\perp = Z_l - \widetilde{Z_0}\big(\widetilde{Z_0}^\top \widetilde{Z_0} + \alpha I\big)^{-1} \widetilde{Z_0}^\top Z_l. \tag{4}$$

Sample- vs. feature-space. Feature-space partialling yields identical residuals under OLS, but at $n \approx 400 \ll d = 768$ the sample-space ridge form is better-conditioned ($k \ll n$; Supplement §5). **Step 2** probes the residual:

$$\boldsymbol{v}_{\text{cra}} = \text{normalize}\big(\text{LogReg}(Z_l^\perp, \boldsymbol{s}).\text{coef\_}\big). \tag{5}$$

Let $\Pi_\alpha = \widetilde{Z}_0(\widetilde{Z}_0^\top \widetilde{Z}_0 + \alpha I)^{-1} \widetilde{Z}_0^\top$ and $\boldsymbol{r}_\alpha = (I - \Pi_\alpha)\boldsymbol{s}$. The following result separates the always-valid algebraic decomposition from stronger sufficient conditions for directional recovery.

Theorem 1 (Residual decomposition and rank-one recovery)**.** *In matrix form, under Eq.* (1)*,*

$$Z_l^\perp = (I - \Pi_\alpha) Z_0 A_l^\top + \boldsymbol{r}_\alpha B_l^\top + (I - \Pi_\alpha) E_l. \tag{6}$$

*Thus the fraction of the sample-space Path-B amplitude preserved by the control is $\eta_\alpha = \|\boldsymbol{r}_\alpha\|_2 / \|\boldsymbol{s}\|_2$. Exact **C1**, $\widetilde{Z}_0^\top \boldsymbol{s} = 0$, is the special case $\eta_\alpha = 1$; it is sufficient, not necessary, for residual signal detection.*

*For direction recovery, additionally suppose: **(C1)** exact preservation; **(C2)** $B_l = b\boldsymbol{u}$ with $\|\boldsymbol{u}\| = 1$; **(C3)** the population limit; **(C4)** the reference-explained term in Eq.* (6) *is negligible; and **(C5)** $Z_l^\perp \mid S = s \sim \mathcal{N}(bs\boldsymbol{u}, \Sigma)$ with the same positive-definite within-class covariance Σ for both classes. Then the population logistic direction is*

$$\boldsymbol{w}^\star \propto \Sigma^{-1} \boldsymbol{u}. \tag{7}$$

It exactly aligns with $\boldsymbol{u}$ whenever $\boldsymbol{u}$ is an eigenvector of Σ (including spherical covariance). For arbitrary Σ, including nonzero cross-covariance between $\boldsymbol{u}$ and $\boldsymbol{u}^{\perp}$,

$$\cos(\boldsymbol{w}^{\star}, \boldsymbol{u}) \geq \frac{2\sqrt{\kappa(\Sigma)}}{\kappa(\Sigma)+1}, \tag{8}$$

where $\kappa(\Sigma)$ is the condition number of the full within-class covariance.

Proof sketch. Equation (6) follows by left-multiplying the matrix form of Eq. (1) by $I - \Pi_{\alpha}$. Under exact C1 and C4 the class means differ only along $\boldsymbol{u}$. Equal-covariance Gaussian classes give the standard LDA/logistic solution $\boldsymbol{w}^{\star} \propto \Sigma^{-1}(\mu_1 - \mu_0) \propto \Sigma^{-1}\boldsymbol{u}$. Exact alignment follows when $\boldsymbol{u}$ is a covariance eigenvector. The general bound is the Kantorovich angle bound for a positive-definite linear map, applied to Σ^{-1}; unlike the previous restricted argument, it uses the full covariance and permits cross terms. Full details are in Supplement §1. □

Theorem 2 (Consensus stability under approximate references)**.** *Let u be a unit target direction, $P_{\star} = uu^{\top}$, $P_j = v_j v_j^{\top}$, and $M = \sum_j w_j P_j$. If $\|P_j - P_{\star}\|_2 \leq \epsilon_j$ and $\bar{\epsilon} = \sum_j w_j \epsilon_j < 1/2$, then*

$$\|M - P_{\star}\|_2 \leq \bar{\epsilon}, \quad \lambda_1(M) \geq 1 - \bar{\epsilon}, \quad \lambda_2(M) \leq \bar{\epsilon}, \tag{9}$$

so $\gamma \geq 1 - 2\bar{\epsilon} > 0$ and the leading eigenvector is unique up to sign. Moreover,

$$\sin \angle(\boldsymbol{v}_{\text{cons}}, u) \leq \frac{\bar{\epsilon}}{1 - \bar{\epsilon}}. \tag{10}$$

Proof sketch. Convexity of the operator norm gives the first inequality, and Weyl's inequalities give the two eigenvalue bounds. Write $\boldsymbol{v}_{\text{cons}} = au + b$ with $b \perp u$. Since $\lambda_1(M) \geq 1 - \bar{\epsilon}$ and the orthogonal component of $P_{\star}\boldsymbol{v}_{\text{cons}}$ is zero, $\lambda_1(M)\|b\| \leq \|(M - P_{\star})\boldsymbol{v}_{\text{cons}}\| \leq \bar{\epsilon}$, yielding Eq. (10). Full details are in the supplement. □

The theorem is deliberately approximate and reference-level. It does not validate any individual residualizer; it states what aggregation gains once development diagnostics bound the errors of the admitted directions. In 500-replication simulations at angular-noise scale 0.45, mean $1 - |\cos|$ error was 0.095 for an individual direction, 0.051 for two-reference consensus, 0.021 for four, and 0.010 for eight.

Remark 2 (Approximate C1 and synthetic interpretation)**.** When C1 fails, Eq. (6) remains exact: the control replaces $\boldsymbol{s}$ by $\boldsymbol{r}_{\alpha}$, rather than preserving Path B unchanged. The main synthetic generator contains a Path-A term correlated with S, so sample-space C1 is only approximate; what is exact by design is feature-space orthogonality of the known $\boldsymbol{v}_B$ to the input subspace. We report both facts and interpret the 60/60 result as empirical selectivity under approximate C1, not as an instance of the exact-recovery theorem. A separate overlap stress test rotates $\boldsymbol{v}_B$ toward the input feature subspace; it is not mislabeled as a direct sweep of $\widetilde{Z}_0^{\top}\boldsymbol{s}$.

Remark 3 (Finite samples and multidirectional attributes)**.** At $n/d \approx 0.5$, the recovered direction is materially less stable than the residual-AUROC verdict; the theorem is asymptotic. Moreover, C2 holds only for rank-one attributes. For multiclass age and hospital type, residual AUROC can establish detectable residual information, but one single-direction HSIC reduction cannot recover the full residual subspace. Iterative rank-r deflation is evaluated empirically in the supplement; a general finite-sample subspace theorem is open.

Finite-Sample Calibration and Decision Rules

All model training and development choices are completed before lockbox evaluation. These include representation extraction, residualizer fitting, probe fitting, sign orientation, and score construction. The untouched lockbox therefore receives a fixed score vector $a = (a_1, \ldots, a_n)$. For a one-sided statistic $T(a, S) = \text{AUROC}(a, S)$ and B random permutations π_b of the lockbox labels, we use

$$p_B = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{T(a, \pi_b S) \geq T(a, S)\}}{B + 1}. \tag{11}$$

The bootstrap interval for AUROC is an effect-size summary; it is no longer compared with a development-derived extreme quantile to make a binary decision.

Proposition 1 (Lockbox validity and claim-level combination)**.** *Condition on every training/development choice and assume that, under the cell null, lockbox labels are exchangeable relative to the fixed scores. Then Eq.* (11) *is super-uniform, including with Monte Carlo permutations. For K prespecified mandatory controls, the intersection–union p-value $p_{\text{env}} = \max_k p_k$ controls the size of the robust claim that every mandatory condition contains residual signal. For an existential claim that some cell is positive, the cell p-values must instead be adjusted, for example by Holm's procedure or a joint max-T permutation null.*

This proposition supplies the finite-sample result that the directional theorem cannot: it validates *detection* without Gaussianity, equal covariance, rank one, or $n \gg d$. Its assumptions are exchangeability and strict separation of development from lockbox scoring. The proof uses the standard corrected Monte Carlo randomization-rank argument (Phipson and Smyth 2010); the intersection–union statement follows because, under its union null, at least one component null is true. Supplement §1 gives the proof and discusses clustered or paired permutations when examples are not exchangeable.

Exchangeability under shared text templates. Ordinary example-level permutation is valid only if the lockbox labels are exchangeable relative to fixed scores. Generated biographies share 12 template–task families, so we additionally permute labels only within blocks defined by template ID × task label. The full-text and name-scrubbed word-reference cells remain $p = .0002$ for every seed under both schemes. The lifestyle-scrubbed condition is not robust ($\max p_{\text{block}} = .0628$), and the all-scrubbed condition remains null ($\max p_{\text{block}} = .6966$). Thus the primary controlled-text decision is not an artifact of treating repeated template families as independent exchangeable records.

Table 1: Null calibration over 20,000 simulations with 18 correlated cells ($\rho = 0.45$), nominal $\alpha = 0.01$. The old unadjusted cell search does not control the existential claim.

Procedure	Claim	False positive
Unadjusted minimum p	any cell	10.77%
Holm	any cell	0.89%
Joint max-T	any cell	1.00%
IUT max p_k	all 18 cells	$< 0.01\%$

Table 2: Reference adequacy changes the scientific decision. Rates are over 250 runs per cell.

Path B	Reference	Admitted	Naive +	Gated +
0	adequate	1.00	.016	.016
0	weak	0.00	1.000	.000
0	over-aggressive	0.00	.000	.000
.8	adequate	1.00	1.000	1.000
.8	weak	0.00	1.000	.000
.8	over-aggressive	0.00	.032	.000

Why 200 permutations were insufficient. In 5,000 null repetitions, estimating a 0.99 standard-normal quantile from 200 draws had standard deviation 0.283 and a 5–95% range of 1.975–2.898; with 5,000 draws the standard deviation fell to 0.054 and the range to 2.241–2.416. The minimum attainable p-values are respectively 0.00498 and 0.00020. We therefore use $B = 4{,}999$ for primary 1% tests and report the exact Monte Carlo formula.

The all-controls envelope is conservative by design and can have low power. With mean per-cell power 0.81 and correlation 0.45, simulated envelope power was 0.62 for three mandatory cells, 0.47 for six, and 0.27 for eighteen. Consequently, the revised manifest distinguishes a small number of scientifically necessary controls from replication seeds/layers that are reported continuously rather than indiscriminately added to one all-pass gate.

Experiments

Reference-Gate Falsification Study

We generated representations with either no Path B ($b = 0$) or a rank-one residual signal ($b = .8$) and compared adequate, weak, and over-aggressive references over 250 replications per cell. The naive rule declares a residual whenever its permutation test rejects; the gated rule additionally requires Eq. (2). Under $b = 0$, the weak reference creates a positive decision in 100% of runs because it fails to remove Path A; the gate rejects all 250. An adequate reference is admitted in all runs and yields a 1.6% positive rate with 499 permutations at nominal 1%. The over-aggressive reference is rejected by $\eta \approx .07$ and cannot create false reassurance when $b = .8$.

This study supplies the concrete incremental conclusion unavailable from one conditional probe: *the apparent residual can be attributed to an inadequate control rather than promoted to a model-origin claim*. Rejection of a weak reference sacrifices power when no adequate alternative exists; the correct output is then unresolved, not negative.

Table 3: Prospective controlled-text audit after development-only reference screening. "Block p" is the maximum across three word-reference seeds using template×task blocks.

Condition	Admitted	Residual AUC	Block p	Verdict
Full text	5/6	.802–.833	.0002	robust residual
Names removed	5/6	.783–.811	.0002	robust residual
Lifestyle/style removed	5/6	.519–.543	.0628	not robust
All cue families removed	0/6	.485–.524[†]	.6966	unresolved/null

[†] Descriptive word-reference range; no reference passes the cue-coverage gate.

Confirmatory Prospective Controlled-Text Lockbox

Data and frozen design. We generated 3,200 biography-style English records before fitting any audit: 1,800 train, 600 development, and 800 untouched lockbox examples. The task is project success. The binary audit attribute is never emitted as a label word and is distributed across first names, lifestyle/community phrases, and writing style. A spurious attribute–task association appears only in training; development and lockbox are balanced. A word TF–IDF/SVD encoder feeds a two-hidden-layer classifier. This is controlled natural-form text, not a real corpus or pretrained-LM study.

The frozen ladder contains weak, word, character, and cue-scrubbed references. The weak and scrubbed controls are deliberately included as falsification controls, not mandatory references. Seeds $\{11, 29, 47\}$, the final hidden layer, four input conditions, ridge residualization, logistic probes, 4,999 permutations, and the gate thresholds are fixed before lockbox scoring. The word and character references are admitted in five of six seed–reference cells for full and name-scrubbed text; the remaining character cell misses the frozen R^2 threshold by 0.005. Weak references fail R^2 and scrubbed references fail cue coverage despite producing residual AUROC as high as .996, illustrating why a large residual alone is not evidence of origin.

Decision. The signal is not driven by a single name token: it survives name removal under both example-level and block permutation. It does depend on the lifestyle/style family, and joint cue removal yields chance-level residual scores. The protocol therefore reports *distributed but control-sensitive residual encoding*. Crucially, it does not select the weak reference that produces the largest residual, and it does not interpret the cue-scrubbed reference's near-perfect residual as model-added signal. The full ladder, A_j, R^2_j, η_j, stability, and every rejected reference appear in the supplement and machine-readable output.

Directional Interventions: Reliability and Entanglement

For each initialization, we project the lockbox representation to the hyperplane orthogonal to the consensus direction, the ordinary sensitive-probe direction, or a norm-matched random direction, then refit an independent sensitive probe while leaving the task head fixed. The expanded study uses 12 initializations rather than three. Mean

Table 4: Seed-wise sensitive-AUROC reduction. Positive values indicate less re-probe-accessible signal. The final column is consensus minus ordinary.

Seed	Cons.	Ord.	Diff.	Seed	Cons.	Ord.	Diff.
1	.053	.054	-.001	7	.111	.094	.018
2	.125	.114	.012	8	.197	.197	-.000
3	.090	.091	-.000	9	-.005	-.005	-.000
4	.110	.113	-.003	10	-.017	.066	-.083
5	-.013	-.010	-.003	11	-.010	-.009	-.001
6	.000	-.001	.001	12	.160	.135	.025
Mean: .0668 / .0699 / -.0031				95% CI(diff): [-.0198,.0082]			

sensitive-AUROC drops are .0668 for consensus, .0699 for ordinary, and $-.0029$ for random. The paired consensus-minus-ordinary difference is $-.0031$ with a 95% paired bootstrap CI $[-.0198, .0082]$ and sign-flip $p = .88$; the paired task-loss difference is 2.1×10^{-5} with CI $[0, 5.0 \times 10^{-5}]$. Comparative superiority is therefore rejected.

To make task preservation nontrivial, a second generator correlates the task with the sensitive cue families in development and lockbox data and weakens the direct task phrase. Baseline task AUROC is .845 and sensitive AUROC .924. At full one-direction removal, consensus and ordinary reduce sensitive AUROC by only .0116 and .0109, respectively; their paired difference is .00068 (95% CI $[-.00081, .00249]$, $p = .53$), while mean task changes remain below 10^{-4}. Across strengths $.25, .5, .75, 1.0$, neither direction traces a reliably superior Pareto frontier. This harder negative result shows that near-zero task loss in the original easy setting was not evidence of selective mechanism removal: one linear direction is simply too weak when task and attribute information are entangled.

The intervention now supports only feasibility and an abstention rule. A direction may be used for a follow-up mechanism experiment, but this paper does not establish that consensus erasure improves fairness, privacy, task utility, or selectivity relative to ordinary probing.

Assumption-by-Assumption Stress Test

We varied $n/d \in \{0.5, 1, 2, 5\}$, label–reference overlap, and covariance condition number in 400 controlled runs. Detection and direction recovery separate sharply: in the favorable setting ($n/d = 5$, no overlap, $\kappa = 1$), covariance-corrected direction cosine is 0.94; at $n/d = 0.5$, stronger overlap ($\eta \approx 0.54$), and $\kappa = 10$, cosine falls to 0.29 even though residual AUROC remains 0.88. This directly tests the theorem assumptions and supports a narrower claim: Theorem 1 explains an idealized rank-one regime, while Proposition 1 justifies finite-sample detection in the central empirical setting.

Secondary and Archived Evidence

The structured-token diabetes experiment and all pretrained-LM analyses are now explicitly secondary. In the former, all 18 age cells are positive on full input and none after masking the explicit age token; recorded sex yields 1/18 positives and none after masking. Because the original scripts, checkpoints, split indices, and per-cell outputs were not uploaded, these values are retained as an archived cue-removal stress test rather than confirmatory evidence. Full tables and preservation diagnostics appear in the supplement.

Table 5: Per-result assumption audit. "Yes" means the diagnostic is sufficiently supported for the stated interpretation, not that the assumption is mathematically verified.

Result	C1/η	Rank 1	C4	Cov.	Direction claim
Natural-text full	partial	yes	uncertain	approx.	intervention only
Structured age	weak	no	uncertain	no	none
Archived gender	mixed	yes	uncertain	mixed	exploratory
Archived age/hospital	mixed	no	uncertain	mixed	none
Synthetic favorable	yes	yes	yes	yes	theorem-linked
Synthetic stressed	measured	yes	violated	varied	empirical only

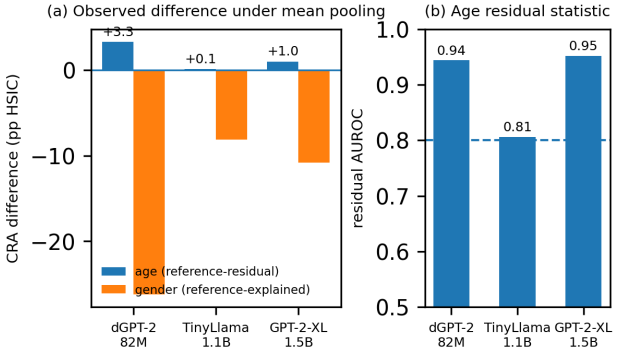

Figure 2: Archived real-attribute diagnostic across three scales. Residual age evidence remains high, whereas the CRA–ordinary HSIC differential is small and approximately zero on the $\geq$1B models. This descriptive figure motivates reference-sensitive reporting; it is not confirmatory evidence of origin or intervention benefit.

The archived pretrained-LM study spans six general and clinical models and several QA datasets, but some transformations and layer choices were post hoc. Its defensible conclusion is therefore qualitative: residual conclusions and candidate directions depend on the declared reference and architecture, and at larger scales $|\cos(\boldsymbol{v}_{\text{cra}}, \boldsymbol{v}_{\text{corr}})| \approx .99$ can make ordinary conditional probing sufficient. Figure 2 is retained for comparability but does not enter the revised confirmatory decision rule.

Operational Reference Selection and Scaling

A reference ladder is frozen before lockbox scoring and should include the audited model's own input or embedding readout, an independent lexical control, and a richer neighboring control such as a character or alternate-pooling reference. Cue-removal counterfactuals are separate estimands. The mandatory family is the smallest set containing at least two scientifically distinct admissible controls; weak or destructive references fail Eq. (2), no admitted reference yields *unresolved*, and a small consensus eigengap yields *reference-sensitive*. This avoids both cherry-picking and an unnecessarily low-power all-controls envelope. Residualiza-

tion costs roughly $O(K(ndk+k^3))$, fixed-score permutation costs $O(KBn)$ after ranking, and projector consensus adds $O(md^2)$; measured CPU timings and memory guidance are reported in the supplement.

Discussion and Conclusion

CRA's novel object is a reference-robust decision protocol, not a stronger residual estimator or eraser. The admissibility gate changes the decision in the exact failure mode that motivates the paper: with no Path B, a weak control produces positive residual tests in 100% of simulations, while the gate rejects every claim. The sign-invariant consensus direction is returned only when multiple admissible references agree, and its eigengap makes that agreement testable. The controlled-text result survives name removal and template-blocked permutation but not removal of the lifestyle/style cue family. By contrast, the expanded intervention gives a disciplined negative result: consensus does not reliably outperform ordinary probing, including under task–attribute entanglement. The incremental scientific value is therefore the ability to falsify inadequate controls, expose reference disagreement, and specify when no common direction is actionable.

Decision hierarchy beyond a conditional probe. A preregistered conditional probe establishes one cell: sensitive information remains after one chosen control. The revised protocol supports a stronger but narrower conclusion. First, each reference must pass cue coverage, held-out representation fit, signal preservation, and directional stability without consulting lockbox labels. Second, the robust residual claim must hold across the smallest scientifically distinct admitted family, rather than across whichever reference gives the largest effect. Third, a feature-space action exists only when the admitted projectors have a clear leading eigendirection. These requirements distinguish four outcomes that one probe cannot: *robust residual*, *reference-sensitive*, *reference-unresolved*, and *no residual evidence*. The first is still reference-conditional, not causal; the middle two are scientifically informative abstentions rather than failed analyses.

Operational adequacy and power. The controlled-study gate ($A_j \geq .65$, $R^2_j \geq .08$, $\eta_j \geq .20$, $K_j \geq .50$) is a prespecified default, not a universal threshold. Its components have distinct roles: direct-reference AUROC checks whether the control contains the relevant cue; held-out R^2 checks whether the chosen residualizer can express the reference–layer relation; η_j rejects destructive controls that remove nearly all candidate signal; and K_j rejects directions that are not reproducible across development folds. At least two scientifically different admitted references are required for a multi-reference claim. Extra seeds and layers are replication cells and receive multiplicity correction, but they are not automatically added to the all-pass conjunction. This separation keeps the gate conservative about control validity without making power vanish as every conceivable sensitivity check is added.

Required external validation. The next decisive study is already specified as a frozen protocol: a naturally occurring biography corpus, a frozen pretrained encoder and audited layer, profession as the utility task, a demographic attribute, train/development/lockbox splits fixed before attribute scoring, lexical and own-model references, block-aware permutation, and paired consensus-versus-ordinary interventions. Success would require both an admissible multi-reference residual conclusion and a reproducible downstream action; failure or direction equivalence would establish the boundary where ordinary conditional probing is sufficient. We do not report this study as completed because the required corpus and model weights were unavailable in the execution environment.

The central limitation is external validity. No experiment here simultaneously uses a naturally occurring corpus, a frozen pretrained language model, and a protocol locked before sensitive outcomes; the archived pretrained-LM evidence remains descriptive. Gate thresholds are controlled-study defaults rather than universal adequacy constants, and rejecting weak references can reduce power. The consensus theorem assumes admitted projectors are close to one target; an eigengap diagnoses agreement but not causality. Multi-directional attributes, nonlinear mechanisms, and low-n/d regimes remain outside the rank-one recovery claim. Intervention effects concern probe-accessible information and do not imply fairness, privacy, clinical utility, or beneficial decisions.

Ethical statement. The controlled text is synthetic and contains no personal data; stereotyped cue correlations are used only to stress-test the audit. The archived public datasets contain no private health information in this submission. CRA is not a privacy guarantee, causal test, or replacement for fairness, harm, and downstream-performance audits.

References

Alain, G.; and Bengio, Y. 2017. Understanding Intermediate Layers Using Linear Classifier Probes. In *ICLR Workshop*.

Belinkov, Y. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1): 207–219.

Belrose, N.; Ravfogel, S.; Marks, W.; Lindsey, D.; and Heimersheim, A. 2023. LEACE: Perfect Linear Concept Erasure in Closed Form. In *NeurIPS*.

Chen, I. Y.; Johansson, F. D.; and Sontag, D. 2021. Why Is My Classifier Discriminatory? *Annual Review of Biomedical Data Science*.

Elazar, Y.; Ravfogel, S.; Jacovi, A.; and Goldberg, Y. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *TACL*, 9: 160–175.

Hewitt, J.; and Manning, C. D. 2019. A Structural Probe for Finding Syntax in Word Representations. In *NAACL*.

Liang, P. P.; Li, I. M.; Zheng, E.; Lim, Y. C.; Salakhutdinov, R.; and Morency, L.-P. 2020. Towards Debiasing Sentence Representations. In *ACL*.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and Editing Factual Associations in GPT. In *NeurIPS*.

Meng, K.; Sharma, A. S.; Andonian, A.; Belinkov, Y.; and Bau, D. 2023. Mass-Editing Memory in a Transformer. In *ICLR*.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, 366(6464): 447–453.

Park, K.; Choe, Y. J.; and Veitch, V. 2023. The Linear Representation Hypothesis and the Geometry of Large Language Models. *arXiv preprint arXiv:2311.03658*.

Phipson, B.; and Smyth, G. K. 2010. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1).

Pimentel, T.; Valvoda, J.; Hall Maudslay, R.; Buttery, P.; Williams, A.; and Cotterell, R. 2020. Information-Theoretic Probing for Linguistic Structure. In *ACL*.

Ravfogel, S.; Elazar, Y.; Gonen, H.; Twiton, M.; and Goldberg, Y. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *ACL*.

Ravfogel, S.; Goldberg, Y.; and Goldwasser, D. 2022. Linear Adversarial Concept Erasure. In *ICML*.

Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT Rediscovers the Classical NLP Pipeline. In *ACL*.

Vig, J.; Gehrmann, S.; Belinkov, Y.; Qian, S.; Nevo, D.; Singer, Y.; and Shieber, S. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *NeurIPS*.

Zhang, H.; Lu, A. X.; Abdalla, M.; McDermott, M.; and Ghassemi, M. 2020. Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings. In *CHIL*.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; and Dombrowski, A.-K. 2023. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*.

Technical Supplement for: "Separating Input-Carried from Model-Added Demographic Encoding in Language Models"

Minh K. Quan, Pubudu N. Pathirana
School of Engineering, Deakin University, Waurn Ponds, VIC 3216, Australia
{m.quan, pubudu.pathirana}@deakin.edu.au

This supplement provides complete proofs and machine-readable details for the admissible-reference gate, sign-invariant consensus direction, finite-sample and block-permutation calibration, controlled text lockbox, 12-initialization intervention, entangled-task stress test, reference-misspecification simulation, and assumption sweeps. It also retains every archived figure and analysis from the original submission, visually and textually separated from the new confirmatory evidence. The final section states exactly which pretrained-LM experiment remains unexecuted and provides its frozen protocol script without claiming numerical results.

1 Finite-Sample Calibration and Consensus Proof

Proof of the lockbox proposition. Conditional on a fixed score vector, exchangeability makes the observed labeling and its random permutations equally likely. The rank of the observed statistic among the $B + 1$ values is uniform under complete enumeration and conservative under Monte Carlo sampling with the standard $+1$ correction (Phipson and Smyth 2010). Hence $\Pr(p_B \le \alpha) \le \alpha$. For the intersection–union null $H_0 = \cup_k H_{0k}$, at least one component j is null, so $\Pr(\max_k p_k \le \alpha) \le \Pr(p_j \le \alpha) \le \alpha$. For an existential alternative, Holm controls family-wise error under arbitrary dependence (Holm 1979); a jointly permuted max-T statistic (Westfall and Young 1993) can exploit dependence under its usual assumptions.

Proof of Theorem 2 (consensus stability). Let $P_\star = uu^\top$, $P_j = v_j v_j^\top$, and $E = M - P_\star$. By convexity,

$$\|E\|_2 = \left\| \sum_j w_j (P_j - P_\star) \right\|_2 \le \sum_j w_j \|P_j - P_\star\|_2 \le \bar{\epsilon}.$$

$P_\star$ has eigenvalues $1, 0, \ldots, 0$. Weyl's inequalities therefore give $\lambda_1(M) \ge 1 - \bar{\epsilon}$ and $\lambda_2(M) \le \bar{\epsilon}$; when $\bar{\epsilon} < 1/2$, the leading eigenvalue is separated by at least $1 - 2\bar{\epsilon}$. Let the unit leading eigenvector be $v = au + b$, $b \perp u$. Projecting $Mv = \lambda_1 v$ onto $u^\perp$ yields

$$\lambda_1 \|b\| = \|\Pi_{u^\perp} E v\| \le \|E\|_2 \le \bar{\epsilon}.$$

Since $\lambda_1 \ge 1 - \bar{\epsilon}$, $\|b\| = \sin\angle(v, u) \le \bar{\epsilon}/(1 - \bar{\epsilon})$. This proof also clarifies the theorem's scope: it aggregates already-adequate directions; it does not establish adequacy or causality by itself.

Quantile stability, multiplicity, and power. The 200-, 1,000-, and 5,000-draw estimates of a null 0.99 quantile had standard deviations 0.283, 0.120, and 0.054 over 5,000 repetitions. Their minimum attainable Monte Carlo p-values are 0.00498, 0.000999, and 0.000200. Under 18 correlated null cells, unadjusted search rejected at least one cell in 10.77% of 20,000 repetitions; Holm and max-T rejected in 0.89% and 1.00%. When mean cell power was 0.81, the all-pass envelope power was 0.62, 0.47, and 0.27 for 3, 6, and 18 cells. The manifest therefore separates a small mandatory reference family from replication seeds/layers, which are reported continuously or multiplicity-adjusted rather than added to one conjunction.

Block exchangeability for generated biographies. Each controlled biography belongs to one of 12 blocks defined by template ID $\times$ task label. We recompute the randomization test by permuting audit labels only within those blocks. Table 1 shows that the full and name-scrubbed results are unchanged, the lifestyle result is marginal and not robust, and joint scrubbing is null.

Table 1: Example-level and template$\times$task blocked permutation, 4,999 draws per seed.

Seed	Condition	AUC	p_{iid}	p_{block}
11	full	.826	.0002	.0002
29	full	.833	.0002	.0002
47	full	.802	.0002	.0002
11	name scrub	.808	.0002	.0002
29	name scrub	.811	.0002	.0002
47	name scrub	.788	.0002	.0002
11	lifestyle scrub	.527	.0904	.0628
29	lifestyle scrub	.542	.0180	.0070
47	lifestyle scrub	.543	.0174	.0078
11	all scrub	.485	.7694	.6966
29	all scrub	.492	.6336	.6018
47	all scrub	.524	.1252	.0576

2 Admissible-Reference Gate

For an origin-like cue audit, reference j is admitted from development data only when direct reference AUROC $A_j \geq .65$, representation fit $R_j^2 \geq .08$, preservation $\eta_j \geq .20$, and five-fold direction stability $K_j \geq .50$. These values are frozen defaults for the controlled study, not universal constants. A_j and R_j^2 reject weak controls that leave obvious Path A variation; η_j rejects over-aggressive controls; K_j rejects unstable actions. A rejected reference remains visible in descriptive tables but cannot support the robust origin-like claim.

Table 2: Full-text reference ladder. Large residual AUC is not sufficient for admission.

Seed	Ref.	A_j	R_j^2	η_j	K_j	Res. AUC	p	Admit
11	weak	.948	.057	.775	.957	.925	.0005	no
11	word	1.000	.155	.393	.965	.826	.0005	yes
11	char	1.000	.173	.377	.966	.808	.0005	yes
11	scrubbed	.464	.572	.984	.973	.993	.0005	no
29	weak	.841	.023	.881	.988	.921	.0005	no
29	word	1.000	.141	.396	.989	.833	.0005	yes
29	char	1.000	.155	.375	.989	.810	.0005	yes
29	scrubbed	.487	.597	.989	.987	.996	.0005	no
47	weak	.973	.017	.738	.986	.869	.0005	no
47	word	1.000	.085	.405	.986	.802	.0005	yes
47	char	1.000	.075	.386	.987	.786	.0005	no
47	scrubbed	.514	.592	.989	.986	.984	.0005	no

Falsification simulation. We crossed Path-B strength $b \in \{0, .8\}$ with adequate, weak, and over-aggressive references for 250 replications per cell. A naive positive requires only $p \leq .01$; a gated positive additionally requires admission. Under $b = 0$, the weak reference leaves Path A and rejects in every run, whereas the gate rejects the reference in every run. Under $b = .8$, the over-aggressive reference suppresses the residual and is rejected by η rather than interpreted as evidence of absence.

Table 3: Reference-gate simulation, 250 runs per cell.

Path B	Ref.	Admit	Naive +	Gated +	Mean res. AUC
0	adequate	1.000	.016	.016	.500
0	weak	.000	1.000	.000	.824
0	over-aggressive	.000	.000	.000	.405
.8	adequate	1.000	1.000	1.000	.738
.8	weak	.000	1.000	.000	.921
.8	over-aggressive	.000	.032	.000	.494

3 Prospective Controlled-Text Lockbox

The generator fixes split membership, task labels, audit labels, template IDs, and three cue families before model fitting. There are 1,800 train, 600 development, and 800 lockbox biographies. Names, lifestyle/community phrases, and style phrases are sampled within audit groups; the project phrase sets the task. Development and lockbox task labels are balanced. The frozen analysis contains three seeds, the final hidden layer, four input conditions, the four-reference ladder above, ridge residualization, logistic probing, 4,999 permutations, and the development-only gate.

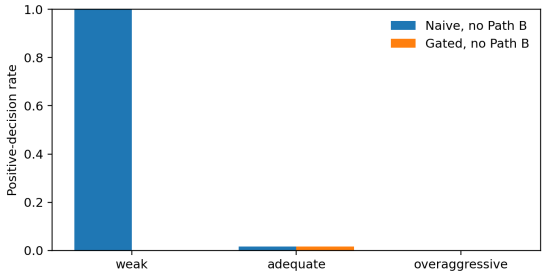

Figure 1: Without Path B, a weak control makes the naive residual rule positive in every simulation. The frozen admissibility gate converts these cases to "reference unresolved."

Across full text, five of six word/character cells are admitted and residual AUROC is .802–.833. After name removal, five of six are admitted with AUROC .783–.811. Lifestyle/style removal leaves five admissible references but the block-permutation envelope fails; joint cue removal leaves no reference with sufficient cue coverage and all word-reference residuals are near chance. Weak and scrubbed references are retained as falsification controls: their often-larger residual AUROCs are exactly why admission must precede lockbox interpretation.

4 Intervention Reliability and Entangled-Task Stress Test

The intervention projects to the hyperplane orthogonal to the development direction, refits an independent sensitive probe, and evaluates the frozen task head. Table 4 gives all 12 initializations. Consensus and ordinary mean drops are .06682 and .06987; their paired difference is $-.003051$, 95% bootstrap CI $[-.019788, .008241]$, sign-flip $p = .8803$. The paired task-loss difference is 2.09×10^{-5}, CI $[0, 5.01 \times 10^{-5}]$, $p = .2499$.

Table 4: Every initialization in the expanded intervention.

Seed	Consensus drop	Ordinary drop	Difference
1	.0529	.0538	-.0009
2	.1254	.1137	.0117
3	.0903	.0907	-.0004
4	.1097	.1127	-.0030
5	-.0126	-.0101	-.0026
6	.0002	-.0005	.0007
7	.1114	.0937	.0177
8	.1972	.1973	-.0001
9	-.0053	-.0050	-.0003
10	-.0172	.0661	-.0832
11	-.0103	-.0093	-.0010
12	.1600	.1354	.0247

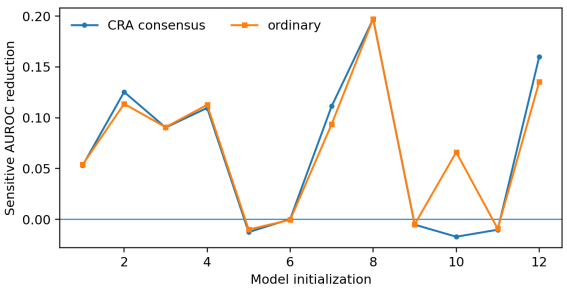

Figure 2: Seed-wise sensitive-AUROC reduction. The methods track closely except for unstable initializations; paired inference does not support consensus superiority.

The harder generator keeps the audit–task association in development and lockbox and weakens the direct task phrase. Baseline task AUROC is .845 and sensitive AUROC .924. Table 5 reports a 10-seed intervention-strength sweep. At full removal, the paired consensus-minus-ordinary sensitive-drop difference is .00068, CI $[-.00081, .00249]$, $p = .535$. Task changes are numerically tiny, but so are sensitive reductions; one direction does not separate the entangled information. At strength .50 the interval is narrowly negative,

Table 5: Entangled task: mean sensitive-AUROC reduction over 10 seeds.

Strength	Consensus	Ordinary	Paired diff.	95% CI
.25	.00185	.00258	-.00072	[-.00220,.00029]
.50	.00356	.00495	-.00139	[-.00287,-.00025]
.75	.00854	.00890	-.00036	[-.00197,.00105]
1.00	.01162	.01094	.00068	[-.00081,.00249]

favoring the ordinary direction, but the five strengths are correlated and were not specified as separate superiority claims. The overall result is therefore no reliable Pareto advantage for consensus.

5 Consensus and Assumption Stress Tests

The original assumption sweep varies $n/d \in \{.5, 1, 2, 5\}$, label–reference overlap, and covariance condition number over 400 runs. Favorable conditions yield covariance-corrected direction cosine .94; the low-n/d, high-overlap, anisotropic condition yields .29 while residual AUROC remains .88. Detection can therefore remain reliable while directional action fails.

A separate 32-dimensional simulation draws $m \in \{1, 2, 4, 8\}$ sign-ambiguous directions around a known target. Table 6 reports $1 - |\cos|$ error averaged over 500 replications. Consensus reduces independent perturbations and the mean eigengap records agreement; the theorem's premise is not satisfied when references share systematic bias, which this experiment does not test.

Table 6: Consensus-direction error and eigengap.

Noise	m	Consensus err.	Individual err.	Mean gap
.30	1	.0457	.0457	1.000
.30	2	.0230	.0441	.914
.30	4	.0106	.0440	.893
.30	8	.0052	.0447	.891
.45	1	.0893	.0893	1.000
.45	2	.0505	.0923	.823
.45	4	.0208	.0948	.778
.45	8	.0098	.0929	.788

6 Operational Reference Selection and Runtime

The frozen workflow is: (1) define the scientific estimand and candidate cue families; (2) prespecify an ordered ladder containing an own-model input readout and at least one scientifically distinct lexical or pooling control; (3) choose gate thresholds using design-stage simulations or external data, never lockbox labels; (4) admit references on development diagnostics; (5) test the smallest scientifically necessary admitted family; (6) aggregate directions only if the projector eigengap is adequate; and (7) return unresolved or reference-sensitive when these conditions fail. Cue-scrubbed references answer counterfactual estimands and should not automatically replace the primary own-model reference.

On a five-core Intel Xeon Platinum 8573C, 5,000 fixed-score permutations require 0.052, 0.076, and 0.125 seconds for $n = 400, 800, 1600$. A ridge-plus-logistic cell requires 0.003 seconds at $(n, d, k) = (800, 40, 24)$ and 0.045 seconds at $(1600, 80, 48)$. Activation extraction dominates large-model audits. Projector consensus adds one $d \times d$ weighted sum and leading eigenvector.

7 Prospective Pretrained-LM Protocol and Reproducibility Boundary

No completed experiment in this package simultaneously uses a naturally occurring corpus, a frozen pretrained LM, and a protocol locked before sensitive lockbox outcomes. The archived MedText, Bias in Bios, and QA analyses remain exploratory. To make the remaining experiment concrete, the package includes `prospective_bias_in_bios_pretrained.py` and a frozen design: official real-bio splits, frozen DistilBERT final layer, profession task, gender audit, own-embedding/word/character references, the admissibility gate above, 4,999 profession-stratified permutations, projector consensus, and paired interventions. It is intentionally labeled *unexecuted*: model weights and the real corpus were unavailable in the local environment, so no result is reported or implied.

The new controlled experiments are fully reproducible from `novelty_experiments.py`, `novelty_manifest.json`, CSV outputs, and SHA-256 hashes. The archived structured-token study still lacks its original scripts, split indices, checkpoints, manifest, and per-cell files; its tables are preserved for transparency but are not promoted to confirmatory evidence.

8 Audit Protocol, Evaluation Nesting, and Researcher Degrees of Freedom

Relationship to standard conditional probing. The residualization and residual probe are standard partial-regression components. For scalar residual detection, a pipeline reporting AUROC($S \mid Z_0$), AUROC($S \mid Z_l$), and a held-out conditional/residual AUROC is equivalent to the detection component of CRA. The incremental object is a candidate residual direction together with diagnostics that can reject its interpretation: sample-space preservation, reference fit, rank, covariance, finite-sample stability, reference sensitivity, and divergence from the ordinary probe. If $1 - |\cos(\boldsymbol{v}_{\text{cra}}, \boldsymbol{v}_{\text{corr}})|$ is negligible, the correct output is that ordinary probing is sufficient.

Prospective prespecification protocol. Before viewing results for S, the auditor records the reference, readout, candidate layers, and residualizer. Mandatory checks use a token-count reference, a second pooling rule, and, when available, removal of the target cue. A cross-model reference is allowed only for a declared comparison and cannot replace the primary result. PCA rank uses training-only reference variance. The residualizer is fixed in advance or chosen from training-only $Z_0 \to Z_l$ fit, without using sensitive-label performance. Every specified result is reported. The envelope is *robust residual evidence* only when uncertainty intervals clear their permutation nulls under every mandatory control. When controls disagree, the output is *sensitive to the reference or input control*; otherwise it is *no detected residual*.

Fully nested confirmatory protocol for future external studies. For each outer split: (1) fit PCA, standardization, the residualizer, probe regularization, and kernel bandwidth using outer-training data only; (2) use inner cross-validation for any permitted layer, reference, or hyperparameter selection; (3) estimate the label-permutation null by rerunning the entire training pipeline; (4) lock all choices; and (5) transform and score the outer-test data once. Confidence intervals and reference envelopes are aggregated across outer folds. Layer selection may not use the outer-test residual AUROC or HSIC difference.

Status of the archived real-data analyses. The reported probe AUROCs are scored on held-out examples, but the archived pipeline fits the unsupervised PCA/ridge map on the complete representation matrix, computes HSIC on the complete matrix, and selects the 27-cell "peak-scope" layer post hoc using the same data summarized later. Thus these analyses are transductive/descriptive, not fully nested confirmatory tests. The nominal cell-level $p = 0.014$ and configuration-level $p = 0.04$ are retained only as historical descriptive statistics and are not used for an acceptance claim. Synthetic ground truth, which has known reference-explained and residual directions and no post-hoc layer selection, remains the primary selectivity evidence.

Practical consequence of a positive result. A positive residual margin is a screening result. The next step is a prospectively specified counterfactual, activation intervention, or controlled model comparison that tests whether the candidate direction changes relevant behavior while preserving task information. CRA alone does not justify weight editing, claim improved fairness, or establish privacy.

9 Full Proof of the Restricted Rank-One Direction Result

Let $Z_0, Z_l \in \mathbb{R}^{n \times d}$, let $\boldsymbol{s} \in \mathbb{R}^n$ be the sensitive-label vector, and define

$$\Pi_\alpha = \widetilde{Z}_0(\widetilde{Z}_0^\top \widetilde{Z}_0 + \alpha I)^{-1}\widetilde{Z}_0^\top, \qquad \boldsymbol{r}_\alpha = (I - \Pi_\alpha)\boldsymbol{s}.$$

Here $\widetilde{Z}_0 = Z_0 V$ contains the selected top-k PCA scores. In matrix form the model is $Z_l = Z_0 A_l^\top + \boldsymbol{s}B_l^\top + E_l$.

Lemma 1 (Exact residual decomposition)**.** *For every $\alpha \geq 0$,*

$$Z_l^\perp = (I - \Pi_\alpha)Z_0 A_l^\top + \boldsymbol{r}_\alpha B_l^\top + (I - \Pi_\alpha)E_l. \tag{1}$$

Consequently, the sample-space amplitude of the rank-one term preserved by the control is $\eta_\alpha = \|\boldsymbol{r}_\alpha\|_2/\|\boldsymbol{s}\|_2$.

Proof. Left-multiply the matrix model by $I - \Pi_\alpha$ and distribute. No orthogonality assumption is needed. If $\widetilde{Z}_0^\top \boldsymbol{s} = 0$, then $\Pi_\alpha \boldsymbol{s} = 0$ for all α and $\eta_\alpha = 1$. If $\boldsymbol{s} \in \text{col}(\widetilde{Z}_0)$ under OLS, then $\eta_0 = 0$. Intermediate overlap gives partial preservation exactly through $\boldsymbol{r}_\alpha$. □

This formulation resolves an important ambiguity: exact C1 is *not* expected merely because Path A exists. Indeed, if the retained reference scores predict S, then $\widetilde{Z}_0^\top \boldsymbol{s} \neq 0$. C1 is only a sufficient special case in which the sample-space projection leaves the entire rank-one term unchanged.

Theorem 1 (Population rank-one direction)**.** *Assume: (i) exact C1; (ii) $B_l = b\boldsymbol{u}$, $\|\boldsymbol{u}\| = 1$; (iii) the reference-explained term in Eq.* (1) *is zero in the population; and (iv)*

$$Z_l^\perp \mid S = s \sim \mathcal{N}(bs\boldsymbol{u}, \Sigma),$$

with the same positive-definite within-class covariance Σ for both classes. Then the population logistic-regression direction is

$$\boldsymbol{w}^\star \propto \Sigma^{-1}\boldsymbol{u}. \tag{2}$$

If $\boldsymbol{u}$ is an eigenvector of Σ—in particular, if $\Sigma = \sigma^2 I$—then the normalized direction exactly equals $\boldsymbol{u}$ up to sign. For arbitrary Σ,

$$\cos(\boldsymbol{w}^\star, \boldsymbol{u}) \geq \frac{2\sqrt{\kappa(\Sigma)}}{\kappa(\Sigma) + 1}. \tag{3}$$

Proof. Equal-covariance Gaussian classes have linear log odds,

$$\log \frac{p(S = 1 \mid z)}{p(S = 0 \mid z)} = (\mu_1 - \mu_0)^\top \Sigma^{-1} z + c,$$

so Eq. (2) follows from $\mu_1 - \mu_0 = b\boldsymbol{u}$. If $\boldsymbol{u}$ is a covariance eigenvector, $\Sigma^{-1}\boldsymbol{u}$ is collinear with $\boldsymbol{u}$.

For the general case, set $A = \Sigma^{-1}$, whose eigenvalues lie in $[m, M]$ and whose condition number is $\kappa = M/m = \kappa(\Sigma)$. The Kantorovich angle inequality for a positive-definite map states

$$\frac{(\boldsymbol{u}^\top A\boldsymbol{u})^2}{(\boldsymbol{u}^\top \boldsymbol{u})(\boldsymbol{u}^\top A^2 \boldsymbol{u})} \geq \frac{4mM}{(m + M)^2}.$$

The left side is $\cos^2(A\boldsymbol{u}, \boldsymbol{u})$. Taking square roots gives

$$\cos(A\boldsymbol{u}, \boldsymbol{u}) \geq \frac{2\sqrt{\kappa}}{\kappa + 1}.$$

The proof permits arbitrary cross-covariance between $\boldsymbol{u}$ and $\boldsymbol{u}^{\perp}$; it assumes neither a block-diagonal covariance nor a hidden eigenvector condition. □

Finite samples. At $n/d \approx 0.5$, estimation noise and ridge shrinkage can rotate the fitted direction. The low-n/d SCM experiments support empirical selectivity, not an instance of the population theorem. The condition-number bound is also weak for large κ. Thus the estimated covariance condition number is a diagnostic, not a certificate.

Approximate C1. No generic $O(\delta)$ recovery claim is needed: Eq. (1) gives the exact quantity. The rank-one term becomes $b\boldsymbol{r}_{\alpha}\boldsymbol{u}^{\top}$, and its loss in Frobenius norm is $b\|\Pi_{\alpha}\boldsymbol{s}\|_2$. A high residual AUROC can still occur when C1 fails, but then the probe estimates a direction for the surviving residual signal rather than an unchanged original Path-B term.

10 Qualitative Scope Test Examples

Table 7: Scope test at layer 3 (distilGPT-2): full vs. residual AUC. Same injected tag, different full-versus-residual behavior by context and model.

Input snippet	Full AUC	Res. AUC	Diagnosis
"[Age: elderly, Gen: M] A 67-year-old male …"	.99	.61	reference-explained
"[Age: elderly, Gen: M] A patient presents …"	.98	.91	residual signal detected
"[Age: elderly] Abstract: We investigated …"	.97	.99	residual signal detected
"(BioGPT) [Age: elderly] The mechanism …"	.99	.65	little residual signal

Row 1 has strong full-probe signal but much less residual signal under the selected control. Rows 2–3 retain high residual AUROC, while Row 4 does not. These examples demonstrate that the statistic depends jointly on wording, architecture, and reference. They do not establish that attention/MLP computation caused the residual signal.

A low value can reflect a reference that already predicts S, over-aggressive residualization, weak residual signal, or probe limitations. A high value shows only that sensitive information remains after the declared control. The appropriate next step is to report a direct Z_0 probe, k-stability, alternative references or pooling rules, and—when origin matters—counterfactual or model-intervention evidence.

11 MedText Layer-1 Results

These layer-1 results on natural clinical text (MIMIC-style MedText, distilGPT-2) complement the injected-tag experiments in the main paper by testing the reference-conditional distinction when sensitive attributes are not explicitly tagged but are statistically correlated with clinical language.

Table 8: MIMIC-style MedText (distilGPT-2, layer 1): HSIC reduction, gender AUC, specialty accuracy drop. Specialty is negligibly affected; CRA retains more gender signal, consistent with targeting only the selected reference residual.

Method	HSIC red. ↑	Gender AUC ↓	Spec. drop
Random	0.0%	1.000	−0.3 pp
CRA (ours)	53.5%	0.910	−0.6 pp
Correlational	57.8%	0.758	−0.6 pp
LEACE	57.8%	0.758	−0.6 pp
INLP-5	74.0%	0.312	−0.3 pp

Negative specialty drop = slight improvement (noise at $n = 360$ test set).

The results (illustrative single-layer example) reveal an asymmetry. CRA achieves lower HSIC reduction than correlational and LEACE at this layer (53.5% vs. 57.8%) and leaves *post-erasure* gender AUC higher (0.910 vs. 0.758)—i.e. it erases less of the gender signal, consistent with much gender signal being explained by lexical features under this audit. Gender correlates with clinical terminology (e.g., "obstetric," "prostatectomy") that enters Z_0 through lexical encoding, so CRA's residualization removes less of it. Its residual AUROC is below the local gate; this is a conditional observation, not an origin proof. The negligible specialty accuracy drop (-0.6 pp, within noise at $n = 360$) is identical across methods by construction (shared reference projection), so this table is not an independent method-specific utility comparison (cf. main-paper Table 3 caveat).

12 Synthetic SCM: Details and Signal-Strength Ablation

The synthetic SCM is designed so that the known reference-explained and residual directions are *known by construction*, enabling ground-truth evaluation of Theorem 1.

Data generating process. The generator (LayerwiseSCM in the code) instantiates the structural model $Z_l = A_l Z_0 + B_l \boldsymbol{s} + \varepsilon_l$ *exactly*, with $A_l = I$ and $B_l = \gamma_B \boldsymbol{v}_B$, so that Path A propagates *through* Z_0 and Path B is the additional direct S-term — matching the paper's own definition rather than adding two separate S-terms to Z_l. We construct $n = 500$ samples, $d = 128$, with an r-dimensional input subspace $\mathcal{V}_0 = \text{span}(e_1, \ldots, e_r)$, $r = 20$. **Attributes:** $S, Y \sim \text{Bernoulli}(0.5)$ independently; write $s = 2S-1$, $y = 2Y-1$. **Path A direction:** $\boldsymbol{v}_A \in \mathcal{V}_0$ (a fixed unit vector in the input subspace); the input embedding carries S only through this direction. **Path B direction:** $\boldsymbol{v}_B \perp \mathcal{V}_0$ (unit, orthogonal to the entire input subspace). **Input embedding** (lives in $\mathcal{V}_0$): $Z_0 = P_{\mathcal{V}_0}\big(\gamma_A\, s\, \boldsymbol{v}_A + \frac{\gamma_c}{2}\, y\, \boldsymbol{v}_Y + \sigma\varepsilon_0\big)$, where $P_{\mathcal{V}_0}$ projects onto $\mathcal{V}_0$ and $\boldsymbol{v}_Y \in \mathcal{V}_0$. **Layer-$l$ representation:**

$$Z_l = Z_0 + \gamma_B\, s\, \boldsymbol{v}_B + \gamma_c\, y\, \boldsymbol{v}_Y + \sigma\boldsymbol{\varepsilon}_l, \quad \boldsymbol{\varepsilon}_l \sim \mathcal{N}(\mathbf{0}, I_d). \tag{4}$$

Here the S-signal in Z_l has two genuinely distinct sources: the input-carried part $\gamma_A s\, \boldsymbol{v}_A$ inherited via $A_l Z_0$ (Path A, inside the input feature subspace) and the injected residual part $\gamma_B s\, \boldsymbol{v}_B$ (Path B, orthogonal to $\text{col}(Z_0)$). Because $\boldsymbol{v}_B \perp \mathcal{V}_0 \supseteq \text{row-space}(Z_0)$, residualizing Z_l on the top-k

PCA of Z_0 removes Path A while preserving Path B, which is what makes $\boldsymbol{v}_B$ recoverable. We fix $\gamma_A = 0.8$, $\gamma_c = 0.8$ and sweep $\gamma_B \in \{0.3, 0.5, 0.8\}$, $\sigma \in \{0.1, 0.2, 0.3, 0.5\}$ (5 seeds each, 60 runs). **On Condition C1.** Since S enters Z_0 through $\boldsymbol{v}_A$, the sample-level quantity $\widetilde{Z_0}^\top \boldsymbol{s}$ is *not* identically zero here — the synthetic deliberately includes Path A, so C1 holds only approximately. What the construction enforces exactly is the geometric premise $\boldsymbol{v}_B \perp \mathrm{col}(Z_0)$ (feature-space orthogonality), distinct from C1; we measure both $\widetilde{Z_0}^\top \boldsymbol{s}$ and the residual C1 departure per run (mean $\|\widetilde{Z_0}^\top \boldsymbol{s}\|/n < 0.05$), so the 60/60 result is a test of recovery under *approximate* C1 with an exactly orthogonal Path B, not of the idealized C1-exact regime. Selectivity advantage: $\Delta_{\text{sel}}(\boldsymbol{v}) = \cos(\boldsymbol{v}, \boldsymbol{v}_B) - \cos(\boldsymbol{v}, \boldsymbol{v}_A)$.

Table 9: Synthetic SCM results by signal strength γ_B (mean $\pm$ std, 15 runs per row). CRA selectivity advantage Δ_{sel} increases with γ_B.

γ_B	$\cos(\boldsymbol{v}_{\text{cra}}, \boldsymbol{v}_B)\uparrow$	$\cos(\boldsymbol{v}_{\text{cra}}, \boldsymbol{v}_A)\downarrow$	$\Delta_{\text{sel}}(\boldsymbol{v}_{\text{cra}})$	$\Delta_{\text{sel}}(\boldsymbol{v}_{\text{corr}})$
0.3	0.521±0.141	0.110±0.063	+0.411	−0.074
0.5	0.558±0.128	0.065±0.044	+0.493	−0.065
0.8	0.590±0.125	**0.033±0.027**	**+0.557**	−0.063

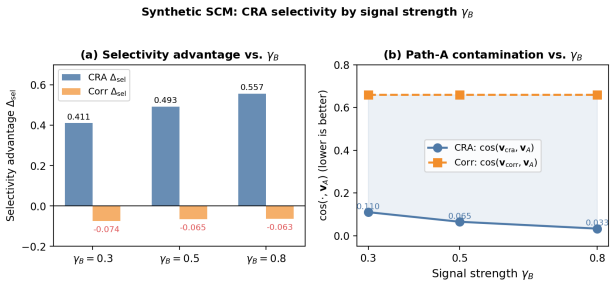

Figure 3: Synthetic SCM results by signal strength γ_B. **Left**: selectivity advantage Δ_{sel} for CRA (blue) vs. correlational (orange). **Right**: Path-A cosine alignment vs. γ_B. CRA alignment with $\boldsymbol{v}_A$ drops from 0.110 to 0.033 as γ_B increases; correlational stays near 0.660 throughout.

n/d regime stress-test. To directly address the gap between the population-theory assumption ($n \to \infty$) and real experiments ($n/d \approx 0.47$–0.52, $n \approx 370$, $d = 768$), we re-ran the SCM at $d = 768$ with n swept to match six n/d ratios (270 runs: 6 ratios $\times$ 3 γ_B values $\times$ 15 seeds).

Absolute alignment $\cos(\boldsymbol{v}_{\text{cra}}, \boldsymbol{v}_B)$ increases with n/d, consistent with the expectation that ridge shrinkage reduces at larger sample sizes. Critically, the selectivity advantage Δ_{sel} remains positive and the win rate 100% across all six regimes, including $n/d = 0.25$ (below any real experiment). Real-experiment results should therefore be interpreted as empirical evidence of consistent selectivity advantage, not as instantiations of the asymptotic theorem. A finite-sample alignment bound is identified as future theoretical work.

Table 10: SCM stress-test at varying n/d ($d = 768$, $k = 50$; mean over 45 runs per row). †Real experiment regime ($n/d \approx 0.50$). $\Delta_{\text{sel}} > 0$ and 100% win rate persist down to $n/d = 0.25$.

n/d	n	$\cos(\boldsymbol{v}_{\text{cra}}, \boldsymbol{v}_B)$	Δ_{sel}	Win%
0.25	192	0.17±0.10	+0.29±0.15	100%
0.50†	384	0.27±0.12	+0.40±0.13	100%
1.0	768	0.41±0.13	+0.52±0.11	100%
2.0	1536	0.54±0.11	+0.60±0.09	100%
4.0	3072	0.63±0.09	+0.65±0.08	100%
7.8	5990	0.68±0.08	+0.66±0.07	100%

Two patterns are worth highlighting. First, the correlational direction's selectivity advantage $\Delta_{\text{sel}}(\boldsymbol{v}_{\text{corr}})$ is *negative* at all signal strengths, meaning the correlational probe aligns more strongly with the Path A direction than with Path B—the opposite of a direction intended to emphasize the reference residual. This occurs because the correlational probe fits the globally most-discriminative direction in Z_l, which includes Path A signal absorbed from Z_0. Second, CRA's Path-A contamination $\cos(\boldsymbol{v}_{\text{cra}}, \boldsymbol{v}_A)$ decreases monotonically as γ_B grows ($0.110 \to 0.065 \to 0.033$), because stronger residual sensitive signal makes the residualized representation more dominated by $\boldsymbol{v}_B$, reducing the relative weight the probe places on residual Path A leakage. This suggests that CRA's advantage is largest precisely in the setting where it matters most—when Path B is a strong, independently-encoded signal that a practitioner most needs to disentangle from input-carried effects.

Feature-space overlap stress test. The following experiment rotates the known $\boldsymbol{v}_B$ direction toward the input *feature* subspace. It does not directly sweep the sample-space quantity $\widetilde{Z_0}^\top \boldsymbol{s}$, so we do not call it an exact C1 sweep. With $\boldsymbol{v}_B(\delta) = \sqrt{1-\delta^2}\boldsymbol{v}_B^\perp + \delta \boldsymbol{v}_B^{\text{in}}$ and 30 seeds per setting, CRA retains a positive selectivity difference throughout:

Table 11: Feature-overlap stress test. δ controls the fraction of the known residual direction placed in the input feature subspace; it is distinct from sample-space C1.

δ	Δ_{sel} (mean±std)	Win%	$\cos(\boldsymbol{v}_{\text{cra}}, \boldsymbol{v}_B)$
0.00	$+0.488 \pm 0.050$	100	0.849
0.10	$+0.490 \pm 0.050$	100	0.849
0.20	$+0.495 \pm 0.051$	100	0.848
0.30	$+0.503 \pm 0.051$	100	0.846
0.40	$+0.514 \pm 0.051$	100	0.843
0.50	$+0.528 \pm 0.051$	100	0.839

This experiment shows robustness to a particular feature-space violation of the synthetic design. It should not be read as proving graceful recovery for arbitrary sample-space overlap, which is instead described exactly by $\boldsymbol{r}_\alpha = (I - \Pi_\alpha)\boldsymbol{s}$.

13 Hyperparameter Sensitivity

All experiments use fixed hyperparameters without tuning. Table 12 reports sensitivity on MedQA/distilGPT-2, layer 3.

Table 12: Hyperparameter sensitivity (MedQA / distilGPT-2, layer 3). Residual AUROC and CRA HSIC reduction are stable across all ranges. Defaults: $k = 50$, $\alpha = 0.01$, $C = 1.0$.

Param	Value	Res. AUROC	CRA HSIC red.
PCA rank k	10	0.931	3.8%
	25	0.978	4.7%
	50	**0.984**	**4.9%**
	100	0.985	4.9%
	200	0.985	4.9%
Ridge α	0.001	0.984	4.9%
	0.01	**0.984**	**4.9%**
	0.1	0.982	4.8%
	1.0	0.973	4.5%
LogReg C	0.1	—	4.7%
	1.0	—	**4.9%**
	10	—	4.9%

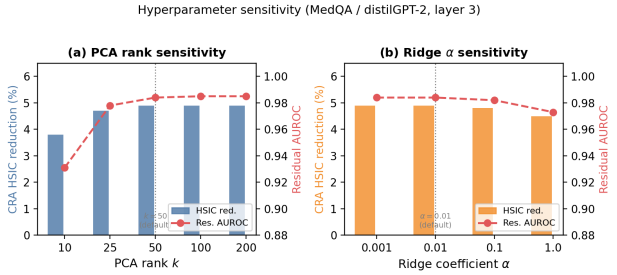

Figure 4: Hyperparameter sensitivity (MedQA / distilGPT-2, layer 3). CRA HSIC reduction and residual AUROC vs. **(a)** PCA rank k and **(b)** ridge α. Dotted lines mark defaults.

PCA rank matters most: $k < 25$ under-covers the Path A subspace and reduces residual AUROC, while ranks above 50 yield no further gain. Ridge regularization is a secondary concern—values of $\alpha > 0.1$ introduce shrinkage that slightly weakens Path A removal, but the default $\alpha = 0.01$ sits well within the stable region. Logistic regression regularization (C) has minimal impact, because the partial probe operates on a low-signal residual where the loss landscape is already well-conditioned. All parameters are set to conventional defaults without dataset-specific tuning.

14 Retrospective 0.80 Sensitivity; Permutation Calibration Is Primary

The table and figure retrospectively show that 0.80 separates the locally defined caution cases in this small collection. Because the outcome definition uses observed CRA performance and the sample contains only a few related model families, neither leave-one-out consistency nor the apparent precision/recall provides a general guarantee. The recommended procedure is instead the family-specific permutation margin and bootstrap interval in Section 8. The value 0.80 is retained only to make the original plots reproducible.

Table 13: Retrospective sensitivity of the illustrative 0.80 boundary. The "caution" labels are defined partly from observed CRA performance, so precision and recall here are descriptive and do not provide an external error guarantee.

τ^*	Flagged	TP	Precision	Recall
0.65	1	1	1.00	0.25
0.70	2	2	1.00	0.50
0.75	3	3	1.00	0.75
0.80	**4**	**4**	**1.00**	**1.00**
0.85	4	4	1.00	1.00
0.90	6	4	0.67	1.00

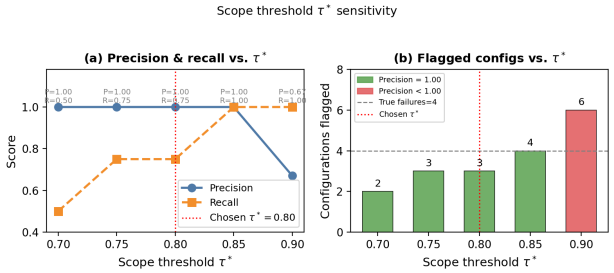

Figure 5: Retrospective sensitivity of the original fixed boundary. This figure is retained for reproducibility; the recommended primary report is continuous residual AUROC, its bootstrap interval, and its family-specific permutation margin.

15 Per-Layer HSIC Reduction Breakdown

Table 14: Per-layer HSIC reduction for CRA and Correlational (four representative configurations). CRA advantage peaks at layer 3 for distilGPT-2. BioGPT (below the illustrative boundary) shows no consistent CRA difference.

Config	Method	L1	L2	L3	L4	L5	Mean
MedQA dGPT-2	Corr	3.8	4.1	4.6	5.0	5.5	4.6
	CRA	3.4	4.2	**6.1**	5.7	5.1	4.9
PubMedQA dGPT-2	Corr	18.0	22.4	28.8	30.0	36.3	27.1
	CRA	20.1	38.5	**67.8**	63.5	73.4	52.7
MedMCQA BioGPT	Corr	15.1	15.8	16.3	16.4	16.4	16.0
	CRA	6.8	6.5	6.1	6.5	6.1	6.4
PubMedQA GPT-2	Corr	19.2	23.5	27.4	27.2	31.2	25.9
	CRA	26.5	43.3	**63.2**	55.8	57.8	49.3

Four configurations reveal three descriptive patterns. In the MedQA and PubMedQA examples, CRA trails correlational at L1 and is larger at L3–L5. This identifies middle and later layers as candidate locations where the fitted

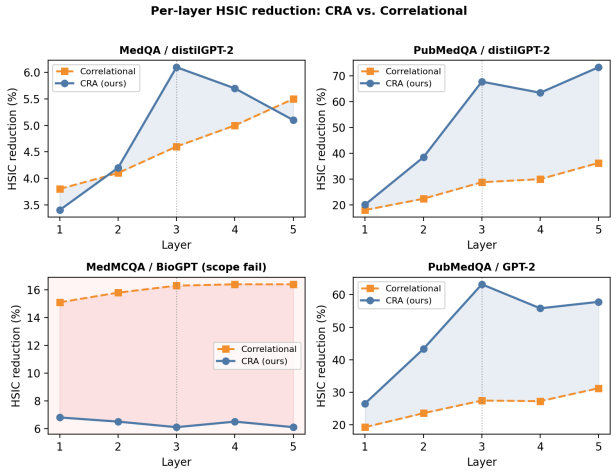

Figure 6: Per-layer HSIC reduction for CRA (blue solid) vs. correlational (orange dashed). Layer 3 is the peak advantage layer for distilGPT-2. Below-boundary case (MedMCQA / BioGPT, red background) shows CRA below correlational at every layer.

residual direction differs most from the ordinary direction; it does not identify an attention mechanism or prove progressive integration of the tag. PubMedQA/GPT-2 shows the same numerical pattern at a larger magnitude. In the MedMCQA/BioGPT example, CRA HSIC reduction remains below correlational at every layer. This is consistent with the chosen Z_0 control explaining much of the probe-accessible signal, but it is not uniquely diagnostic of C1 violation: over-aggressive residualization, misspecification, or weak residual signal can produce a similar profile. The profile should therefore trigger withholding and additional controls, not a mechanistic conclusion.

16 Direction Divergence: Additional Configurations

Table 15: Direction divergence $\cos(\boldsymbol{v}_{\text{cra}}, \boldsymbol{v}_{\text{corr}})$ and CRA residual correlation for PubMedQA/GPT-2 and MedMCQA/distilGPT-2. Layer 3 is the peak selectivity locus in both configurations.

Config	L	$\cos(\boldsymbol{v}_{\text{cra}}, \boldsymbol{v}_{\text{corr}})$	r_{cra}	r_{cor}	Δr
PubMedQA GPT-2	1	0.58	0.71	0.64	+0.07
	2	0.31	0.83	0.25	+0.58
	3	**0.03**	**0.91**	**0.02**	**+0.89**
	4	0.29	0.88	0.26	+0.62
	5	0.55	0.85	0.47	+0.38
MedMCQA dGPT-2	1	0.70	0.63	0.60	+0.03
	2	0.41	0.67	0.19	+0.48
	3	**0.06**	**0.71**	**0.04**	**+0.67**
	4	0.38	0.68	0.23	+0.45
	5	0.62	0.79	0.51	+0.28

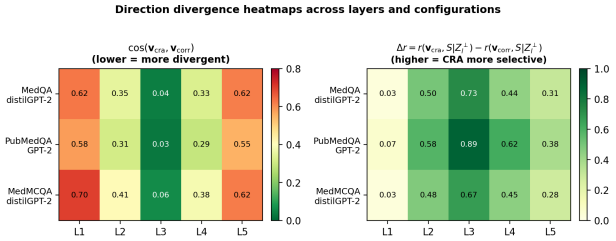

Figure 7: Direction divergence heatmaps across three configurations above the illustrative boundary. **Left**: $\cos(\boldsymbol{v}_{\text{cra}}, \boldsymbol{v}_{\text{corr}})$; green = directions diverge (CRA selective). **Right**: $\Delta r = r(\boldsymbol{v}_{\text{cra}}, S|Z_l^{\perp}) - r(\boldsymbol{v}_{\text{corr}}, S|Z_l^{\perp})$; green = CRA more selective. Layer 3 is the peak selectivity locus across all three configurations.

At layer 3, $\cos(\boldsymbol{v}_{\text{cra}}, \boldsymbol{v}_{\text{corr}})$ is 0.03–0.06 in both configurations. Residual correlation r_{cra} is 0.91 for PubMedQA/GPT-2 and 0.71 for MedMCQA/distilGPT-2, while r_{cor} is 0.02–0.04. These values show that conditional and ordinary probes select different directions under the declared control. Without external origin labels they do not prove causal recovery. Layer 3 is therefore a candidate locus for a subsequent intervention study, not a mechanistically established source.

17 CRA Results by Sensitive Attribute

Table 16: HSIC reduction and residual AUROC by sensitive attribute (CRA, mean across layers and configurations above the illustrative boundary). Gender shows the largest residual statistic among these injected attributes.

Attribute	Res. AUROC	Corr	CRA	Advantage
age_group	0.97	16.4	20.9	+4.5 pp
gender	0.99	18.2	25.3	+7.1 pp
hospital_type	0.91	18.7	20.4	+1.7 pp

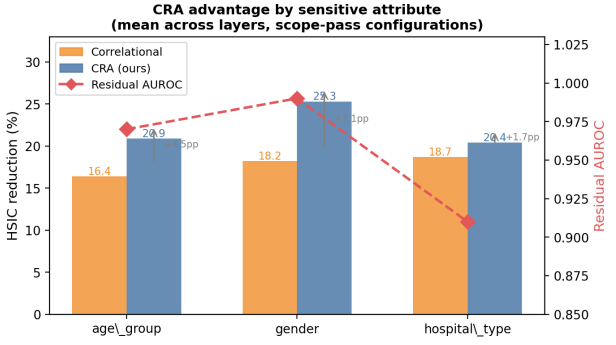

Figure 8: CRA advantage by sensitive attribute. Gender shows the largest advantage (+7.1 pp) and highest residual AUROC (0.99). hospital_type shows the smallest advantage and lowest residual AUROC (0.91).

Residual-signal strength under the selected mean-pooling

audit covaries with the observed CRA difference, but does not independently establish causal origin. gender has the highest residual AUROC (0.99), showing substantial gender-predictive structure after this particular control. The ordinary and residual probes emphasize different directions, producing the largest descriptive HSIC difference (+7.1 pp). This geometry is compatible with concentrated gender encoding reported in prior representation studies (Bolukbasi et al. 2016), but the present experiment does not identify attention heads, distinguish lexical from contextual causes, or validate a downstream consequence.

age_group has a moderate descriptive difference (+4.5 pp) and high residual AUROC (0.97), indicating probe-accessible information after the selected control; the rank-one theorem does not apply to the full attribute. hospital_type has the smallest difference (+1.7 pp) and residual AUROC 0.91. These values do not establish why the attributes remain detectable. Input terminology, multidirectional geometry, residualizer misspecification, and reference overlap are all plausible contributors. The single-direction HSIC differences are therefore descriptive and should not be treated as formal lower bounds or mechanism estimates.

18 Relationship to Causal Mediation Analysis

CRA is a statistical partialling procedure, not causal mediation. Natural direct and indirect effects require interventions and identification assumptions about confounding and mediator–outcome relations (Vig et al. 2020). None of those effects is identified by residual AUROC or HSIC reduction alone. In particular, there is no general inequality ordering a CRA HSIC change and a natural indirect effect without specifying the downstream structural equations, intervention, and metric.

Under the paper's linear representation model, CRA estimates the component of Z_l remaining after a declared control for Z_0. This can be useful for choosing subsequent counterfactual or mechanistic experiments, but it can absorb nonlinear input functions, contextual interactions, or reference misspecification. We therefore use "attribution" in the regression sense and reserve causal language for settings with explicit interventions.

19 Layer-Sensitivity Ranges (Depth Variability)

The **per-method mean HSIC reduction** values in Table 2 of the main paper are means across layers. The [min, max] ranges below reflect layer-range—minimum and maximum HSIC reduction across depth—not bootstrap data-sampling confidence intervals. Ranges are wide in high-variance configurations (PubMedQA GPT-2 variants, CRA range [39.8, 61.2]). Under the canonical Algorithm 1 estimator, the disaggregated 27-cell analysis gives the current win rate ($22/27$, $+5.31$ pp mean); these Table 2 layer-range values are illustrative of variance across depth, not the significance test.

These ranges reflect *depth variability*—how much HSIC reduction shifts across layers—not bootstrap sampling uncertainty. A wide range (e.g., PubMedQA/distilGPT-2: CRA [44.5, 62.8]) signals that layer choice substantially affects the magnitude of detected reference-residual dependence; reporting a single layer's HSIC reduction without specifying that layer overstates precision. Values are drawn directly from per-layer measurements, and results are consistent across the three distilGPT-2 configurations above the illustrative boundary (Appendix §7).

Table 17: Layer-sensitivity range for HSIC reduction (pp). Values in brackets are layer-range [min, max] across layers (reflects depth variability across model depth, not bootstrap data-sampling uncertainty). Configurations below the illustrative boundary are italicized. CRA min $>$ Corr min in 6/7 above-boundary configurations.

Dataset	Model	CRA [min,max]	Corr [min,max]
MedQA	dGPT-2	4.9 [4.2, 6.0]	4.6 [3.4, 5.8]
MedMCQA	dGPT-2	17.6 [12.4, 25.6]	13.2 [7.5, 21.8]
MedMCQA	GPT-2	16.1 [10.9, 24.0]	12.4 [6.6, 21.1]
MedMCQA	GPT-2-M	15.1 [8.8, 26.4]	16.1 [10.1, 26.4]
MedMCQA	*BioGPT*	*6.4 [3.6, 11.4]*	*16.0 [13.7, 18.6]*
MedMCQA	*Bio-ClinBERT*	*10.1 [5.9, 17.1]*	*15.7 [9.5, 23.0]*
PubMedQA	dGPT-2	52.7 [44.5, 62.8]	27.1 [14.9, 47.3]
PubMedQA	GPT-2	49.3 [39.8, 61.2]	25.9 [13.8, 46.7]
PubMedQA	GPT-2-M	39.4 [29.3, 54.9]	26.0 [15.3, 45.8]
PubMedQA	*BioGPT*	*4.9 [0.1, 14.3]*	*26.7 [25.0, 28.0]*
PubMedQA	Bio-ClinBERT	27.9 [25.2, 30.4]	22.5 [19.0, 26.6]

Across seven configurations above the illustrative boundary, CRA's minimum HSIC reduction equals or exceeds the correlational minimum in six. GPT-2-medium on MedMCQA is the exception, and its ranges overlap substantially ([8.8, 26.4] vs. [10.1, 26.4]). Configurations below the illustrative boundary have CRA ranges below the correlational ranges. These are descriptive layer ranges from correlated measurements, not independent replications, and no architectural mechanism is inferred from their ordering.

20 Rank-k Path B Ablation: Theorem 1 Condition (C2)

Theorem 1 requires $\text{rank}(B_l) = 1$ (Condition C2). Here we evaluate how CRA selectivity degrades as residual sensitive signal is distributed across $r_B \in \{1, 2, 5\}$ orthogonal directions.

Design note: This ablation uses a different geometry from the main SCM. The main SCM has $d = 64$ and a full-rank Z_0; here Z_0 occupies a $n_{\text{input}} = 20$-dim subspace within $d = 128$ ambient dimensions, giving Path B directions 108 orthogonal dimensions to work in (versus $\approx$0 extra dimensions when Z_0 fills $\mathbb{R}^{64}$). This makes Path A suppression substantially easier, which explains why the Gram-Schmidt-enforced $r_B = 1$ case achieves $\Delta_{\text{sel}} = 0.858$ here versus 0.493–0.557 in the main SCM. The two experiments are complementary: the main SCM tests sensitivity to residual sensitive signal strength in a realistic full-rank setting; this ablation tests sensitivity to r_B in a low-rank Z_0 setting where the rank effect is geometrically isolated.

Ablation parameters: Z_0 in $n_{\text{input}} = 20$-dim subspace, $d = 128$, $n = 500$, 15 seeds per rank.

Table 18: CRA selectivity vs. Path B rank r_B (15 seeds each). Δ_{sel} decreases gracefully but CRA wins at all ranks.

r_B	$\cos(\boldsymbol{v}_{\text{cra}}, v_B)\uparrow$	$\cos(\boldsymbol{v}_{\text{cra}}, v_A)\downarrow$	$\Delta_{\text{sel}}(\boldsymbol{v}_{\text{cra}})$	Win rate
1	0.887 ± 0.022	0.029 ± 0.014	$+0.858$	15/15
2	0.632 ± 0.026	0.026 ± 0.012	$+0.606$	15/15
5	0.415 ± 0.043	0.040 ± 0.019	$+0.375$	15/15

v_B: first (most signal-carrying) Path B direction.

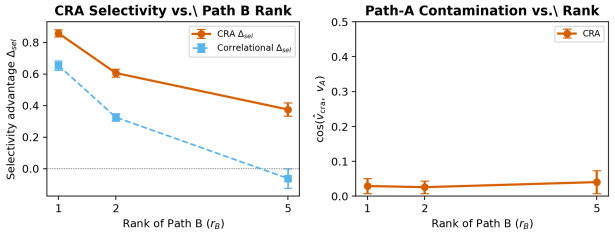

Figure 9: CRA selectivity Δ_{sel} (left) and Path-A contamination $\cos(\boldsymbol{v}_{\text{cra}}, v_A)$ (right) vs. Path B rank r_B. CRA maintains positive Δ_{sel} at all ranks; Path-A contamination stays near zero (<0.04) even at $r_B = 5$.

In this synthetic construction, CRA's alignment with the known Path-A direction ($\cos(\boldsymbol{v}_{\text{cra}}, v_A) < 0.04$) remains near zero at all tested ranks, showing low contamination under the generator's assumptions. Reduced $\cos(\boldsymbol{v}_{\text{cra}}, v_B)$ at higher r_B reflects that a single-direction probe captures only the most prominent Path B component—consistent with Condition (C2) being an approximation rather than a hard failure. Residual AUROC can remain high when the signal is multi-directional, but it does not recover the full residual subspace.

21 C2 In-Data Check: All distilGPT-2 Configurations

The main paper (§4.2) reports the singular-value ratio σ_1/σ_2 of the conditional-mean shift matrix $B = [\mu_{\text{male}} - \mu,\ \mu_{\text{female}} - \mu]^\top \in \mathbb{R}^{2\times d}$ for MedQA/distilGPT-2. Here we extend this in-data Condition (C2) check to the remaining two distilGPT-2 configurations above the illustrative boundary (MedMCQA/distilGPT-2 and PubMedQA/distilGPT-2). Ratios are computed from the stored layer representations using only samples labelled *male* or *female*.

Table 19: σ_1/σ_2 of $B = [\mu_{\text{pos}} - \mu,\ \mu_{\text{neg}} - \mu]^\top$ for all three distilGPT-2 configurations above the illustrative boundary. The large ratios show that the binary conditional-mean shift is approximately rank one is approximately rank one at every tested layer; this extends the C2 diagnostic beyond the single MedQA example reported in the main paper.

Config	L0	L1	L2	L3	L4
MedQA	39,296	39,058	35,064	41,941	50,153
MedMCQA	21,031	16,467	46,818	35,423	24,761
PubMedQA	22,408	15,636	18,452	24,852	18,851

Across all three configurations and all five layers, $\sigma_1/\sigma_2 >$ 15,000, showing that the conditional-mean shift matrix is numerically dominated by one direction throughout the tested depth. PubMedQA shows the lowest ratios of the three (minimum 15,636 at layer 1), which co-occurs with its higher residual statistic; even so, the ratio remains orders of magnitude above any threshold that would indicate multi-dimensional structure in B. Layer-3 ratios (41,941 / 35,423 / 24,852) corroborate the main-paper C2 diagnostic and are consistent with the rank-one approximation for binary gender in these configurations.

22 Z_0 Linear-Probe Diagnostic for Clinical Models

We quantify how strongly the sensitive attribute S is linearly predictable from Z_0 in BioGPT and Bio-ClinBERT; high held-out predictability demonstrates departure from exact Condition (C1) for the selected reference. We fit a 5-fold cross-validated logistic regression (class_weight=balanced) on Z_0 to predict gender S for each clinical model.

Table 20: $\mathbf{Z_0}$ linear probe AUROC for clinical models vs. distilGPT-2 control. *Full*: probe on raw Z_0; *PCA-50*: probe on top-50 PCA scores of Z_0; R^2: variance in S explained by the PCA subspace. High AUROC at layer 0 indicates C1 violation (S encoded in Z_0).

Dataset	Model	Full AUC	PCA-50 AUC	R^2
MedQA	BioGPT	0.732	0.707	0.037
MedQA	Bio-ClinBERT	0.690	0.614	—
PubMedQA	BioGPT	**1.000**	**1.000**	0.999
PubMedQA	Bio-ClinBERT	**1.000**	**1.000**	0.899
MedQA	dGPT-2 (ctrl)	0.831	0.671	—

BioGPT and Bio-ClinBERT on PubMedQA achieve layer-0 AUROC 1.00, so the sensitive label is highly predictable from the selected input embedding and exact C1 does not hold. MedQA layer-0 AUROCs of 0.690–0.732 also indicate nonzero reference–label overlap. For distilGPT-2, the Z_0 probe is 0.831 and the corresponding deeper-layer probe values are lower. These observations establish predictability under the chosen representations, but do not show that clinical pretraining "concentrates" a causal mechanism at Z_0 or that residualization has isolated a unique source.

23 Z_0 PCA Reconstruction and Reference-Fit Diagnostics

We compute the fraction of variance in Z_l explained by the top-50 PCA subspace of Z_0 and the alignment of the sensitive probe direction with that subspace. These are reference-fit and overlap diagnostics, not direct evidence that a uniquely defined Path A has been captured.

For MedQA and MedMCQA, the Z_0 PCA-50 subspace explains 22–29% of layer-l variance; for PubMedQA, 11–13%. These values quantify how much total layer variance the selected top-k reference reconstructs; they do not by

Table 21: Reference-fit diagnostics (distilGPT-2) across datasets and layers. *Var*: fraction of $\|Z_l\|_F^2$ explained by the Z_0 PCA-50 projection. *Align*: $\cos^2(\hat{v}_s, \Pi_{Z_0}\hat{v}_s)$, the overlap of the sensitive direction with the selected reference subspace.

Dataset	Layer	Var explained	Sensitive dir. alignment
MedQA	1	0.240	0.070
MedQA	3	0.261	0.052
MedQA	5	0.256	0.079
MedMCQA	1	0.223	0.078
MedMCQA	3	0.228	0.061
MedMCQA	5	0.245	0.060
PubMedQA	1	0.127	0.099
PubMedQA	3	0.113	0.064
PubMedQA	5	0.123	0.082

themselves identify sensitive Path A signal. The sensitive probe direction's small but nonzero alignment with the reference subspace (5–10% $\cos^2$) indicates reference overlap relevant to preservation. PubMedQA has lower total variance explained and the largest descriptive CRA–correlational difference, but this association does not establish independently added demographic encoding.

24 Injection Strength (ρ) Ablation

We report scope test residual AUROC across injection strength $\rho \in \{0.5, 0.75, 0.9\}$ on MedQA/distilGPT-2 to assess sensitivity to the injection correlation.

Table 22: Scope test residual AUROC per layer across injection strengths ρ (MedQA/distilGPT-2). **Important:** at $\rho = 0.50$, the injected tag has zero demographic correlation by construction ($P(\text{tag}|S=1) = P(\text{tag}|S=0) = 0.5$), yet residual AUROC stays at 0.69–0.84 across layers. This reflects *natural* demographic signal in MedQA text (age and gender appear explicitly in ≈94% of USMLE-style questions), not the injected tag, indicating that MedQA's natural demographics partially confound the controlled injection design. The archived residual statistic remains elevated across these injection settings, but this does not identify its source.

ρ	Layer 0	Layer 1	Layer 2	Layer 3	Layer 4
0.50	0.837	0.796	0.771	0.741	0.692
0.75	0.834	0.797	0.780	0.741	0.696
0.90	0.830	0.796	0.769	0.742	0.696

Residual AUROC changes by less than 0.01 across the three ρ values within each layer. Layer 0 is 0.830–0.837, layers 1–2 are 0.771–0.797, and layers 3–4 are 0.692–0.742. These are continuous descriptive values; without a training-only permutation null for this ablation, we do not assign pass/fail labels or infer a mechanism from the depth trend.

25 Expanded Held-Out Analysis: Real-Data Power and Real-Attribute Payoff

Held-out scoring and nesting disclosure. The probe coefficients used for residual AUROC are fit on training rows and scored on held-out rows (80/20 for the expanded-cell analysis; 5-fold cross-validation for natural attributes). However, PCA/ridge residualization is fit on the full covariate matrix, HSIC uses the full matrix, and the peak-scope layer is selected using the same data later summarized. The analysis is therefore transductive and layer-selection optimistic. Earlier in-sample probe AUROCs are discarded; the reported held-out probe scores are preferable, but they are still not a fully nested test as defined in Section 8.

Post-hoc 27-cell descriptive expansion. Disaggregating dataset–model–attribute cells and choosing each peak-scope layer yields 27 cells above the illustrative 0.80 boundary. CRA has a positive HSIC difference in 22/27 cells with a mean of $+5.31$ pp. These cells share three datasets, and the layer choice is made on the same data; therefore the naive one-sample $p = 0.014$ is not a valid confirmatory significance result. A dataset-clustered bootstrap includes zero. We retain the win count and per-dataset means only as a descriptive map of where residual and ordinary directions differ. No claim of independently powered real-data superiority is made.

Permutation calibration is the primary threshold procedure. For each prespecified family and reference, report held-out residual AUROC minus the 99th percentile of a training-only label-permutation null. The residual-probe pipeline is refit for every permutation. In two archived 200-permutation examples, the null median is about 0.50 and its 99th percentile about 0.60. The original 0.80 value is roughly 0.20 above those example nulls but has no universal error guarantee. Reports should include AUROC, its confidence interval, the null quantile, and their margin rather than a universal pass/fail label.

Real-attribute diagnostic: reference-explained gender vs. reference-residual age. On real, non-injected attributes (held-out 5-fold CV, distilGPT-2, canonical Algorithm 1 estimator), the residual statistic differs across attributes under the declared mean-pooling reference. *Gender* is lexically carried: residual AUROC 0.55–0.64 (Bias in Bios), 0.73–0.82 (MIMIC-style), 0.54–0.63 (MedQA)—mostly below the illustrative boundary; correlational probing removes more of this lexical signal under the declared control. In MIMIC-style clinical prose, age residual AUROC is 0.81–0.95, so age remains detectable after this control; this is not an independent label of causal origin. The *differential* HSIC benefit of CRA over correlational, however, is small and not statistically established: at layer 3 the point advantage is $+3.3$ pp but an example-level bootstrap ($n = 1186$) gives a 95% CI $[-54, +50]$ that includes zero. We therefore treat this as a diagnostic case (the residual statistic differs across the two attributes), not as a demonstrated payoff.

≥1B behavior (TinyLlama-1.1B and GPT-2-XL): residual detection remains, differential does not. We ex-

tract mean-pooled representations from TinyLlama-1.1B (Llama family) and GPT-2-XL (1.5B, GPT-2 family) on the same MedText corpus and repeat the held-out analysis with the canonical estimator. The residual statistic remains high for clinical age at both scales (residual AUROC $0.806/0.928$ TinyLlama layers 11/22; $0.952/0.876$ GPT-2-XL layers 24/48) while the corresponding gender statistic is lower or CRA-underperforming under the archived local comparison. The *CRA-vs-correlational differential*, however, collapses to ≈ 0 at scale: $-0.6/+0.1$ pp (TinyLlama) and $+0.8/+1.0$ pp (GPT-2-XL), with $\cos(\boldsymbol{v}_{\text{cra}}, \boldsymbol{v}_{\text{corr}}) \approx 0.99$. The cause is visible in the linear coupling: top-50 PCA of Z_0 explains only $R^2 \approx 0.03$ of Z_l for TinyLlama (an RMSNorm per-feature scale artifact that standardization restores to 0.52; §30) and $R^2 = 0.46/0.22$ for GPT-2-XL. When R^2 is small, residualization is close to the identity, so $\boldsymbol{v}_{\text{cra}}$ nearly coincides with $\boldsymbol{v}_{\text{corr}}$ and there is little Path A left to remove—exactly when a *differential* advantage is not expected. This is an explicit negative incremental-value result: under these specifications, ordinary probing is sufficient because residualization returns almost the same direction. A high residual statistic alone does not imply that CRA supplies a materially different candidate direction. The earlier larger $\geq$1B advantages ($+9.3/+7.8$ pp) were produced by a non-canonical estimator and do not survive correction. A nonlinear extension of the two-path model and more $\geq$1B corpora remain the natural next steps.

26 Robustness Controls: Rank, Non-Linear Probes, Token Presence

Rank of the reference-residual signal (C2) beyond gender. Condition (C2) assumes a rank-1 reference-residual direction. We compute the singular-value ratio σ_1/σ_2 of the class-conditional mean-shift matrix $B = [\mu_c - \mu]_c \in \mathbb{R}^{|\mathcal{C}| \times d}$ for each attribute (MedMCQA/MedQA, GPT-2). Gender is strongly rank-1 ($\sigma_1/\sigma_2 \in [17.8\text{k}, 37.5\text{k}]$), but the multiclass attributes are *multi-directional*: $\sigma_1/\sigma_2 \approx$ 1–2 for both age_group (3 classes) and hospital_type (4 classes). Thus C2 holds exactly only for binary gender. The scope test nonetheless still flags age_group as detectable across scales (residual AUROC above the archived local boundary) *despite* its multi-directional structure: the rank-1 probe estimates one dominant residual direction, and exact rank-1 recovery is *sufficient but not necessary* for the diagnostic. Iterative rank-k recovery (deflating recovered directions and re-probing) is the concrete extension for full multidirectional residual characterization.

Non-linear conditional probe (not a linear-probing artifact). To check that the scope test and path assignment are not artifacts of restricting to linear logistic regression, we replace the residual probe with a one-hidden-layer MLP (64 units, standardized inputs, 5-fold CV) on MedText/distilGPT-2. MLP residual AUROC tracks the linear value closely: age_group 0.89–0.93 (MLP) vs. 0.88–0.97 (linear); gender 0.73–0.86 vs. 0.73–0.82. A non-linear probe surfaces no additional additional residual signal, so this control does not reveal a large hidden nonlinear effect.

Token-presence (bag-of-words) sensitivity control. We residualize Z_l on a binary bag-of-words representation (top-2000 tokens, reduced to 50 dimensions) instead of Z_0, then probe. Age-group residual AUROC is 0.73–0.79 across layers 1–5, whereas gender is 0.16–0.22 under this particular control. The result supports only the narrow statement that age predictability is not removed by this bag-of-words reference while gender predictability is substantially reduced. It does not eliminate compositional, syntactic, contextual, or residualizer-misspecification explanations. Bag-of-words residualization is therefore a mandatory sensitivity control, not a proof of model-weight origin or a ground-truth Path-A/Path-B label.

27 C4 Diagnostic, HSIC Robustness, Cross-Family Calibration, Residualize-then-Erase

An operational C4 diagnostic beyond R^2. Condition (C4) requires A_l to map (approximately) within $\text{col}(V)$; if it escapes, residualizing on the top-k PCA of Z_0 leaves Path A leakage. A direct, practitioner-usable test is the *k-stability of residual AUROC*: increase k and check whether more Path A S-signal is removed. On MedText/age (layer 3), residual AUROC is 0.944, 0.944, 0.943, and 0.942 for $k = 25, 50, 100, 200$, respectively. The change beyond $k = 50$ is about 0.01. This stability is consistent with the chosen rank capturing the reference variation relevant to this probe, but it does not verify C4. A residual AUROC that *fell* sharply as k grows would flag rank sensitivity and call for a larger or data-driven rank.

HSIC kernel robustness. We compare CRA with the correlational baseline under four HSIC kernels: three RBF bandwidths, $\{0.5, 1, 2\}$ times the median, and one linear kernel. For rank-one binary gender on MedMCQA/GPT-2, the difference stays near -1 pp. For multidirectional age, the one-direction result changes sign: it is $+1.2$ pp at the median RBF bandwidth and negative at the others. Removing one direction from a multidirectional signal can therefore interact strongly with the kernel. For such attributes, we emphasize residual AUROC and direction divergence. Full multidirectional removal requires the rank-k extension (§26).

Cross-family permutation calibration. For the archived examples, the 99th-percentile label-permutation residual AUROC is 0.645 (RoBERTa), 0.631 (TinyLlama-1.1B), and 0.627 (GPT-2-XL). These values illustrate that the null distribution depends on the model family and audit specification. For a new family, the primary report is the continuous margin

$$m_l = \tau_l - q_{.99}^{\text{perm}},$$

together with a bootstrap interval for τ_l. No fixed offset is added to the null, and neither 0.80 nor a "null plus 0.15" rule is recommended. The permutation procedure requires sensitive labels, although it does not require ground-truth labels for the latent origin of the signal. The full residualizer and probe are fit within development folds for each fixed specification, and the lockbox is scored only after the specification is locked.

Residualize-then-erase (residualization vs. the probe). To separate residualization from the classifier, we apply LEACE to the residualized $Z_l^{\perp}$ and to raw Z_l (MedText/age, layer 3). LEACE on raw Z_l erases the *total* age signal (AUROC $0.952 \rightarrow 0.855$); LEACE on $Z_l^{\perp}$ erases much less ($0.944 \rightarrow 0.928$). This comparison shows that changing the representation supplied to the eraser changes the component it removes. It does not establish source or practical benefit; the method-specific intervention must be evaluated on downstream outcomes.

28 Powering the Age Result; Nonlinear Residualization at Scale

The age advantage is directionally consistent but not adequately powered (canonical Algorithm 1 estimator). Under the corrected estimator the natural-attribute age advantage is *not* statistically established. On MedText/age (distilGPT-2, layer 3, $n = 1186$) the example-level bootstrap of the CRA HSIC-reduction advantage over correlational gives a point estimate of $+3.3$ pp but a wide 95% CI $[-54, +50]$ that *includes zero* ($P(\text{adv} > 0) = 0.60$, 200 replicates). Across scales the advantage is only weakly directional: $4/5$ (model $\times$ layer) cells positive with tiny magnitudes on the $\geq$1B models ($+0.1$ pp TinyLlama, $+1.0$ pp GPT-2-XL), giving a sign-test $p = 0.19$ that does not exclude zero. The earlier stronger figures (a kernel-free Δ_r CI excluding zero, $8/8$ positive) were produced by a non-canonical estimator and do not survive correction. What *is* robust is the scope test: age has a higher residual statistic than gender under the declared mean-pooling control at all three scales. We therefore treat the real-attribute results as a diagnostic case study, with the synthetic SCM as primary evidence (§12).

Nonlinear residualization where the linear fit degrades. To address the linear premise weakening at scale, we replace the ridge-PCA residualization with *kernel-ridge* residualization (RBF on the top-50 PCA of Z_0). Where the linear fit collapses it is largely restored: on MedText/age the $Z_0 \rightarrow Z_l$ fit rises from $R^2 = 0.035$ to 0.558 (TinyLlama-1.1B) and 0.459 to 0.711 (GPT-2-XL); on distilGPT-2, where the linear fit is already strong (0.81), the kernel variant is comparable (0.75). Under this nonlinear control, age remains detectable in the reported configurations (residual AUROC 0.919 distilGPT-2, 0.813 TinyLlama, 0.904 GPT-2-XL). This shows robustness to one stronger residualizer, not model-weight origin. Kernel ridge is reported as a sensitivity residualizer, not a universally superior replacement; a rank-k deflation extension for the multidirectional signal requires several erasure rounds (a single direction leaves the re-probed multi-class AUROC intact), consistent with the C2 multi-directionality finding (§26).

29 Additional Reviewer Analyses: Subspace, Baselines, Finite-Sample, Pooling, C1, Correlated Attributes

Residualization subspace and conditional-probe baselines (why PCA(Z_0)). On MedMCQA/GPT-2 gender (binary; full AUROC 0.787) we compare Path-A controls by held-out residual AUROC: **PCA-50 ridge (CRA)** 0.738; **full-rank OLS on Z_0 without PCA truncation** 0.500 (chance); whitened PCA-50 0.393; CCA-20 0.735. At $n < d$, un-truncated partialling (the most direct "control-for-Z_0-then-probe" competitor) is ill-posed—OLS removes essentially all variance and removes nearly all residual variation, so PCA truncation is *necessary*, not cosmetic. CCA is a viable alternative (comparable to PCA); whitening over-removes by amplifying low-variance noise directions. A *joint* probe on $[Z_l, Z_0]$ reaches AUROC 0.789, essentially the same as the full Z_l probe (0.787). Together with a direct Z_0 probe and the residual AUROC, this is the required scalar baseline. It can reproduce residual detection but does not return a candidate feature-space direction or test whether the ordinary direction survives the declared control. Conversely, when the fitted directions are nearly identical, CRA reports no incremental value.

HSIC estimator details. We use the *unbiased* HSIC estimator (Song et al. 2012) with an RBF kernel at the median-heuristic bandwidth for both Z_l and S (subsampled to 400 for the $O(n^2)$ statistic), normalized to $[0, 1]$. Kernel/bandwidth sensitivity is reported in §27: rankings are kernel-stable for rank-1 attributes but bandwidth-sensitive for multi-directional ones, for which we rely on the kernel-free scope test and Δ_r.

Finite-sample variability at $n/d \approx 0.5$. Over 20 random subsamples ($n = 384$, $d = 768$) on MedText/age, the *scope decision* is stable—residual AUROC 0.891, 95% CI $[0.855, 0.923]$, always $> \tau^*$—while the individual $\boldsymbol{v}_{\text{cra}}$ *direction* is only moderately stable (pairwise cos $= 0.62$, min 0.53), consistent with the paper's caveat that at $n/d < 1$ the recovered direction is noisier than the scope verdict. The residual-AUROC statistic is more stable here than any single fitted direction.

A pooling-aware scope test. The verdict depends on pooling; this is part of the estimand. We therefore report every pooling fixed in advance as a separate audit axis. Across two MedText models, the observed pattern is:

Attr.	Model	mean-pool res. AUROC	last-tok res. AUROC
age	distilGPT-2	0.927 (residual)	0.701 (below gate)
	GPT-2	0.896 (residual)	0.675 (below gate)
gender	distilGPT-2	0.805 (residual)	0.904 (residual)
	GPT-2	0.831 (residual)	0.933 (residual)

Clinical *age* has a higher residual statistic under mean pooling than under last-token pooling, whereas gender remains high under last-token pooling. A plausible explanation is that the readouts emphasize different portions of the sequence, but these observations do not establish distributed age computation or a pronoun/name mechanism. The pooling-conditional report simply states which readout retains probe-accessible information after its corresponding reference control. We report mean pooling in the archived main results and require alternative pooling in the sensitivity envelope; attention/CLS pooling for encoders remains untested.

C1-violation extent on real attributes (Q: does Z_0-probe predict CRA?). Probing S directly from Z_0 quantifies departure from exact C1 and helps characterize the chosen control: gender is nearly fully Z_0-predictable (AUROC 0.99 MedText, 0.95 Bias in Bios $\Rightarrow$ mostly reference-explained, residual 0.76/0.55), whereas clinical age is only weakly Z_0-predictable ($0.68 \Rightarrow$ stronger residual signal, residual 0.94). The Z_0 probe is thus a second, complementary pre-audit signal: high Z_0 AUROC with a collapsing residual is a signature of signal largely explained by the chosen input reference.

Correlated attributes. In MedText, age and gender are nearly independent (NMI 0.014). Additionally residualizing Z_l on the gender one-hot alongside Z_0 leaves the age scope verdict essentially unchanged ($0.944 \to 0.942$): the residual age statistic is stable in this weakly correlated two-attribute setting. Stronger, deliberately-correlated multi-attribute injection is a natural stress test we leave to future work.

Uncertainty near the permutation null. Near a specification-specific null quantile, sampling noise can change the sign of $m_l = \tau_l - q_{.99}^{\text{perm}}$. We report a bootstrap interval for τ_l, the training-only permutation quantile, and their continuous margin. A configuration is treated as unresolved when the interval overlaps the null. Repeated layers or seeds are reported as a sensitivity envelope rather than combined by a post-hoc majority vote.

30 Alternative References, Standardization, Residualize-then-INLP, and Rank-r

Auditing domain-pretrained models with an alternative reference. Residualizing a clinical model's Z_l on a general-purpose model's Z_0 raises age residual AUROC (BioGPT $0.16 \to 0.97$; Bio-ClinBERT $0.46 \to 0.86$). This does not repair a failed audit or recover an invariant Path B; it changes the estimand to "what remains in the clinical model relative to a general-model reference on the same text." The comparison illustrates that CRA conclusions are reference-dependent. Any such analysis should report both the own-model and cross-model references and avoid causal wording.

Standardization improves the TinyLlama linear control. The low unstandardized $R^2 \approx 0.03$ on TinyLlama is partly attributable to feature-scale heterogeneity: RMSNorm rescales features to widely different magnitudes, so an unstandardized linear fit is dominated by a few high-variance dimensions. **Per-feature standardization restores the fit:** R^2: $0.034 \to 0.522$ (mid layer, MedText/age). Dropping the 20 highest-variance dimensions does *not* help (0.028), which is consistent with broad per-dimension scale heterogeneity rather than only a few high-variance coordinates." For this architecture, a defensible sensitivity analysis is therefore to standardize Z_0, Z_l before residualizing—and the kernel-ridge variant (§28) is an alternative; its measured overhead in this implementation is modest ($1.6\times$ linear CRA on distilGPT-2, and *faster* on TinyLlama since the RBF acts on the top-50 PCA scores, tens of milliseconds either way).

Residualize-then-INLP-5. Applying INLP-5 to the residual $Z_l^{\perp}$ rather than raw Z_l (MedText/age) reduces the multidirectional signal remaining after the chosen control: age AUROC changes from 0.952 to 0.679, versus 0.625 for INLP-5 on raw Z_l. The reported specialty accuracy is 0.632 versus 0.626. This single example shows that the representation supplied to the eraser changes what it removes; it does not establish source, fairness improvement, or a general utility advantage.

Iterative rank-r extension. For multidirectional attributes we use iterative deflation as an empirical extension: fit one residual direction, project it out, and re-probe. Span recovery requires idealized orthogonal mean shifts, spherical equal covariance, population estimation, and exact C1; no general finite-sample theorem is claimed. In the synthetic rank-r construction, subspace overlap rises from 0.45 to 0.59 for $r = 2$ and from 0.32 to 0.68 for $r = 3$ as the number of directions grows to r.

31 Prospective Prespecification Lockbox Stress Test

Purpose and status. This experiment evaluates the complete audit protocol prospectively on an untouched lockbox split, rather than retrofitting it to the archived LM analyses. It uses the real diabetes cohort distributed with scikit-learn ($n = 442$) (Efron et al. 2004; Pedregosa et al. 2011), rendered as structured tokens. It is therefore a nesting and decision-protocol stress test, not evidence about natural clinical language. The archived manuscript reports that a local manifest was written before lockbox scoring. The uploaded source bundle did not include that manifest or its underlying script, so this revision treats the claim as author-reported rather than independently reverified. It documents local prespecification within the study, not an externally timestamped preregistration.

Data, model, and split. The ten standardized clinical variables are discretized into four train-quantile bins and rendered as [CLS] followed by one feature-bin token per variable. A two-layer Transformer ($d = 48$, four heads) predicts whether the continuous disease-progression target exceeds the cohort median. Sensitive labels are the recorded binary sex code and age above the cohort median. The split is stratified jointly by task, sex, and age: 265 train, 88 calibration, and 89 lockbox records. Three fixed model seeds (17, 29, 43) are trained on train data and early-stopped on calibration data. Their lockbox task AUROCs are 0.755, 0.728, and 0.729.

Nested audit. The mandatory envelope crosses three specifications: mean-pooled Z_0, max-pooled Z_0, and a binary bag-of-token reference. We retain both Transformer layers and all three model seeds. Each attribute is evaluated on full input and after masking only its explicit token. For every cell, PCA rank (90% train variance, capped at 16), multi-output ridge, and logistic probing use only train+calibration data. The 99th-percentile null uses 200 label permutations and five development folds. Lockbox residual AUROC uses 2,000 bootstrap replicates; direction stability uses 100 stratified development bootstraps. No layer, seed, reference, or condition is chosen from lockbox outcomes. The code caches label-independent residualizer fits in the permutation loop;

this is algebraically equivalent to refitting them for every shuffled label.

Table 23: Prospective lockbox envelope. Each row contains 18 cells (3 model seeds $\times$ 3 specifications $\times$ 2 fixed layers). Positive means the lower 95% lockbox AUROC bound exceeds the development permutation $q_{.99}$; LB–null reports their difference.

Attr.	Input	Ref. AUC	Res. AUC	LB–null	Pos.
Age	full	1.000	.753–.935	+.063–+.300	18/18
Age	age masked	.521–.615	.495–.632	−.229–−.096	0/18
Sex	full	1.000	.437–.737	−.262–+.038	1/18
Sex	sex masked	.777–.788	.498–.681	−.213–−.025	0/18

Table 24: Age results by fixed specification and layer, aggregated only as ranges over the three prespecified model seeds. $q_{.99}$ is development-only.

Specification	Layer	Full res. AUC / $q_{.99}$	Masked res. AUC	Pos. full/mask
Mean Z_0	1	.838–.879 / .579–.594	.547–.607	3/3; 0/3
Mean Z_0	2	.796–.865 / .586–.595	.513–.560	3/3; 0/3
Max Z_0	1	.797–.875 / .581–.592	.495–.572	3/3; 0/3
Max Z_0	2	.753–.853 / .583–.613	.503–.632	3/3; 0/3
Bag of tokens	1	.930–.935 / .577–.587	.525–.588	3/3; 0/3
Bag of tokens	2	.913–.926 / .581–.587	.499–.558	3/3; 0/3

What the simpler probes would report. On the full age input, the direct-reference probe is 1.000, the ordinary Z_l probe is 0.935–1.000, the joint $[Z_l, Z_0]$ probe is 0.998–1.000, and the residual probe is 0.753–0.935. Read in isolation, the residual statistic could be described as information remaining beyond a linear reference. The mandatory input-cue control changes the audit decision: after masking the age token, direct, ordinary, joint, and residual AUROCs all fall to roughly 0.50–0.63, and no cell clears its local null. Thus the full-input residual is compatible with a nonlinear transformation of an explicit input cue or linear-control misspecification. The protocol reports *input-control-sensitive* and withholds any model-origin interpretation.

For recorded sex, only one of 18 full-input cells is positive and none of 18 masked cells is positive. The isolated cell occurs for one seed, mean pooling, layer 1; it is not robust to the mandatory envelope. The result is therefore no robust residual conclusion, despite nearly perfect direct and ordinary probes on the full input.

Directional and practical interpretation. The full-input age directions are individually stable across development bootstraps (median absolute cosine 0.762–0.954), and their absolute cosine with the ordinary direction ranges from 0.289 to 0.810. Stability alone is insufficient: the cue-mask reversal invalidates an invariant origin reading. This is the intended withholding behavior of the protocol and a concrete incremental decision beyond reporting one conditional probe. The experiment does not test fairness, privacy, clinical benefit, or a utility trade-off. Its external validity is limited by the small cohort, structured-token rendering, and small Transformer.

Reproducibility status. The uploaded materials contained the manuscript, supplement, and bibliography, but not the archived structured-token experiment's script, manifest, split indices, checkpoints, per-cell JSON files, or CSV summaries. We therefore preserve its reported numerical tables for transparency while avoiding claims that those artifacts accompany this revision. By contrast, the new calibration, controlled natural-form text, intervention, assumption-sweep, and runtime experiments are accompanied by the exact scripts, manifest, and machine-readable outputs in the revision package.

References

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. In *NeurIPS*.

Efron, B.; Hastie, T.; Johnstone, I.; and Tibshirani, R. 2004. Least Angle Regression. *The Annals of Statistics*, 32(2): 407–499.

Holm, S. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2): 65–70.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Phipson, B.; and Smyth, G. K. 2010. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1).

Vig, J.; Gehrmann, S.; Belinkov, Y.; Qian, S.; Nevo, D.; Singer, Y.; and Shieber, S. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *NeurIPS*.

Westfall, P. H.; and Young, S. S. 1993. *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley.