

Steering from Random Quotients: A Sample-Complexity Theory of Polysemantic Contrast Pairs

Khanh Nguyen^{1,3,*}, Luong Doan^{3,1}, Hai Nguyen^{1,3}, Hung Mai^{3,2},
Ngoc Vu³, Phong Ho³, Nhung Duong³, Tuan Do^{3,1,*},

¹Phenikaa University, Ha Noi, Vietnam, ²National Economics University, Ha Noi, Vietnam,

³B0Labs, N2TP Technology Solutions JSC, Ha Noi, Vietnam, *Corresponding authors

khanh.nguyenquoc@phenikaa-uni.edu.vn, luong.doan.wk@gmail.com, hai.nguyenduc@phenikaa-uni.edu.vn,

{pqhung.mai, maingoc.vu, phong.ho, nhung.duong}@n2tp.com, tuan.dongoc@phenikaa-uni.edu.vn

Abstract

Activation steering often averages activation differences between positive and negative prompts to construct a steering vector. This difference-in-means (DiM) estimator is simple, but a contrast pair can coactivate the intended concept together with topic, style, length, refusal, or truthfulness-related nuisance features. We model each pair as revealing the target vector through a quotient that removes its active nuisance span. The resulting polysemantic information operator determines which target components are identifiable and the number of contrast pairs needed to recover them. We establish identifiability, sample-complexity bounds, a coactivation phase diagram, and sparse-dictionary recovery guarantees. Synthetic and controlled experiments support the predicted quotient-recovery behavior. Public-data and Pythia-410M pilots instead evaluate recovery of the quotient estimand induced by chosen nuisance supports, not an externally validated semantic target. A controlled TruthfulQA counter-confounder illustrates the main NLP lesson: raw behavioral margin can reward a style-entangled DiM vector even when quotienting removes the injected style component.

1 Introduction

Activation steering changes language-model behavior without fine-tuning. A common approach is the difference-in-means (DiM) estimator, which averages positive-minus-negative activations over contrast pairs such as truthful versus false completions or harmless versus harmful responses.

Such contrast pairs rarely isolate a single property. Truthful and false answers may also differ in length, hedging, topic, or named entities; refusals can correlate with politeness; and safety directions can overlap with style. Under polysemanticity, DiM may therefore produce a behaviorally useful direction that remains statistically confounded.

We formalize this setting as random quotient steering: each contrast pair reveals the target only after the nuisance span active in that pair has been quotiented out. This view yields identifiability conditions, finite-sample rates, and sparse-dictionary guarantees for SAE-style steering. The theory is conditional on observed or credibly estimated nuisance supports: it characterizes recovery once those supports are supplied, but it does not discover them. Synthetic experiments test the model directly, whereas the public-data and Pythia-410M experiments are held-out quotient-recovery pilots under task-specific support constructions. A controlled TruthfulQA counter-confounder then shows why raw log-probability margin can prefer a style-entangled DiM vector. Our contribution is thus a conditional theory of when contrast pairs identify a target representation, together with an explicit account of when correlated prompt artifacts remain indistinguishable from that target.

2 Related Work

Activation steering and representation engineering provide the immediate NLP context. Activation Addition introduced optimization-free steering with activation-space vectors from prompt contrasts, while Contrastive Activation Addition scaled this idea to datasets of positive and negative behavior examples and explicitly used mean activation differences (Turner et al., 2023; Rinsky et al., 2024). Representation Engineering frames such methods as a pipeline for representation identification, operationalization, and control, and the recent TMLR taxonomy highlights reliability, multi-concept control, and evaluation as central open issues (Zou et al., 2023; Wehner et al., 2025). Related work on latent steering vectors, inference-time truth interventions, unsupervised latent knowledge, truth geometry, and model editing shows that linear internal directions can track NLP properties such as

truthfulness and factual associations (Subramani et al., 2022; Li et al., 2023; Burns et al., 2023; Marks and Tegmark, 2024; Meng et al., 2022). Our paper asks a complementary statistical question: when does a contrast-derived steering vector identify the intended concept rather than a correlated nuisance?

Superposition and sparse-feature work motivate our nuisance model. Toy superposition models explain how many features can share representation dimensions, and dictionary-learning work shows that sparse autoencoder features can be more interpretable than raw neuron coordinates (Elhage et al., 2022; Bricken et al., 2023; Huben et al., 2024). SAE-targeted steering and feature-guided activation additions are close empirical neighbors because they construct steering directions through sparse features (Chalnev et al., 2024; Soo et al., 2025). We differ by giving sample-complexity and lower-bound statements for sparse steering when active polysemantic confounders induce random quotient views.

Projection-based concept erasure is close in linear algebra but different in objective. Debiasing, INLP, and LEACE remove known protected or concept directions from representations (Bolukbasi et al., 2016; Ravfogel et al., 2020; Belrose et al., 2023). Random quotient steering instead estimates a target direction through pair-dependent nuisance spans. The same projectors appear, but the question is not how to erase a concept; it is which parts of a steering direction remain identifiable after arbitrary active nuisance is quotiented out.

Causal abstraction and safety work clarify the stakes of this distinction. Identifiability of causal abstractions studies when latent variables can be recovered from interventional or contrastive structure (Li et al., 2025). Recent safety analysis shows that steering vectors can overlap with refusal-related directions and alter jailbreak attack success (Li et al., 2026). Our matched-nuisance lower bound uses standard nonasymptotic tools, including matrix concentration, Lasso compatibility, Assouad, and Fano (Tropp, 2012; Wainwright, 2019; Buhlmann and van de Geer, 2011; Yu, 1997), but the experiment is steering-specific: nuisance amplitudes hide differences inside active spans, so only quotient-visible components contribute information.

3 Problem Setup

Let $h(x) \in \mathbb{R}^d$ denote a representation extracted from a fixed layer and token position of a trained model. A contrast pair (x_+^i, x_-^i) for a concept c produces the activation difference

$$Y_i = h(x_+^i) - h(x_-^i). \quad (1)$$

The target steering direction is a nonzero vector $\beta \in \mathbb{R}^d$. The unit direction is $\theta = \beta / \|\beta\|_2$, and $b = \|\beta\|_2$ denotes the signal norm.

The model includes a dictionary of nuisance directions $\Gamma = [\gamma_1, \dots, \gamma_p]$. For a support set $S \subseteq [p]$, define the active nuisance subspace $U_S = \text{span}\{\gamma_j : j \in S\}$. Let Π_{U_S} be the orthogonal projector onto U_S , and let $P_S = I - \Pi_{U_S}$ be the quotient projector onto $U_S^\perp$. The random support S_i is sampled from a concept-specific coactivation law Π_c . The contrast observation is

$$Y_i = \beta + u_i + \xi_i, \quad u_i \in U_{S_i}. \quad (2)$$

Nuisance amplitudes u_i are arbitrary within their active spans. If they were assumed mean-zero, the problem would reduce to ordinary mean estimation; here the nuisance can be adversarial or systematically biased, so the only safe information in pair i is the projection

$$P_i Y_i = P_i \beta + P_i \xi_i, \quad P_i = P_{S_i}. \quad (3)$$

The polysemantic information operator is

$$\mathcal{I}_c = \mathbb{E}_{S \sim \Pi_c} [P_S]. \quad (4)$$

For direction recovery, the relevant tangent space at θ is $T = \theta^\perp$, and the tangent information operator is $\mathcal{I}_{c,T} = P_T \mathcal{I}_c P_T|_T$ with $P_T = I - \theta\theta^\top$.

The theoretical results use the following standing assumptions. First, $S_1, \dots, S_m$ are independent draws from Π_c , and the corresponding projectors $P_i = P_{S_i}$ are observed. In an application, a practitioner must construct each candidate nuisance span independently of the target contrast, for example from controlled or null contrasts, known generation factors or metadata, SAE dictionary features, or randomized environments. The resulting span determines P_i , which is then supplied to QLS. Thus, the framework evaluates recovery conditional on that support construction; it does not infer nuisance supports from unconstrained contrast pairs. Second, each P_S is an orthogonal projector. Third, unless a theorem states otherwise, the noise variables are independent, mean-zero, and independent

of the supports; the high-probability upper bounds and minimax lower bounds are stated for Gaussian noise $\xi_i \sim N(0, \sigma^2 I_d)$. Fourth, nuisance amplitudes are unrestricted subject only to $u_i \in U_{S_i}$. Fifth, all statements are made on the identifiable subspace of $\mathcal{I}_c$; if $\mathcal{I}_c$ has a nullspace, components inside that nullspace are not identifiable from contrast pairs with arbitrary active nuisance.

These assumptions separate the positive and negative messages. With observed or controlled supports, quotient estimators remove active nuisance spans. Without support information, directions that always lie in active spans are unidentifiable.

In practice, we recommend the following use protocol. First, pre-specify candidate nuisances from the data-generation process or domain knowledge. Second, estimate their directions from separate null contrasts whenever possible. Third, inspect the spectrum of the empirical information operator to verify that the target remains visible and well conditioned after projection. Fourth, perturb plausible support choices to assess sensitivity. Finally, report downstream behavior together with direct measurements of nuisance leakage. If the supports cannot be justified or audited, QLS should be described as a projection heuristic rather than as recovery of a deconfounded semantic direction.

4 Estimators

The dense estimator is quotient least squares. With

$$A_m = \sum_{i=1}^m P_i, \quad r_m = \sum_{i=1}^m P_i Y_i, \quad (5)$$

set

$$\tilde{\beta} = A_m^\dagger r_m, \quad \hat{\theta} = \frac{\tilde{\beta}}{\|\tilde{\beta}\|_2}. \quad (6)$$

This is least squares for the projected equations $P_i Y_i = P_i \beta + P_i \xi_i$; when A_m is singular the pseudoinverse returns the minimum-norm solution on the visible range.

The sparse estimator assumes that β is sparse in a dictionary $B \in \mathbb{R}^{d \times p}$: $\beta = B\alpha^*$ with $\|\alpha^*\|_0 \leq s$. Sparse-Steer solves the projected Lasso

$$\hat{\alpha} \in \arg \min_{\alpha \in \mathbb{R}^p} \frac{1}{2m} \sum_{i=1}^m \|P_i(Y_i - B\alpha)\|_2^2 + \lambda \|\alpha\|_1, \quad (7)$$

and sets $\hat{\beta} = B\hat{\alpha}$. Here $\theta = \beta/\|\beta\|_2$ always denotes the target direction, whereas α^* denotes its sparse dictionary coefficients. The key population

quantity is not the ordinary Gram $B^\top B$, but the quotient Gram $B^\top \mathcal{I}_c B$ and its restricted compatibility over sparse cones.

5 Theory

5.1 Identifiability

The nullspace of $\mathcal{I}_c$ is exactly the set of directions hidden by the random quotient experiment.

Theorem 1 (Support-Free Impossibility). *Let $\mathcal{S}_c = \{S : \Pi_c(S) > 0\}$ denote the population support of the coactivation law. The population unidentifiable subspace is*

$$\mathcal{N} = \bigcap_{S \in \mathcal{S}_c} U_S = \{v : P_S v = 0 \text{ a.s. under } \Pi_c\}. \quad (8)$$

For any $v \in \mathcal{N}$, the parameters β and $\beta + v$ induce the same law for the observed contrast differences after a nuisance reparametrization. Consequently, components of β inside $\mathcal{N}$ are not identifiable. The finite-sample statement is that if a nonzero v satisfies $v \in U_{S_i}$ for every observed pair, then β and $\beta + v$ are indistinguishable from $Y_{1:m}$ alone, and as $m \rightarrow \infty$ this set of finite-sample-absorbable directions shrinks to $\mathcal{N}$.

Proof Sketch. Direct nuisance absorption. For $\beta' = \beta + v$ with $v \in U_{S_i}$, set $u'_i = u_i - v$; both u_i and v lie in U_{S_i} , so u'_i is admissible and $\beta' + u'_i = \beta + u_i$. The noise law is unchanged, so the data distribution is identical. The characterization $\mathcal{N} = \{v : P_S v = 0 \text{ a.s.}\}$ follows from Theorem 2. The full proof appears in Section A.1.

Theorem 2 (Spectral Phase Transition). *For every $u \in \mathbb{R}^d$,*

$$u^\top \mathcal{I}_c u = 0 \iff P_S u = 0 \text{ almost surely}. \quad (9)$$

Consequently, components in $\ker(\mathcal{I}_c)$ are unidentifiable under unrestricted nuisance. Conversely, if $\mathcal{I}_c$ is invertible on the target subspace, then $\beta = \mathcal{I}_c^{-1} \mathbb{E}[P_S Y]$.

Proof Sketch. Use $u^\top P_S u = \|P_S u\|_2^2$ for orthogonal projectors. Hence $u^\top \mathcal{I}_c u = \mathbb{E}\|P_S u\|_2^2$, which vanishes exactly when $P_S u = 0$ a.s.; the direction is then always inside the active nuisance span and can be absorbed as in Theorem 1. If $\mathcal{I}_c$ is invertible on the target subspace, projecting the observation eliminates nuisance and yields $\mathbb{E}[P_S Y] = \mathcal{I}_c \beta$. Full proof in Section A.2.

5.2 Dense Sample Complexity

The next theorem gives high-probability guarantees for QLS: a full-vector bound on the identifiable subspace, a direction bound for the identifiable direction $\theta_{\mathcal{R}} = \Pi_{\mathcal{R}}\beta / \|\Pi_{\mathcal{R}}\beta\|_2$ (with θ recovered only when $\beta \in \mathcal{R}$), and tangent bounds. The generic unconstrained tangent variance is governed by $\text{tr}(P_T \mathcal{I}_c^{-1} P_T)$; the sharper intrinsic rate $\text{tr}(\mathcal{I}_{c,T}^\dagger)$ is attained under radial/tangent decoupling or by a constrained local direction estimator.

Theorem 3 (QLS Upper Bound). *Let $\xi_i \sim N(0, \sigma^2 I_d)$ independently, let $\mathcal{R} = \text{range}(\mathcal{I}_c)$, and assume $\mathcal{I}_c$ is positive definite on $\mathcal{R}$. If*

$$m\lambda_{\min}(\mathcal{I}_c|_{\mathcal{R}}) \geq C \log(\dim(\mathcal{R})/\delta), \quad (10)$$

then with probability at least $1 - 2\delta$,

$$\|\Pi_{\mathcal{R}}(\tilde{\beta} - \beta)\|_2 \leq C\sigma \sqrt{\frac{\text{tr}((\mathcal{I}_c|_{\mathcal{R}})^\dagger) + \|(\mathcal{I}_c|_{\mathcal{R}})^\dagger\|_{\text{op}} \log(1/\delta)}{m}}. \quad (11)$$

Define the identifiable direction $\theta_{\mathcal{R}} = \Pi_{\mathcal{R}}\beta / \|\Pi_{\mathcal{R}}\beta\|_2$ and write $b_{\mathcal{R}} = \|\Pi_{\mathcal{R}}\beta\|_2$. If the right side is at most $b_{\mathcal{R}}/2$, normalization gives

$$\sin \angle(\hat{\theta}, \theta_{\mathcal{R}}) \leq \frac{2}{b_{\mathcal{R}}} \cdot (\text{right side}). \quad (12)$$

The angle between $\hat{\theta}$ and the original target θ is bounded by the same quantity only under $\beta \in \mathcal{R}$, in which case $b_{\mathcal{R}} = b$ and $\theta_{\mathcal{R}} = \theta$.

Assume $\beta \in \mathcal{R}$ for the tangent bounds below relative to $T = \theta^\perp$; invisible components in $\mathcal{R}^\perp$ otherwise project onto T and create irreducible tangent error. If $\beta \notin \mathcal{R}$, replace T by $T_{\mathcal{R}} = \mathcal{R} \cap \theta_{\mathcal{R}}^\perp$, β by $\Pi_{\mathcal{R}}\beta$, and P_T by $P_{T_{\mathcal{R}}}$. Writing $H_T = P_T(\mathcal{I}_c|_{\mathcal{R}})^\dagger P_T$, generic unconstrained QLS satisfies

$$\|P_T(\tilde{\beta} - \beta)\|_2 \leq C\sigma \sqrt{\frac{\text{tr}(H_T) + \|H_T\|_{\text{op}} \log(1/\delta)}{m}}. \quad (13)$$

The sharper intrinsic-tangent rate

$$\sigma \sqrt{\frac{\text{tr}(\mathcal{I}_{c,T}^\dagger) + \|\mathcal{I}_{c,T}^\dagger\|_{\text{op}} \log(1/\delta)}{m}} \quad (14)$$

is attained by either of the following estimators, with $\mathcal{I}_{c,T}$ assumed positive definite and the analogous matrix Chernoff sample condition:

(i) Unconstrained QLS under empirical decoupling. If $P_T A_m \theta = 0$ —implied by the almost-sure per-sample condition $P_T P_S \theta = 0$ for $S \sim \Pi_c$ —then $\|P_T(\tilde{\beta} - \beta)\|_2$ is bounded by C times the displayed

rate. The weaker population condition $P_T \mathcal{I}_c \theta = 0$ makes only $\mathbb{E}[A_m]$ block diagonal and incurs a Schur-complement inflation that vanishes as $m \rightarrow \infty$.

(ii) Constrained local direction estimator (baseline form). Fix a baseline unit direction θ_0 and a local radius $0 < r \leq b/2$ (so that $\sqrt{b^2 - \|z\|_2^2}$ is real and bounded below by $b\sqrt{3}/2$ for all $\|z\|_2 \leq r$), and let $T_0 = \theta_0^\perp$. Define

$$\hat{z} \in \arg \min_{z \in T_0, \|z\|_2 \leq r} \sum_{i=1}^m \|P_i(Y_i - \sqrt{b^2 - \|z\|_2^2} \theta_0 - z)\|_2^2, \quad (15)$$

$$\hat{\theta}_{\text{loc}} = \frac{\sqrt{b^2 - \|\hat{z}\|_2^2} \theta_0 + \hat{z}}{b}. \quad (16)$$

The target is $z_0 = P_{T_0}\beta$ (not $P_T\beta$, which is identically zero when $\theta = \beta/b$). The exact local information matrix for z at z_0 is the pullback

$$G_{\text{loc}}(z_0) = J(z_0)^\top \mathcal{I}_c J(z_0), \quad (17)$$

$$J(z_0)h = h - \frac{z_0^\top h}{\sqrt{b^2 - \|z_0\|_2^2}} \theta_0, \quad (18)$$

for $h \in T_0$. Under the chart regularity $r_0 \leq b/2$ —which gives $\|J(z)\|_{\text{op}} \leq 2$ and $\|D^2\beta(z)\|_{\text{op}} \leq 2/b$ on $\{\|z\|_2 \leq r_0\}$ —together with the matrix-Chernoff condition $m\lambda_{\min}(G_{\text{loc}}(z_0)) \geq C \log(\dim(T_0)/\delta)$, the sharp baseline-estimator rate is

$$\|\hat{z} - z_0\|_2 \leq C\sigma \left(1 + \frac{r_0}{b}\right) \times \sqrt{\frac{\text{tr}(G_{\text{loc}}(z_0)^{-1}) + \|G_{\text{loc}}(z_0)^{-1}\|_{\text{op}} \log(1/\delta)}{m}} + C \frac{r_0^2}{b}, \quad (19)$$

without any decoupling assumption on A_m . For $\beta \in \mathcal{B}_{\text{loc}}(\theta_0, r_0)$ with $r_0 \leq r$, the linearized model $\beta(z) \approx b\theta_0 + z$ replaces $J(z_0)$ by the identity on T_0 and $G_{\text{loc}}(z_0)$ by $G_0 = E_0^\top \mathcal{I}_c E_0$ where E_0 spans T_0 . In this regime $\|\hat{z} - z_0\|_2$ is bounded by C times the same display with G_0 in place of $G_{\text{loc}}(z_0)$, plus a coupling-aware deterministic bias

$$\frac{r_0^2}{b} \|G_0^{-1} E_0^\top \mathcal{I}_c \theta_0\|_2. \quad (20)$$

This reduces to $O(r_0^2/b)$ under bounded radial-tangent coupling $\|G_0^{-1} E_0^\top \mathcal{I}_c \theta_0\|_2 = O(1)$ or under decoupling $E_0^\top \mathcal{I}_c \theta_0 = 0$; otherwise the bias is amplified by $\|G_0^{-1} E_0^\top \mathcal{I}_c \theta_0\|_2$. Setting $\theta_0 = \theta$ recovers the oracle tangent estimator with $J(z_0) = I$ and $G_{\text{loc}}(0) = \mathcal{I}_{c,T}$.

Under the additional event $\|\tilde{\beta} - \beta\|_2 \leq b/2$, the bound in (i) implies $\sin \angle(\hat{\theta}, \theta) \leq (2/b)\|P_T(\tilde{\beta} - \beta)\|_2$; an analogous direction bound holds for $\hat{\theta}_{\text{loc}}$ in (ii).

Proof Sketch. Condition on the projectors. Since $P_i u_i = 0$, $r_m = A_m \beta + \zeta_m$ with $\zeta_m = \sum_i P_i \xi_i$. Conditionally on $P_{1:m}$, $\zeta_m \sim N(0, \sigma^2 A_m)$, so $\tilde{\beta} - \Pi_{\mathcal{R}} \beta \sim N(0, \sigma^2 A_m^\dagger)$ on the visible range. Matrix Chernoff lower bounds A_m by a constant multiple of $m\mathcal{I}_c$ on $\mathcal{R}$, and Gaussian norm concentration gives the trace-pseudoinverse rate. The generic tangent bound is the Schur block $P_T(\mathcal{I}_c|_{\mathcal{R}})^\dagger P_T$; under decoupling, this block collapses to $\mathcal{I}_{c,T}^\dagger$. Full proof in Section A.3.

Theorem 4 (Matched-Nuisance Minimax Lower Bound). *Fix a baseline unit direction $\theta_0 \in \mathbb{R}^d$ and let $T_0 = \theta_0^\perp$. Let $T_{\text{id}} = T_0 \cap \text{range}(\mathcal{I}_c)$, let $E \in \mathbb{R}^{d \times r}$ be any orthonormal basis matrix for T_{id} , and define the identifiable tangent information matrix $G_T = E^\top \mathcal{I}_c E$. Assume Gaussian noise $\xi_i \sim N(0, \sigma^2 I_d)$, unrestricted active-span nuisance, and $G_T \succ 0$. For the local fixed-norm class*

$$\mathcal{B}_{\text{loc}}(\theta_0, r_0) = \left\{ \sqrt{b^2 - \|z\|_2^2} \theta_0 + z : z \in T_{\text{id}}, \|z\|_2 \leq r_0 \right\}, \quad (21)$$

with $r_0 \leq b/2$, write $\mathcal{U}_\beta = \{(u_i)_{i=1}^m : u_i \in U_{S_i}\}$ and let $P_{\beta,u}^m$ denote the law of $Y_{1:m} = \beta + u_{1:m} + \xi_{1:m}$. The composite-nuisance local minimax risk satisfies

$$\inf_{\hat{\theta}} \sup_{\beta \in \mathcal{B}_{\text{loc}}(\theta_0, r_0)} \sup_{u \in \mathcal{U}_\beta} \mathbb{E}_{P_{\beta,u}^m} \sin^2 \angle(\hat{\theta}, \beta / \|\beta\|_2) \geq c \min \left\{ \frac{r_0^2}{b^2}, \frac{\sigma^2}{mb^2} \text{tr}(G_T^{-1}) \right\}. \quad (22)$$

The inner $\sup_{u \in \mathcal{U}_\beta}$ is essential: the matched-nuisance proof exhibits a worst-case nuisance that hides the active-span component of Δ . The local radius r_0 caps the rate: small r_0 gives r_0^2/b^2 , otherwise the trace term dominates. Setting $r_0 = b/2$ recovers the standard $\min\{1, \dots\}$ form.

Moreover, the bound is irreducible along unidentifiable tangent directions: for any unit $v \in \ker(\mathcal{I}_c) \cap T_0$ and any $0 < \tau \leq b/\sqrt{2}$, the two fixed-norm alternatives

$$\beta_\pm = \sqrt{b^2 - \tau^2} \theta_0 \pm \tau v \quad (23)$$

are observationally indistinguishable under matched nuisance, yielding the sample-size-independent bound

$$\inf_{\hat{\theta}} \max_{\pm} \sup_{u \in \mathcal{U}_{\beta_\pm}} \mathbb{E}_{P_{\beta_\pm, u}^m} \sin^2 \angle(\hat{\theta}, \beta_\pm / b) \geq \tau^2 / b^2, \quad (24)$$

which holds for every m .

Proof Sketch. A quotient Assouad construction. For alternatives β, β' with $\Delta = \beta - \beta'$, take $u_i = 0$ under the first parameter and $u'_i = (I - P_i)\Delta$ under the second (admissible since $I - P_i$ projects onto U_{S_i}). The conditional mean gap is $P_i \Delta$, giving $\mathbb{E}_S \text{KL}(P_\beta^m, P_{\beta'}^m) = \frac{m}{2\sigma^2} \Delta^\top \mathcal{I}_c \Delta$. Packing the tangent eigenbasis of G_T with amplitudes $a_k = \alpha t / \sqrt{\lambda_k}$ and $t^2 = \min\{\sigma^2/m, r_0^2 / \text{tr}(G_T^{-1})\}$ keeps $\|z_\omega\|_2 \leq r_0$ and yields the trace lower bound. The irreducible clause follows because $v \in \ker(\mathcal{I}_c) \cap T_0$ forces $\Delta^\top \mathcal{I}_c \Delta = 0$, so KL is zero for all m . Full proof in Section A.4.

5.3 Explicit Phase Diagram

Corollary 5 (Explicit Polysemantic Phase Diagram). *Suppose a fixed q -dimensional nuisance subspace U coactivates with probability p :*

$$P_S = \begin{cases} I - \Pi_U, & \text{w.p. } p, \\ I, & \text{w.p. } 1 - p. \end{cases} \quad (25)$$

Then $\mathcal{I}_c = p(I - \Pi_U) + (1 - p)I = I - p\Pi_U$. The eigenvalues are $1 - p$ on U and 1 on $U^\perp$. Thus, for $p < 1$,

$$\text{tr}(\mathcal{I}_c^\dagger) = d - q + \frac{q}{1 - p}. \quad (26)$$

As $p \rightarrow 1$, complexity inside U diverges; at $p = 1$ these components are unidentifiable. Any tangent component lying in U inherits the same $1/(1 - p)$ blow-up.

Proof Sketch. Spectral: $\mathcal{I}_c = I - p\Pi_U$ has eigenvalue $1 - p$ on U and 1 on $U^\perp$. Taking the pseudoinverse trace gives the display; the $p = 1$ case follows from Theorem 2. Full proof in Section A.5.

5.4 Sparse-Steer

Theorem 6 (Sparse-Steer Upper Bound). *Let $J = \text{supp}(\alpha^*)$ and define the cone*

$$\mathcal{C}(J) = \{\Delta : \|\Delta_{J^c}\|_1 \leq 3\|\Delta_J\|_1\}. \quad (27)$$

Let $\Phi_s > 0$ denote the empirical quotient compatibility constant. Assume the empirical quotient

compatibility inequality

$$\frac{1}{m} \sum_i \|P_i B \Delta\|_2^2 \geq \frac{\Phi_s}{2} \|B \Delta\|_2 \frac{\|\Delta_J\|_1}{\sqrt{s}} \quad (28)$$

holds for all $\Delta \in \mathcal{C}(J)$. If the score $g = \frac{1}{m} \sum_i B^\top P_i \xi_i$ satisfies $\lambda \geq 2\|g\|_\infty$, then the projected Lasso obeys

$$\|\hat{\beta} - \beta\|_2 \leq \frac{6\lambda\sqrt{s}}{\Phi_s}. \quad (29)$$

For Gaussian or sub-Gaussian noise, define the empirical projected column norm $L_{\max} = \max_{j \in [p]} \left(\frac{1}{m} \sum_{i=1}^m \|P_i b_j\|_2^2 \right)^{1/2}$, where b_j is the j -th column of B . Then $\lambda \asymp \sigma L_{\max} \sqrt{\log(p/\delta)/m}$ satisfies $\lambda \geq 2\|g\|_\infty$ on the relevant event and yields

$$\|\hat{\beta} - \beta\|_2 \lesssim \frac{\sigma L_{\max}}{\Phi_s} \sqrt{\frac{s \log(p/\delta)}{m}}. \quad (30)$$

The standard $L_{\max}$ -free rate follows by rescaling projected dictionary columns to $L_{\max} \leq 1$.

Proof Sketch. Lasso basic inequality after projection. Projected equations remove nuisance, the score condition gives the usual sparse cone, and quotient compatibility converts projected prediction error into $\|B(\hat{\alpha} - \alpha^*)\|_2$ error. The score variance scales with $L_{\max}$, which sets λ . Full proof in Section A.6.

Proposition 7 (Sparse Quotient Lower Bound). *The following bound parallels Theorem 4 for the sparse projected experiment. For fixed support J , if $G_J = B_J^\top \mathcal{I}_c B_J \succ 0$ and $B_J^\top B_J \succeq \kappa_J I$, then*

$$\inf_{\hat{\beta}} \sup_{\alpha: \text{supp}(\alpha) \subseteq J} \sup_{u \in \mathcal{U}_{B\alpha}} \mathbb{E}_{P_{B\alpha, u}^m} \|\hat{\beta} - B\alpha\|_2^2 \geq c \kappa_J \frac{\sigma^2}{m} \text{tr}(G_J^\dagger). \quad (31)$$

For unknown support, under atom separation and incoherence, sparse direction recovery incurs the usual $s \log(p/s)$ support-search factor with quotient visibility in the denominator; under the same composite-nuisance supremum.

Proof Sketch. Matched-nuisance quotient divergence on sparse packings. A fixed-support Assouad construction gives the $\text{tr}(G_J^\dagger)$ term; Varshamov–Gilbert over separated supports gives the $s \log(p/s)$ factor. Full proof in Section A.7.

5.5 Multi-Environment Deconfounding

When pair-specific supports are unobserved, variation across randomized environments can provide an alternative source of identifying information under the rank conditions below.

Proposition 8 (Multi-Environment Deconfounding). *Suppose environment means satisfy $\mu_e = \beta + \sum_{j=1}^p \pi_{ej} \gamma_j$, or stacked $M = \mathbf{1}\beta^\top + \Pi\Gamma^\top$. If $w^\top \mathbf{1} = 1$ and $w^\top \Pi\Gamma^\top = 0$, then $w^\top M = \beta^\top$. Let $C = \Pi\Gamma^\top$ denote the confounder mean matrix, and let R_C, R_Π project onto $\text{col}(C)^\perp, \text{col}(\Pi)^\perp$ respectively. Two regimes apply.*

Known Γ (sharp condition). *A feasible w exists iff $\mathbf{1} \notin \text{col}(C)$, equivalently $R_C \mathbf{1} \neq 0$. The minimum-norm feasible weight is $w^* = R_C \mathbf{1} / \|R_C \mathbf{1}\|_2^2$.*

Unknown or worst-case Γ (dictionary-uniform sufficient condition). *If w must annihilate Π itself, the feasibility condition is the stronger $R_\Pi \mathbf{1} \neq 0$ with $w^* = R_\Pi \mathbf{1} / \|R_\Pi \mathbf{1}\|_2^2$. Since $\text{col}(C) \subseteq \text{col}(\Pi)$, $R_\Pi \mathbf{1} \neq 0$ implies $R_C \mathbf{1} \neq 0$ but the converse can fail.*

In either regime, $\|w^*\|_2$ grows as $\mathbf{1}$ approaches the relevant column span; at the boundary, no annihilating weighted-average estimator exists. If only the nuisance column span is known, β is unidentifiable by environment weighting at the boundary; if C is known exactly, β can be recovered by subtraction $M - C = \mathbf{1}\beta^\top$. Environment variation can substitute for support observation under the right rank condition.

Proof Sketch. Linear algebra. Any feasible w annihilates the confounder mean matrix while preserving the target row mean, and the displayed w^* is the minimum-norm feasible vector by projection and Cauchy–Schwarz. Full proof in Section A.8.

6 Experiments and Analysis

Experiments address two questions. Quotient-recovery experiments measure how accurately an estimator recovers a target defined by the experimental support construction. Functional steering experiments measure downstream log-probability margin. These outcomes need not agree: a downstream metric may reward a nuisance component that QLS removes. We therefore interpret them as an objective trade-off, not as evidence that either estimator is universally superior.

6.1 Experimental Protocol

The synthetic experiment uses the exact random quotient model with representation dimension $d = 96$, nuisance-subspace dimension $q = 16$, noise $\sigma = 0.35$, nuisance scale 4.0, coactivation probabilities $p \in \{0, 0.25, 0.5, 0.75, 0.9, 0.97\}$, sample sizes $m \in \{16, 32, 64, 128, 256, 512, 1024\}$, and seeds 17, 42, 137. The target vector and confounder subspace are randomly generated. DiM averages raw Y_i , while QLS uses the known projector for each sample.

The public text pilots use SMS Spam (Almeida et al., 2011), IMDB sentiment (Maas et al., 2011), and TruthfulQA (TQA) (Lin et al., 2022). Text is represented by TF-IDF followed by truncated SVD and standardization. Confounder supports are heuristic and task-specific. SMS uses lexical features such as currency, all-caps tokens, message length, exclamation marks, and URL-like patterns. IMDB uses review length, exclamation, and related style indicators. TruthfulQA uses category and answer-style features. We call these experiments *held-out quotient-recovery pilots*. They report $\sin \angle(\hat{\beta}_m, \beta_*)$, where β_* is a large held-out estimate of the quotient estimand induced by the chosen nuisance supports and available pair pool. It is not an externally grounded semantic target; the pilots test finite-sample recovery under the chosen quotient construction.

The small-LLM pilot uses cached EleutherAI/pythia-410m hidden states on TruthfulQA (CUDA). Final hidden states for truthful and false answer completions are projected from 1024 to 192 dimensions by a fixed Gaussian orthoprojector. Contrast pairs use the same question with best answer versus first incorrect answer.

The sparse experiments evaluate two dictionary settings. The theorem-aligned synthetic setting uses $B = I$ with $d = 192$, sparsity $s = 8$, 32 confounders, sample sizes $m \in \{128, 256, 512, 1024, 2048, 4096\}$, and seeds 17, 42, 137, 233, 404. The learned-dictionary setting trains a 384-component SAE-style decoder dictionary on Pythia TruthfulQA hidden states on CUDA and then runs Sparse-Steer in that dictionary.

The functional TruthfulQA experiments patch Pythia-410M activations and measure truthful-versus-false answer log-probability margin. The validation-selected sweep uses layers

Experiment	Setting	Baseline	QLS/Sparse
Synthetic	$p=0, m=1024$	0.099	0.099
Synthetic	$p=0.97, m=1024$	0.464	0.258
SMS Spam	$m=1024$	0.798	0.195
IMDB	$m=1024$	0.157	0.156
TQA TF-IDF	$m=512$	0.196	0.128
TQA Pythia	$m=512$	0.205	0.175
Sparse $B=I$	$m=128$	0.236	0.128
Sparse $B=I$	$m=512$	0.118	0.062
Pythia SAE	$m=32$	0.743	0.663
Pythia SAE	$m=512$	0.166	0.243

Table 1: Angle-recovery and sparse-steering summary. The baseline column reports DiM except in the sparse rows, where it reports dense QLS; the final column reports QLS or Sparse-Steer, respectively. Values are descriptive point estimates; seed counts are reported in Section B.

$\{6, 10, 14, 18, 22\}$, patch modes last_prompt, all_prompt, and answer_prefix, signed alphas $\{-8, -4, -2, -1, -0.5, 0.5, 1, 2, 4, 8\}$, and seeds 17, 42, 137. It selects each method’s configuration on validation and evaluates once on held-out test.

The controlled counter-confounder experiments inject an explicit nuisance. Training contrasts pair truthful answers with Certainly, and false answers with Maybe,. Null contrasts that change only the prefix estimate the style direction. The functional counter-style test reverses the cue. The representation diagnostic uses a held-out *prefix-balanced operational target*: the average truth-versus-false difference across both prefixes. Balancing removes the injected prefix association, but this operational target is not claimed to be absolute semantic ground truth.

6.2 Angle Recovery and Sparse Steering

The synthetic phase diagram matches predicted behavior. At $m = 1024$, $p = 0$ both methods give mean sine angle 0.099. At $p = 0.97$, QLS reaches 0.258 while DiM stays at 0.464. Figure 1 (left) visualizes the separation; Table 1 places the result alongside the public-data and sparse-steering pilots. As $p \rightarrow 1$, nuisance visibility shrinks by $1 - p$ and sample complexity grows, matching Corollary 5.

Under the chosen quotient targets, SMS is the clearest public-data example: at $m = 1024$, DiM is 0.798 versus QLS 0.195. TruthfulQA shows consistent gains: TF-IDF QLS improves from 0.196 to 0.128 at $m = 512$; Pythia hidden-state QLS from 0.205 to 0.175. IMDB is neutral (0.157 versus 0.156), suggesting the constructed IMDB confounders do not create persistent bias. These comparisons establish estimator recovery relative to β_* , not the semantic validity of the support construc-

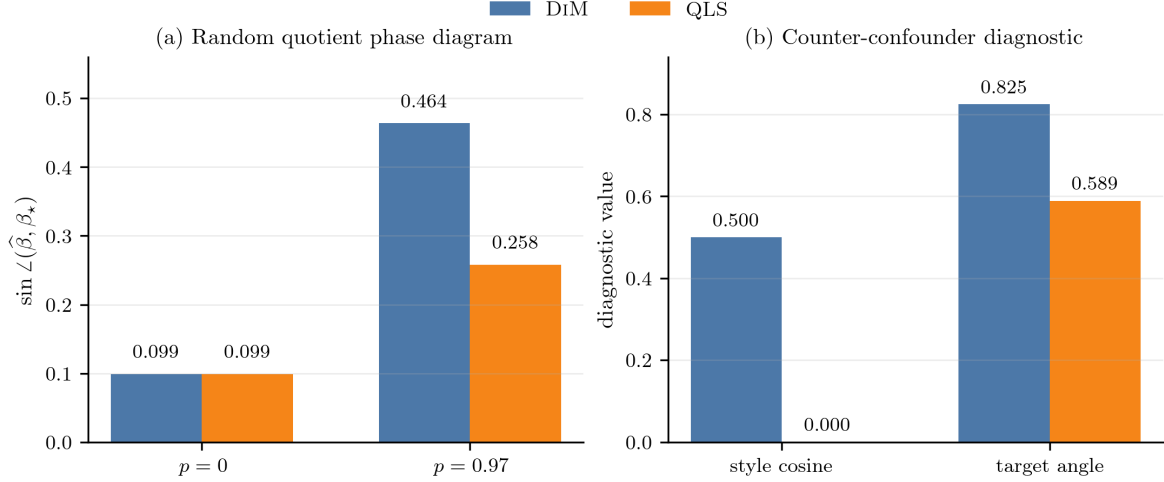


Figure 1: Quotient-recovery results across synthetic, public-text, Pythia hidden-state, and controlled counter-confounder regimes. Public-text and Pythia targets are held-out quotient estimates under chosen supports; the controlled target is prefix-balanced.

tion.

Sparse-Steer behaves as predicted in the compatible synthetic setting. With $B = I$ and $s = 8$, it beats full-space QLS at small and medium m : 0.128 versus 0.236 at $m = 128$, 0.062 versus 0.118 at $m = 512$. The learned Pythia SAE dictionary is more delicate: it improves over DiM at small m (0.663 versus 0.743 at $m = 32$) but does not dominate QLS at larger m (0.243 versus 0.166 at $m = 512$). Sparse recovery depends on dictionary compatibility after quotienting, as predicted by Theorem 6.

6.3 Functional Steering and Counter-Confounding

Functional steering does not yield a universal QLS win. In the validation-selected TruthfulQA sweep, the no-steer test margin is -0.502 ; DiM improves it to -0.262 ± 0.005 , QLS to -0.273 ± 0.003 . Accuracy is similar (DiM 0.419 ± 0.017 , QLS 0.416 ± 0.005). Both improve the baseline; raw margin slightly favors DiM.

The controlled counter-confounder experiments explain why. On the reversed-style functional test, no-steer margin is -0.382 , DiM reaches -0.081 , QLS -0.207 . The selected DiM vector has style cosine 0.500 while QLS has style cosine approximately zero. Figure 1 (right) and Table 2 show the split: DiM scores better on raw margin while QLS removes the style component and is closer to the prefix-balanced operational target.

Difference-in-means estimates whatever systematic components are present in raw contrast dif-

Counter-confounder metric	DiM	QLS
Reversed-style margin	-0.081	-0.207
Style cosine of vector	0.500	≈ 0
Angle to prefix-balanced target	0.825	0.589

Table 2: Counter-confounder trade-off between raw behavioral margin and prefix leakage.

ferences; when those include nuisance directions, DiM can retain components that improve the behavioral metric. QLS instead estimates the quotient target through projected moments. In this experiment it reduces style leakage and approaches the prefix-balanced operational target, but achieves a lower raw margin than DiM and essentially the same accuracy in the unmodified TruthfulQA sweep. The result is therefore an objective trade-off between raw margin and nuisance removal, not behavioral superiority of QLS.

The synthetic experiment matches the mathematical model exactly. The public text and Pythia angle experiments use heuristic supports and serve as pilots. The controlled counter-confounder experiment aligns most closely with the theory because the nuisance is injected and estimated from null contrasts. The SAE experiment shows that sparse dictionaries depend on quotient compatibility, an empirical question.

7 Limitations

The theory assumes observed or credibly estimated active supports. Under unrestricted nuisance, di-

rections always coactivated with the concept can be absorbed into nuisance amplitudes (Theorem 1). Supports may come from metadata, prompt construction, null contrasts, SAE activations, or randomized environments (Proposition 8), but QLS does not discover them. If a chosen support is not credible, the result should be interpreted only as recovery of the quotient induced by that construction.

A deterministic misspecification decomposition makes this dependence explicit. Let $\hat{P}_i$ be orthogonal estimated projectors, $\hat{A} = \sum_{i=1}^m \hat{P}_i$, and $\hat{R} = \text{range}(\hat{A})$. Because $P_i u_i = 0$, the corresponding QLS estimate satisfies

$$\hat{\beta} - \Pi_{\hat{R}} \beta = \hat{A}^\dagger \sum_{i=1}^m \hat{P}_i \xi_i + \hat{A}^\dagger \sum_{i=1}^m (\hat{P}_i - P_i) u_i. \quad (32)$$

The first term is sampling noise; the second is nuisance leakage caused by projector error. If $\lambda_{\min}^+(\hat{A}) \geq m\mu > 0$ and $\|\hat{P}_i - P_i\|_{\text{op}} \leq \epsilon_i$, then

$$\left\| \hat{A}^\dagger \sum_{i=1}^m (\hat{P}_i - P_i) u_i \right\|_2 \leq \frac{1}{m\mu} \sum_{i=1}^m \epsilon_i \|u_i\|_2. \quad (33)$$

This is an analytical diagnostic, not an empirical sensitivity analysis or a statistical guarantee for estimated projectors. Support-perturbation ablations remain future work.

The QLS upper bound (Theorem 3) takes three forms. The full-vector bound on $\mathcal{R} = \text{range}(\mathcal{I}_c)$ yields a direction guarantee for $\theta_{\mathcal{R}} = \Pi_{\mathcal{R}} \beta / \|\Pi_{\mathcal{R}} \beta\|_2$; the original θ is recovered only under $\beta \in \mathcal{R}$. The generic unconstrained tangent variance is controlled by $\text{tr}(P_T(\mathcal{I}_c|_{\mathcal{R}})^\dagger P_T)$. The sharper intrinsic tangent trace $\text{tr}(\mathcal{I}_{c,T}^\dagger)$ holds under empirical decoupling $P_T A_m \theta = 0$, the almost-sure condition $P_T P_S \theta = 0$, or for a constrained local direction estimator; the weaker population condition $P_T \mathcal{I}_c \theta = 0$ alone incurs a Schur-complement inflation in finite samples. The matched-nuisance lower bound (Theorem 4) matches the intrinsic trace on T_{id} , is truncated by r_0 , and is sample-size-independent along $\ker(\mathcal{I}_c) \cap T_0$.

Public-data confounder supports are heuristic. SMS and TruthfulQA show large angle reductions relative to held-out quotient targets, whereas IMDB is neutral. These results evaluate the estimator under the selected supports; they do not show that those supports isolate externally validated semantic ground truth.

QLS does not universally beat DiM on raw log-probability margin: the controlled experiments show this happens because DiM retains a nuisance style component the raw metric rewards. The empirical claim is about deconfounded direction recovery, not behavioral dominance.

The sparse lower bound (Theorem 7) requires dictionary separation or incoherence. Without such geometry, support recovery is ill-posed independently of the quotient model. The Pythia SAE results show that learned dictionaries require validation and tuning before Sparse-Steer dominates dense QLS.

The empirical scope is also narrow: the language-model results are Pythia-410M/TruthfulQA pilots. Validation on larger models, additional steering tasks, and independently grounded semantic targets is needed before claiming broad external generality. These extensions require new experiments and are left for future work.

8 Conclusion

Contrast pairs under polysemanticity are observations of a target direction through random quotient spaces determined by active nuisance spans. This perspective yields an identifiability condition (Theorems 1 and 2), sample-complexity bounds in terms of pseudoinverse traces of $\mathcal{I}_c$ (Theorem 3), a matched-nuisance minimax lower bound (Theorem 4), a Sparse-Steer recovery theorem (Theorem 6), and an explicit phase diagram for coactivated confounders (Corollary 5). Synthetic and controlled experiments support the estimator-level prediction, while public-data and Pythia pilots measure recovery of quotient targets induced by chosen supports. The behavioral results reveal a trade-off: DiM can obtain a higher raw margin by retaining a rewarded nuisance component, whereas QLS can reduce that leakage. The framework is therefore best read as a conditional theory of recovery under credible nuisance supports, with broader semantic and model-level validation left for future work.

References

- Tiago A. Almeida, José María G. Hidalgo, and Akebo Yamakami. 2011. [Contributions to the study of sms spam filtering: new collection and results](#). In *Proceedings of the 11th ACM Symposium on Document Engineering, DocEng '11*, page 259–262, New York, NY, USA. Association for Computing Machinery.

- Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. [LEACE: Perfect linear concept erasure in closed form](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. [Man is to computer programmer as woman is to homemaker? debiasing word embeddings](#). In *NIPS*, pages 4349–4357.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. [Towards monosemanticity: Decomposing language models with dictionary learning](#).
- Peter Buhlmann and Sara van de Geer. 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.
- Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. [Discovering latent knowledge in language models without supervision](#). In *The Eleventh International Conference on Learning Representations*.
- Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. 2024. [Improving steering vectors by targeting sparse autoencoder features](#). *arXiv preprint arXiv:2411.02193*.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, and Chris Olah. 2022. [Toy models of superposition](#).
- Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. [Sparse autoencoders find highly interpretable features in language models](#). In *The Twelfth International Conference on Learning Representations*.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. [Inference-time intervention: Eliciting truthful answers from a language model](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Xiusi Li, Sékou-Oumar Kaba, and Siamak Ravanbakhsh. 2025. [On the identifiability of causal abstractions](#). In *The 28th International Conference on Artificial Intelligence and Statistics*.
- Yuxiao Li, Alina Fastowski, Efstratios Zaradoulas, Bardh Prenkaj, and Gjergji Kasneci. 2026. [Analysing the safety pitfalls of steering vectors](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 11182–11204, San Diego, California, United States. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [Truthfulqa: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. [Learning word vectors for sentiment analysis](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.
- Samuel Marks and Max Tegmark. 2024. [The geometry of truth: Emergent linear structure in large language model representations of true/false datasets](#). In *First Conference on Language Modeling*.
- Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). In *Advances in Neural Information Processing Systems*.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. [Null it out: Guarding protected attributes by iterative nullspace projection](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. [Steering llama 2 via contrastive activation addition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.
- Samuel Soo, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Ming YAN. 2025. [Interpretable steering of large language models with feature guided activation additions](#). In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.
- Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. [Extracting latent steering vectors from pretrained language models](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.
- Joel A. Tropp. 2012. [User-friendly tail bounds for sums of random matrices](#). *Foundations and Trends in Machine Learning*, 5(4):299–340.
- Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. [Activation addition: Steering language models without optimization](#). *arXiv preprint arXiv:2308.10248*.
- Martin J. Wainwright. 2019. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. 2025. [Taxonomy, opportunities, and challenges of representation engineering for large language models](#). *Transactions on Machine Learning Research*. Survey Certification.

Bin Yu. 1997. *Assouad, Fano, and Le Cam*, pages 423–435. Springer New York, New York, NY.

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

A Full Proofs

A.1 Support-Free Impossibility

Proof of Theorem 1. Let $Y_i = \beta + u_i + \xi_i$ with $u_i \in U_{S_i}$ and unrestricted nuisance. We separate the finite-sample and population statements.

Finite-sample absorbable set. Suppose there is a nonzero v such that $v \in U_{S_i}$ for every observed pair $i = 1, \dots, m$. Fix any admissible realization of supports $S_{1:m}$ and nuisances $u_i \in U_{S_i}$. Define

$$\beta' = \beta + v, \quad u'_i = u_i - v. \quad (34)$$

Because $u_i \in U_{S_i}$ and $v \in U_{S_i}$, linearity of U_{S_i} gives $u'_i \in U_{S_i}$. The conditional mean under the reparametrized model is $\beta' + u'_i = \beta + u_i$. The noise law is unchanged, so for every i the conditional distribution of Y_i is identical under (β, u_i) and (β', u'_i) . Since the noise variables are independent under both parameterizations, the joint law of $Y_{1:m}$ is identical. Any estimator that observes only $Y_{1:m}$ receives the same data distribution under β and β' . Thus β is not identifiable along v .

Population unidentifiable subspace. Let $\mathcal{S}_c = \{S : \Pi_c(S) > 0\}$ and $\mathcal{N} = \bigcap_{S \in \mathcal{S}_c} U_S$. A vector v lies in $\mathcal{N}$ iff $v \in U_S$ for every S in the support of Π_c , equivalently $\Pi_{U_S} v = v$ a.s., equivalently $P_S v = 0$ a.s. The finite-sample argument applies to every $v \in \mathcal{N}$ for every m , so the population unidentifiable subspace contains $\mathcal{N}$. Conversely, Theorem 2 shows that any $v \notin \mathcal{N}$ contributes positive Fisher information through $\mathcal{I}_c$ and is therefore identifiable. Hence the population unidentifiable subspace coincides with $\mathcal{N}$. As $m \rightarrow \infty$, the finite-sample absorbable set shrinks to $\mathcal{N}$ almost surely. $\square$

A.2 Identifiability Phase Transition

Proof of Theorem 2. Let $\mathcal{I}_c = \mathbb{E}[P_S]$ with $P_S = I - \Pi_{U_S}$ an orthogonal projector. Since $P_S = P_S^\top = P_S^2$,

$$u^\top P_S u = u^\top P_S^\top P_S u = \|P_S u\|_2^2. \quad (35)$$

Taking expectation gives $u^\top \mathcal{I}_c u = \mathbb{E}\|P_S u\|_2^2$. The random variable $\|P_S u\|_2^2$ is nonnegative, so its expectation is zero iff it is zero a.s. Hence $u^\top \mathcal{I}_c u = 0 \Leftrightarrow P_S u = 0$ a.s.

$P_S u = 0$ means $u \in U_S$ a.s. For any scalar t , define $\beta' = \beta + tu$ and $u'_i = u_i - tu$. Since $tu \in U_{S_i}$ a.s., $u'_i \in U_{S_i}$ and $\beta' + u'_i = \beta + u_i$, so the component of β in direction u is absorbed by nuisance and is not identifiable.

Conversely, multiplying the observation by P_i gives $P_i Y_i = P_i \beta + P_i u_i + P_i \xi_i$. Since $u_i \in U_{S_i}$ and P_i projects onto $U_{S_i}^\perp$, $P_i u_i = 0$. With mean-zero noise, $\mathbb{E}[P_S Y] = \mathcal{I}_c \beta$. If $\mathcal{I}_c$ is invertible on the target subspace, $\beta = \mathcal{I}_c^{-1} \mathbb{E}[P_S Y]$. The population projected moment identifies exactly the components outside $\ker(\mathcal{I}_c)$. $\square$

A.3 QLS Upper Bound

Proof of Theorem 3. Because $P_i u_i = 0$, $P_i Y_i = P_i \beta + P_i \xi_i$. Summing over i gives $r_m = A_m \beta + \zeta_m$ with $\zeta_m = \sum_i P_i \xi_i$. The pseudoinverse $A_m^\dagger$ acts as zero on $\ker(A_m)$, so

$$\tilde{\beta} = A_m^\dagger r_m = \Pi_{\text{range}(A_m)} \beta + A_m^\dagger \zeta_m, \quad (36)$$

and $\tilde{\beta} - \Pi_{\text{range}(A_m)} \beta = A_m^\dagger \zeta_m$. Condition on $P_{1:m}$. By Gaussian linearity, $\zeta_m \mid P_{1:m} \sim N(0, \sigma^2 A_m)$, since

$$\text{Var}(\zeta_m \mid P_{1:m}) = \sum_i P_i \sigma^2 I_d P_i = \sigma^2 A_m. \quad (37)$$

Hence on $\text{range}(A_m)$, $\tilde{\beta} - \Pi_{\text{range}(A_m)} \beta \mid P_{1:m} \sim N(0, \sigma^2 A_m^\dagger)$.

Identifiable-subspace full-vector bound. Restrict to $\mathcal{R} = \text{range}(\mathcal{I}_c)$. We first show $\text{range}(A_m) = \mathcal{R}$ on the Chernoff event. Since $\mathcal{I}_c$ is PSD, $\mathcal{R}^\perp = \ker(\mathcal{I}_c)$. For any $x \in \ker(\mathcal{I}_c)$, Theorem 2 gives $P_S x = 0$ a.s., so every realized P_i annihilates $\mathcal{R}^\perp$. Hence $\text{range}(P_i) \subseteq \mathcal{R}$ for every i and $\text{range}(A_m) \subseteq \mathcal{R}$. The reverse inclusion follows once the Chernoff event gives $A_{m,\mathcal{R}} \succ 0$. Define

$$X_i = (\mathcal{I}_c|_{\mathcal{R}})^{-1/2} (\Pi_{\mathcal{R}} P_i \Pi_{\mathcal{R}}|_{\mathcal{R}}) (\mathcal{I}_c|_{\mathcal{R}})^{-1/2}. \quad (38)$$

Then $\mathbb{E}[X_i] = I_{\mathcal{R}}$ and $\|X_i\|_{\text{op}} \leq \lambda_{\min}(\mathcal{I}_c|\mathcal{R})^{-1}$. Matrix Chernoff gives $\sum_i X_i \succeq (m/2)I_{\mathcal{R}}$ with probability at least $1 - \delta$ whenever $m\lambda_{\min}(\mathcal{I}_c|\mathcal{R}) \gtrsim \log(\dim(\mathcal{R})/\delta)$. Unwhitening yields

$$A_{m,\mathcal{R}} := \Pi_{\mathcal{R}} A_m \Pi_{\mathcal{R}}|_{\mathcal{R}} \succeq \frac{m}{2} \mathcal{I}_c|_{\mathcal{R}}. \quad (39)$$

On this event $\text{range}(A_m) = \mathcal{R}$ and $A_{m,\mathcal{R}}^\dagger \preceq (2/m)(\mathcal{I}_c|_{\mathcal{R}})^\dagger$. Conditional Gaussian norm concentration applied to $\Pi_{\mathcal{R}}(\tilde{\beta} - \beta)$ gives, with conditional probability at least $1 - \delta$,

$$\begin{aligned} & \|\Pi_{\mathcal{R}}(\tilde{\beta} - \beta)\|_2^2 \\ & \leq C\sigma^2 [\text{tr}(A_{m,\mathcal{R}}^\dagger) + \|A_{m,\mathcal{R}}^\dagger\|_{\text{op}} \log(1/\delta)]. \end{aligned} \quad (40)$$

Combining the Gaussian and Chernoff events proves the stated full-vector bound.

Direction normalization. The error is on $\mathcal{R}$. Write $\beta = \Pi_{\mathcal{R}}\beta + \Pi_{\mathcal{R}^\perp}\beta$. Because $\tilde{\beta} \in \text{range}(A_m) \subseteq \mathcal{R}$ on the Chernoff event, $\tilde{\beta} - \Pi_{\mathcal{R}}\beta \in \mathcal{R}$ and $\Pi_{\mathcal{R}^\perp}\beta$ is invisible. Setting $\theta_{\mathcal{R}} = \Pi_{\mathcal{R}}\beta / \|\Pi_{\mathcal{R}}\beta\|_2$ and $b_{\mathcal{R}} = \|\Pi_{\mathcal{R}}\beta\|_2$, if the right side is at most $b_{\mathcal{R}}/2$, then

$$\sin \angle(\hat{\theta}, \theta_{\mathcal{R}}) \leq \frac{2\|\Pi_{\mathcal{R}}(\tilde{\beta} - \beta)\|_2}{b_{\mathcal{R}}}. \quad (41)$$

The angle to θ collapses to that to $\theta_{\mathcal{R}}$ exactly when $\Pi_{\mathcal{R}^\perp}\beta = 0$; otherwise even infinite-sample QLS leaves an irreducible angle $\angle(\theta, \theta_{\mathcal{R}})$.

Generic unconstrained tangent bound. Without decoupling, the tangent error is the Schur block of $A_m^\dagger$ restricted to T . The matrix Chernoff event gives

$$P_T A_m^\dagger P_T \preceq \frac{2}{m} P_T (\mathcal{I}_c|_{\mathcal{R}})^\dagger P_T, \quad (42)$$

and conditional Gaussian norm concentration on $P_T(\tilde{\beta} - \beta)$ yields the displayed Schur-block tangent rate. The relevant population quantity is the Schur block, not $\mathcal{I}_{c,T}^\dagger$.

Sharper tangent bound under decoupling. Under $P_T A_m \theta = 0$, A_m is block diagonal w.r.t. $\text{span}\{\theta\} \oplus T$, so $P_T A_m^\dagger P_T = (P_T A_m P_T|_T)^\dagger$. A sufficient population-level condition is the almost-sure per-sample $P_T P_S \theta = 0$ for $S \sim \Pi_c$; the strictly weaker $P_T \mathcal{I}_c \theta = 0$ only makes $\mathbb{E}[A_m] = m\mathcal{I}_c$ block diagonal, and the actual tangent covariance is the Schur complement

$$P_T A_m^\dagger P_T = (C_m - g_m a_m^{-1} g_m^\top)^{-1}|_T, \quad (43)$$

with $a_m = \theta^\top A_m \theta$, $C_m = P_T A_m P_T|_T$, $g_m = P_T A_m \theta$, dominated by $(1 - \eta_m)^{-1} C_m^{-1}$ when $\eta_m = \|C_m^{-1/2} g_m\|_2^2 / a_m < 1$.

Under empirical decoupling, define $A_{m,T} = P_T A_m P_T|_T = \sum_i P_T P_i P_T|_T$. The matrices $P_T P_i P_T|_T$ are independent PSD contractions with expectation $\mathcal{I}_{c,T}$. Matrix Chernoff gives $A_{m,T} \succeq (m/2)\mathcal{I}_{c,T}$ with probability at least $1 - \delta$, hence $A_{m,T}^\dagger \preceq (2/m)\mathcal{I}_{c,T}^\dagger$ and

$$\text{tr}(A_{m,T}^\dagger) \leq \frac{2}{m} \text{tr}(\mathcal{I}_{c,T}^\dagger), \quad (44)$$

$$\|A_{m,T}^\dagger\|_{\text{op}} \leq \frac{2}{m} \|\mathcal{I}_{c,T}^\dagger\|_{\text{op}}. \quad (45)$$

Conditional Gaussian norm concentration on $P_T(\tilde{\beta} - \beta)$ gives the sharper tangent display. The same bound is attained by a constrained local direction estimator on the unit sphere.

Direction conclusion. Under $\beta \in \mathcal{R}$,

$$\sin \angle(\hat{\theta}, \theta) = \frac{\|P_T(\tilde{\beta} - \beta)\|_2}{\|\beta + (\tilde{\beta} - \beta)\|_2}. \quad (46)$$

On $\|\tilde{\beta} - \beta\|_2 \leq b/2$ the denominator is at least $b/2$, giving the stated direction bound for both tangent variants. $\square$

A.4 Matched-Nuisance Lower Bound

Proof of Theorem 4. Fix two candidate steering vectors β, β' with $\Delta = \beta - \beta'$. Under the first parameter choose $u_i = 0$; under the second choose $u'_i = (I - P_i)\Delta$. Because $I - P_i$ projects onto U_{S_i} , $u'_i \in U_{S_i}$. The conditional means differ by

$$\mu_i - \mu'_i = \beta - \beta' - (I - P_i)\Delta = P_i \Delta. \quad (47)$$

For Gaussian observations with covariance $\sigma^2 I_d$,

$$\text{KL}(P_\beta^{(i)}, P_{\beta'}^{(i)} | S_i) = \frac{1}{2\sigma^2} \|P_i \Delta\|_2^2. \quad (48)$$

Summing and taking expectation gives $\mathbb{E}_S \text{KL} = \frac{m}{2\sigma^2} \Delta^\top \mathcal{I}_c \Delta$. Similarly $\chi^2(P_\beta^m, P_{\beta'}^m | S_{1:m}) = \exp(\sigma^{-2} \sum_i \|P_i \Delta\|_2^2) - 1$.

Diagonalize $G_T = E^\top \mathcal{I}_c E = \sum_{k=1}^r \lambda_k \tilde{v}_k \tilde{v}_k^\top$ with $\lambda_k > 0$, and lift the eigenvectors to $T_{\text{id}} \subset \mathbb{R}^d$ via $v_k = E \tilde{v}_k$. Choose a radius-aware amplitude scale

$$t^2 = \min\{\sigma^2/m, r_0^2/\text{tr}(G_T^{-1})\}, \quad (49)$$

and for $\omega \in \{-1, +1\}^r$ define $z_\omega = \sum_{k=1}^r \omega_k a_k v_k$ with $a_k = \alpha t / \sqrt{\lambda_k}$. By construction

$$\sum_{k=1}^r a_k^2 = \alpha^2 t^2 \text{tr}(G_T^{-1}) \leq \alpha^2 r_0^2, \quad (50)$$

so $\|z_\omega\|_2 \leq \alpha r_0 \leq r_0$ for $\alpha \leq 1$, and $\beta_\omega = \sqrt{b^2 - \|z_\omega\|_2^2} \theta_0 + z_\omega \in \mathcal{B}_{\text{loc}}(\theta_0, r_0)$. Two vertices differing only in coordinate k have first-order tangent difference $2a_k v_k$ with radial correction $O(\alpha^2 r_0^2/b)$. For sufficiently small α , the quotient KL between neighbors is bounded by an absolute constant:

$$\begin{aligned} & \frac{m}{2\sigma^2} (2a_k v_k)^\top \mathcal{I}_c (2a_k v_k) + O(\alpha^4) \\ &= \frac{2\alpha^2 t^2 m}{\sigma^2} + O(\alpha^4) \leq 2\alpha^2 + O(\alpha^4), \end{aligned} \quad (51)$$

using $v_k^\top \mathcal{I}_c v_k = \lambda_k$ and $t^2 \leq \sigma^2/m$. Assouad's lemma gives

$$\begin{aligned} & \inf_{\hat{\theta}} \sup_{\omega} \mathbb{E} \sin^2 \angle(\hat{\theta}, \beta_\omega/b) \\ & \geq \frac{c'}{b^2} \min\{\sigma^2 \text{tr}(G_T^{-1})/m, r_0^2\}, \end{aligned} \quad (52)$$

exactly the displayed lower bound.

Irreducible impossibility along null tangent directions. Let $v \in \ker(\mathcal{I}_c) \cap T_0$ be any unit vector. By Theorem 2, $P_S v = 0$ a.s. under Π_c , so $v \in U_S$ for almost every S . Fix $0 < \tau \leq b/\sqrt{2}$ and set

$$\beta_\pm = \sqrt{b^2 - \tau^2} \theta_0 \pm \tau v, \quad (53)$$

$$u_\pm = \beta_\pm/b = a\theta_0 \pm \tau v, \quad (54)$$

with $a = \sqrt{1 - t^2}$, $t = \tau/b$. Both $\beta_\pm$ have norm b . The matched-nuisance construction with $\Delta = \beta_+ - \beta_- = 2\tau v$ yields $P_i \Delta = 0$ a.s., so the conditional law of the data is identical under either parameter; in particular $\mathbb{E}_S \text{KL}(P_{\beta_+}^m, P_{\beta_-}^m) = 0$ for every m .

For any unit $\hat{u} \in \mathbb{R}^d$, using orthonormality of θ_0 and v ,

$$\sin^2 \angle(\hat{u}, u_\pm) = 1 - (a(\hat{u}^\top \theta_0) \pm t(\hat{u}^\top v))^2. \quad (55)$$

Averaging,

$$\begin{aligned} & \frac{1}{2} [\sin^2 \angle(\hat{u}, u_+) + \sin^2 \angle(\hat{u}, u_-)] \\ &= 1 - a^2 (\hat{u}^\top \theta_0)^2 - t^2 (\hat{u}^\top v)^2. \end{aligned} \quad (56)$$

Since $\theta_0 \perp v$ and $t \leq 1/\sqrt{2}$ gives $a^2 \geq t^2$,

$$a^2 (\hat{u}^\top \theta_0)^2 + t^2 (\hat{u}^\top v)^2 \leq a^2, \quad (57)$$

hence the average is at least $1 - a^2 = t^2 = \tau^2/b^2$. Taking expectation and using $\max \geq \text{average}$ gives $\max_{\pm} \mathbb{E} \sin^2 \angle(\hat{u}, u_\pm) \geq \tau^2/b^2$, for every m and every estimator. $\square$

A.5 Explicit Phase Diagram

Proof of Corollary 5. Suppose $P_S = I - \Pi_U$ w.p. p and $P_S = I$ w.p. $1 - p$. Then $\mathcal{I}_c = p(I - \Pi_U) + (1 - p)I = I - p\Pi_U$. If $v \in U$, $\Pi_U v = v$, so $\mathcal{I}_c v = (1 - p)v$; if $v \in U^\perp$, $\Pi_U v = 0$, so $\mathcal{I}_c v = v$. The spectrum has q eigenvalues $1 - p$ and $d - q$ eigenvalues 1. For $p < 1$,

$$\text{tr}(\mathcal{I}_c^\dagger) = (d - q) + q/(1 - p). \quad (58)$$

As $p \rightarrow 1$, complexity diverges; at $p = 1$ directions inside U lie in $\ker(\mathcal{I}_c)$ and are unidentifiable by Theorem 2. $\square$

A.6 Sparse-Steer Upper Bound

Proof of Theorem 6. The projected observation is $P_i Y_i = P_i B \alpha^* + P_i \xi_i$. By optimality of $\hat{\alpha}$,

$$\begin{aligned} & \frac{1}{2m} \sum_i \|P_i(Y_i - B\hat{\alpha})\|_2^2 + \lambda \|\hat{\alpha}\|_1 \\ & \leq \frac{1}{2m} \sum_i \|P_i(Y_i - B\alpha^*)\|_2^2 + \lambda \|\alpha^*\|_1. \end{aligned} \quad (59)$$

Substitute $\hat{\alpha} = \alpha^* + \Delta$ and cancel the common noise norm:

$$\begin{aligned} & \frac{1}{2m} \sum_i \|P_i B \Delta\|_2^2 \\ & \leq \langle g, \Delta \rangle + \lambda (\|\alpha^*\|_1 - \|\alpha^* + \Delta\|_1), \end{aligned} \quad (60)$$

with $g = m^{-1} \sum_i B^\top P_i \xi_i$. If $\lambda \geq 2\|g\|_\infty$, then $\langle g, \Delta \rangle \leq \frac{\lambda}{2} (\|\Delta_J\|_1 + \|\Delta_{J^c}\|_1)$. By decomposability of ℓ_1 , $\|\alpha^*\|_1 - \|\alpha^* + \Delta\|_1 \leq \|\Delta_J\|_1 - \|\Delta_{J^c}\|_1$. Therefore

$$\frac{1}{2m} \sum_i \|P_i B \Delta\|_2^2 \leq \frac{3\lambda}{2} \|\Delta_J\|_1 - \frac{\lambda}{2} \|\Delta_{J^c}\|_1. \quad (61)$$

Nonnegativity of the left side gives $\|\Delta_{J^c}\|_1 \leq 3\|\Delta_J\|_1$, so Δ lies in the sparse cone, and $\frac{1}{m} \sum_i \|P_i B \Delta\|_2^2 \leq 3\lambda \|\Delta_J\|_1$. Empirical quotient compatibility gives

$$\frac{\Phi_s}{2} \|B \Delta\|_2 \frac{\|\Delta_J\|_1}{\sqrt{s}} \leq 3\lambda \|\Delta_J\|_1, \quad (62)$$

and cancellation yields $\|B \Delta\|_2 \leq 6\lambda \sqrt{s}/\Phi_s$. Since $\hat{\beta} - \beta = B \Delta$, the sparse upper bound follows.

Choice of λ . The score g has coordinate variance

$$\text{Var}(g_j \mid P_{1:m}) = \frac{\sigma^2}{m} \cdot \frac{1}{m} \sum_i \|P_i b_j\|_2^2. \quad (63)$$

Defining $L_{\max} = \max_j (m^{-1} \sum_i \|P_i b_j\|_2^2)^{1/2}$, a Gaussian maximal inequality conditioned on

$P_{1:m}$ gives $\|g\|_\infty \leq C\sigma L_{\max} \sqrt{\log(p/\delta)/m}$ with probability at least $1 - \delta$. Choosing $\lambda \asymp \sigma L_{\max} \sqrt{\log(p/\delta)/m}$ and substituting yields

$$\|\hat{\beta} - \beta\|_2 \lesssim \frac{\sigma L_{\max}}{\Phi_s} \sqrt{s \log(p/\delta)/m}. \quad (64)$$

The $L_{\max}$ -free rate follows by normalizing projected dictionary columns so that $L_{\max} \leq 1$. $\square$

A.7 Sparse Quotient Lower Bound

Proof of Theorem 7. Fix two coefficient vectors θ, θ' and write $\Delta_\beta = B(\theta - \theta')$. Compare the model at θ with $u_i = 0$ to the model at θ' with $u'_i = (I - P_i)\Delta_\beta$. The conditional mean gap is $P_i\Delta_\beta$, so

$$\mathbb{E}_S \text{KL} = \frac{m}{2\sigma^2} (\theta - \theta')^\top B^\top \mathcal{I}_c B (\theta - \theta'). \quad (65)$$

For fixed support J , diagonalize $G_J = B_J^\top \mathcal{I}_c B_J = \sum_{k=1}^s \lambda_k v_k v_k^\top$. Define $\alpha_\omega = a \sum_{k=1}^s \omega_k \lambda_k^{-1/2} v_k$ with $a = \alpha\sigma/\sqrt{m}$ and $a_k = a/\sqrt{\lambda_k}$. Neighbors differing in coordinate k have quotient KL equal to $2\alpha^2$, so Assouad gives

$$\inf_{\hat{\alpha}} \sup_{\omega} \mathbb{E}_\omega \|B_J(\hat{\alpha} - \alpha_\omega)\|_2^2 \geq c \sum_{k=1}^s \|B_J(a_k v_k)\|_2^2. \quad (66)$$

If $B_J^\top B_J \succeq \kappa_J I_s$, then $\|B_J x\|_2^2 \geq \kappa_J \|x\|_2^2$, giving the fixed-support trace term.

For unknown support, take a subset $\mathcal{A} \subset [p]$ of N separated atoms with quotient visibility at most $\lambda_{\max}^{\mathcal{A}}$. A Varshamov–Gilbert construction gives supports $J_1, \dots, J_M$ of size s with $\log M \geq c_0 s \log(N/s)$ and pairwise symmetric difference at least $s/2$. Setting $\theta^{(\ell)} = \tau$ on J_ℓ , separation and incoherence imply pairwise signal separation at least $c\rho^2\tau^2s$, while expected quotient KL is at most $Cm\lambda_{\max}^{\mathcal{A}}\tau^2s/\sigma^2$. Choosing $\tau^2 = c\sigma^2 \log(N/s)/(m\lambda_{\max}^{\mathcal{A}})$ and applying Fano gives

$$\inf_{\hat{\beta}} \sup_{\|\alpha\|_0 \leq s} \mathbb{E} \|\hat{\beta} - B\alpha\|_2^2 \geq c' \frac{\sigma^2 \rho^2}{m\lambda_{\max}^{\mathcal{A}}} s \log(N/s). \quad (67)$$

Direction-risk bounds follow by placing the packing in a local annulus and dividing by $\|\beta\|_2^2$, up to constants and truncation at one. $\square$

A.8 Multi-Environment Deconfounding

Proof of Proposition 8. Suppose $M = \mathbf{1}\beta^\top + \Pi\Pi^\top$. If $w^\top \mathbf{1} = 1$ and $w^\top \Pi = 0$, then $w^\top M = \beta^\top$.

Let $C = \Pi\Pi^\top$ and R_C the orthogonal projector onto $\text{col}(C)^\perp$. The relevant constraint for identifying β is $w^\top C = 0$, i.e., $w \in \text{col}(C)^\perp$, which

together with $w^\top \mathbf{1} = 1$ yields $w^\top M = \beta^\top$. Equivalently $w = R_C w$, so any feasible w satisfies $w^\top \mathbf{1} = w^\top R_C \mathbf{1}$. A feasible w exists iff $R_C \mathbf{1} \neq 0$. (If C is known exactly, β remains identifiable by subtraction $M - C = \mathbf{1}\beta^\top$.)

Under feasibility $R_C \mathbf{1} \neq 0$, the minimum-norm w solves

$$\min_w \|w\|_2^2 \text{ s.t. } w \in \text{col}(C)^\perp, w^\top \mathbf{1} = 1. \quad (68)$$

Inside $\text{col}(C)^\perp$ only $R_C \mathbf{1}$ contributes to the inner product with $\mathbf{1}$, so by Cauchy–Schwarz the minimizer is parallel to $R_C \mathbf{1}$:

$$w^* = \frac{R_C \mathbf{1}}{\|R_C \mathbf{1}\|_2^2}, \quad \|w^*\|_2^2 = \frac{1}{\|R_C \mathbf{1}\|_2^2}. \quad (69)$$

For the dictionary-uniform regime where the constraint is strengthened to $w^\top \Pi = 0$, the same argument with Π in place of C gives $w^* = R_\Pi \mathbf{1}/\|R_\Pi \mathbf{1}\|_2^2$ and feasibility $R_\Pi \mathbf{1} \neq 0$. Since $\text{col}(C) \subseteq \text{col}(\Pi)$, $R_C \mathbf{1} \neq 0$ is implied by $R_\Pi \mathbf{1} \neq 0$ but is strictly weaker whenever $\text{col}(C) \subsetneq \text{col}(\Pi)$.

If each environment mean has independent isotropic variance proportional to σ^2/m , the variance of the weighted estimator is proportional to $\|w^*\|_2^2$. Sample complexity explodes as $\mathbf{1}$ approaches the relevant column span; at the boundary β becomes unidentifiable by environment weighting (though it remains recoverable by subtraction if C is known exactly). $\square$

B Additional Experimental Details

All public-data pilots (Section 6.2) used three seeds: 17, 42, and 137. The synthetic sparse experiment used five seeds: 17, 42, 137, 233, and 404. The Pythia experiments used cached EleutherAI/pythia-410m weights and were run on CUDA. The Pythia hidden-state quotient pilot projected 1024-dimensional final hidden states to 192 dimensions for numerical stability and speed. The functional sweeps (Section 6.3) used held-out validation to choose layer, patch mode, and alpha separately for DiM and QLS before evaluating on test. The counter-confounder representation diagnostic used only train questions to estimate the style direction and steering vectors, and only test questions to form the prefix-balanced operational target.