Liar, Liar: Beyond Vocabulary Suppression

Aryan Gupta
aryan.cs.app@gmail.com

June 2026

Abstract

Representation-engineering interventions reduce deceptive behavior in language models by injecting a contrastively constructed vector into the residual stream. Whether such interventions manipulate an upstream concept or merely tilt the readout against a small lexicon of deception-coded tokens is not resolved by the standard experimental setup. We formalize the question through a token-conditional projection: for a chosen set T of deception-coded tokens, project the steering vector v_{dec} onto the orthogonal complement of the relevant unembedding subspace to obtain $v^{\perp}$ with zero direct logit contribution at every token in T. The fraction of behavioral steering preserved under this projection is a test statistic for the depth of the underlying representation. The construction has a minimum-norm characterization analogous to LEACE and requires the RMSNorm-corrected effective unembedding rather than the raw unembedding matrix; the naive global-orthogonality formulation is impossible whenever the vocabulary exceeds the residual dimension. We position the construction against the Park–Choe–Veitch causal-inner-product duality, the Arditi et al. weight-orthogonalization recipe, the Venkatesh–Kurapath non-identifiability result, and Nadaf's function-vector decoding gap.

1 Introduction

The headline result of the representation-engineering line is that adding a contrastively constructed vector to a transformer's residual stream at a middle layer steers high-level behavior, with reported reductions in deception, sycophancy, and refusal of harmful requests [1, 2, 3, 4]. Two accounts of this result are consistent with the reported numbers. Under the *deep* account, the steering vector moves an upstream representation of the relevant concept and downstream layers consume that representation. Under the *shallow* account, the steering vector tilts the logit head against a small lexicon of behavior-coded tokens, with no upstream content. The two accounts predict different behavior off-distribution and lead to different conclusions about what the model represents.

The standard experimental setup conflates two paths along which the intervention propagates. Adding v_{dec} at layer $\ell^{\star}$ shifts every downstream activation, including the final residual that the unembedding W_U reads to produce logits. The direct logit-attribution path [15] carries the projection of v_{dec} onto the rows of W_U unchanged through the readout. The indirect path routes v_{dec} through every subsequent attention and feed-forward block before reaching the readout. The reported behavioral effect is the sum of the two contributions, and a measurement that does not separate them cannot answer the depth question.

We separate them by construction. Fix a set T of tokens whose probabilities are diagnostic for the behavior under study; the deception case will use a curated list of words such as "lie", "trick", "true", "honest", or a learned logistic direction over those classes. Let $\widetilde{W}_U$ denote the effective unembedding after the final RMSNorm gain has been folded in. Project v_{dec} onto the orthogonal complement of $\operatorname{span}\{\widetilde{W}_U[t,:] : t \in T\}$ in $\mathbb{R}^d$ to obtain $v^{\perp}$, satisfying $\widetilde{W}_U[T,:]\, v^{\perp} = \mathbf{0}$. Injecting

$v^{\perp}$ at layer $\ell^{\star}$ produces a behavioral change that cannot run through direct readout at the tokens in T and must instead propagate through the indirect path. The fraction of v_{dec}'s steering effect preserved under this substitution quantifies how much of the original intervention rode on the direct path.

This document develops the formal apparatus. The naive global formulation is impossible whenever the vocabulary exceeds the residual dimension, forcing the token-conditional reformulation. The RMSNorm gain before the unembedding shifts the relevant subspace by a diagonal scaling and a rank-one correction along the readout point. The projection admits a minimum-norm characterization in the style of LEACE [5], specialized to the token-conditional setting. The companion document `PLAN.md` specifies the experimental program.

The remainder of the document is organized as follows. Section 2 fixes notation and reviews the residual stream, the linear representation hypothesis, the unembedding and its post-norm correction, and the standard contrastive construction of steering vectors. Section 3 proves that the global formulation is impossible when $V > d$, motivating the token-conditional reformulation in Section 4. Section 5 develops the rank-one logit-difference variant and its connection to mass-mean probes. Section 6 formalizes the direct-versus-indirect path decomposition. Section 7 introduces the test statistic and the hypothesis schema it instantiates. Section 8 gives the minimum-norm characterization. Section 9 positions the construction against the closest prior work. Section 10 states what each hypothesis predicts about the test statistic, and Section 11 catalogues the assumptions under which the construction is meaningful.

2 Preliminaries

2.1 Residual stream

Let $\mathcal{M}$ be a decoder-only transformer with L layers, residual stream width d, vocabulary size V. For an input token sequence $x_{1:n}$, the model maintains a residual stream $h_i^{(\ell)} \in \mathbb{R}^d$ at every layer $\ell \in \{0, 1, \ldots, L\}$ and position $i \in \{1, \ldots, n\}$. Each transformer block updates the residual stream additively,

$$h_i^{(\ell+1)} = h_i^{(\ell)} + \text{Attn}^{(\ell)}(h_{1:i}^{(\ell)})_i + \text{MLP}^{(\ell)}(h_i^{(\ell)} + \text{Attn}^{(\ell)}(h_{1:i}^{(\ell)})_i), \tag{1}$$

where the two contributions are the attention block and the feed-forward block writing into the residual stream. The final residual $h_n^{(L)}$ is the input to the output head.

2.2 Output head and the unembedding

In the standard architecture the output head consists of a normalization layer followed by a linear projection to vocabulary space. Modern transformers use root-mean-square normalization [13],

$$\text{RMSNorm}_{\gamma}(z) := \gamma \odot \frac{z}{\sqrt{d^{-1}\|z\|_2^2 + \varepsilon}}, \tag{2}$$

where $\gamma \in \mathbb{R}^d$ is a learned gain and $\varepsilon > 0$ is a stability constant. Older models use layer normalization [14], which subtracts the per-token mean before scaling and then applies an affine. The logits are

$$\boldsymbol{\ell}(h_n^{(L)}) = W_U \cdot \text{RMSNorm}_{\gamma}(h_n^{(L)}) \in \mathbb{R}^V, \tag{3}$$

with $W_U \in \mathbb{R}^{V \times d}$ the unembedding matrix and rows $W_{U t,:} \in \mathbb{R}^d$, one per token. We will treat the RMSNorm case throughout; the LayerNorm case requires the additional handling of the centering null direction, deferred to Section 11.

For our purposes the relevant object is the linear map a small residual perturbation induces on the pre-softmax logits at the readout. Let $z := h_n^{(L)}$ and let $\mathcal{N}_\gamma(z) := \mathrm{RMSNorm}_\gamma(z)$. For a perturbation δz small relative to $\|z\|_2$, a first-order expansion gives

$$\boldsymbol{\ell}(z + \delta z) - \boldsymbol{\ell}(z) = W_U \cdot J_{\mathcal{N}_\gamma}(z) \cdot \delta z + O(\|\delta z\|^2), \tag{4}$$

where $J_{\mathcal{N}_\gamma}(z) \in \mathbb{R}^{d \times d}$ is the Jacobian of RMSNorm at z. A direct computation gives

$$J_{\mathcal{N}_\gamma}(z) = \frac{1}{\sigma(z)} \operatorname{diag}(\gamma) \left(I_d - \frac{z z^\top}{d\, \sigma(z)^2} \right), \tag{5}$$

with $\sigma(z) := \sqrt{d^{-1}\|z\|_2^2 + \varepsilon}$. Equation (5) follows from differentiating Equation (2) term by term, treating $\sigma(z)$ as the scalar function of z defined above.

2.3 Effective unembedding

The composition that controls direct logit contributions of residual perturbations at the readout is

$$\widetilde{W}_U(z) := W_U \cdot J_{\mathcal{N}_\gamma}(z) \in \mathbb{R}^{V \times d}. \tag{6}$$

We refer to $\widetilde{W}_U(z)$ as the *effective unembedding at* z. The rows of $\widetilde{W}_U(z)$ are token-indexed and live in $\mathbb{R}^d$. They differ from the rows of W_U by a diagonal scaling and a rank-one correction along the readout point z. For an intervention applied at a middle layer $\ell^\star \ll L$, the effective unembedding the intervention couples to is evaluated at the readout point in the unperturbed run; we will denote this by $\widetilde{W}_U{}^\star$ and treat it as a fixed matrix when we project. We return to the sense in which this is justified in Section 11.

2.4 Linear representation hypothesis and steering vectors

The linear representation hypothesis, in the form sharpened by Park, Choe, and Veitch [8], posits that for many semantic features W there exist two associated vectors: a context-side representation $\bar{\lambda}_W \in \mathbb{R}^d$ such that counterfactual context shifts move the residual stream along $\bar{\lambda}_W$, and an unembedding-side representation $\bar{\gamma}_W \in \mathbb{R}^d$ such that counterfactual word shifts move the unembedding-direction vector along $\bar{\gamma}_W$. The two are identified by a causal inner product $\langle \bar{\gamma}_W, \cdot \rangle_C = \bar{\lambda}_W^\top$ induced by the covariance of unembedding rows.

We use the contrastive construction for steering vectors throughout. Given a paired dataset $\mathcal{D} = \{(p_i^+, p_i^-)\}_{i=1}^N$ of positively-coded and negatively-coded prompts (for example, "you should respond honestly" versus "you should lie convincingly" as a system prompt prefix), the mean-difference steering vector at layer $\ell^\star$ is

$$v_{\mathrm{dec}} := \frac{1}{N} \sum_{i=1}^{N} \left(h_{n_i}^{(\ell^\star)}(p_i^-) - h_{n_i}^{(\ell^\star)}(p_i^+) \right) \in \mathbb{R}^d. \tag{7}$$

This is the canonical CAA recipe of [2], the difference-of-means variant of the RepE reading vector [1], and the directional construction of the Arditi refusal direction [4].

2.5 Token-coded sets

Let $T \subset \{1, \ldots, V\}$ denote a deception-coded token set, that is, a finite collection of token indices whose probability ratio is diagnostic for the behavior under study. In the deception case, T is the union of a deceptive subset $T^- = \{$"lie", "trick", "deceive", $\ldots\}$ and an honest subset $T^+ = \{$"true", "honest", "frank", $\ldots\}$. The exact construction of T is a methodological choice we return to in Section 11; for now we treat it as given.

3 Impossibility of the Global Construction

The cleanest version of the question one might ask is: is there a nonzero residual perturbation $v \in \mathbb{R}^d$ such that adding v to the residual stream produces zero change in the logits of every token in the vocabulary? In the architectures of practical interest, the answer is no.

Proposition 3.1 (Triviality of the global null space)**.** *Let $W_U \in \mathbb{R}^{V\times d}$ and suppose* $\operatorname{rank}(W_U) = d$*. Then* $\ker(W_U) = \{0\}$*. In particular, if $V \geq d$ and the rows of W_U span $\mathbb{R}^d$, then no nonzero $v \in \mathbb{R}^d$ satisfies $W_U v = 0$.*

Proof. By the rank-nullity theorem applied to W_U viewed as a linear map $\mathbb{R}^d \to \mathbb{R}^V$, we have $\dim\ker(W_U) = d - \operatorname{rank}(W_U) = 0$. □

Remark 3.2 (Why W_U is full rank in practice)**.** For modern transformers V exceeds d by an order of magnitude: Llama-2-7B has $V = 32{,}000$, $d = 4096$; Llama-2-70B has $V = 32{,}000$, $d = 8192$; Mistral-7B has $V = 32{,}000$, $d = 4096$. The unembedding has full column rank with probability one over standard initializations, and this is confirmed by direct inspection of the released checkpoints.

Proposition 3.1 forces the reformulation that follows. If one wants a nontrivial residual direction with zero direct readout effect, one must restrict the readout: drop the requirement that the direction be invisible to every token, and require only that it be invisible to a chosen subset.

4 Token-Conditional Orthogonalization

4.1 The token-conditional subspace

Fix a token-coded set $T \subset \{1, \ldots, V\}$ of size $k := |T|$, with $k < d$. Let $\widetilde{W}_U{}^\star \in \mathbb{R}^{V\times d}$ be the effective unembedding at the unperturbed readout point as defined in Equation (6). Denote by $\widetilde{W}_U{}^\star[T, :] \in \mathbb{R}^{k\times d}$ the submatrix consisting of the rows of $\widetilde{W}_U{}^\star$ indexed by T. Define

$$S_T \;:=\; \operatorname{row}\left(\widetilde{W}_U{}^\star[T, :]\right) \;=\; \operatorname{span}\left\{\widetilde{W}_{U\,t,:}^\star \;:\; t \in T\right\} \;\subseteq\; \mathbb{R}^d, \tag{8}$$

the row span of $\widetilde{W}_U{}^\star[T, :]$. The dimension of S_T is at most k, and is exactly k when the rows of $\widetilde{W}_U{}^\star[T, :]$ are linearly independent. The relevant subspace for our construction is the orthogonal complement

$$S_T^\perp \;:=\; \{v \in \mathbb{R}^d \;:\; \langle v, w\rangle = 0 \text{ for all } w \in S_T\}. \tag{9}$$

The dimension of $S_T^\perp$ is at least $d - k$, and in particular nontrivial whenever $k < d$. This is the room our intervention has to operate in.

4.2 Projection operators

Let $A := \widetilde{W}_U{}^\star[T, :] \in \mathbb{R}^{k\times d}$. Let A^+ denote the Moore–Penrose pseudoinverse of A. Define

$$P_T \;:=\; A^+A \;\in\; \mathbb{R}^{d\times d}, \qquad P_T^\perp \;:=\; I_d - P_T. \tag{10}$$

Lemma 4.1 (Properties of the projectors)**.** *The matrices P_T and $P_T^\perp$ are orthogonal projectors. They satisfy $P_T^2 = P_T$, $P_T^\top = P_T$, $P_T + P_T^\perp = I_d$, and*

$$\operatorname{range}(P_T) = \operatorname{row}(A) = S_T, \qquad \operatorname{range}(P_T^\perp) = \ker(A) = S_T^\perp. \tag{11}$$

Proof. A standard linear-algebra identity for the Moore–Penrose pseudoinverse gives $A^+A = V_r V_r^\top$ where $A = U_r \Sigma_r V_r^\top$ is the compact SVD with $r = \mathrm{rank}(A) \le k$ and $V_r \in \mathbb{R}^{d\times r}$ has orthonormal columns spanning $\mathrm{row}(A)$. Therefore $P_T = V_r V_r^\top$ is the orthogonal projector onto $\mathrm{row}(A) = S_T$. The complement $P_T^\perp = I_d - V_r V_r^\top$ is the orthogonal projector onto $S_T^\perp$. The identities $P_T^2 = P_T$, $P_T^\top = P_T$, and the range identifications follow. □

4.3 The orthogonalized steering vector

Given the steering vector v_{dec} of Equation (7), define its S_T-orthogonal projection by

$$v^\perp \;:=\; P_T^\perp\, v_{\mathrm{dec}} \;=\; v_{\mathrm{dec}} - A^+ A v_{\mathrm{dec}}. \tag{12}$$

Proposition 4.2 (Direct logit contribution at T vanishes)**.** *Let $v^\perp$ be defined by Equation (12). Then for every $t \in T$,*

$$\big(\widetilde{W}_U{}^\star v^\perp\big)_t \;=\; 0. \tag{13}$$

Proof. By construction, $v^\perp \in S_T^\perp = \ker(A)$, so $Av^\perp = \mathbf{0}$. The t-th coordinate of $\widetilde{W}_U{}^\star v^\perp$ for $t \in T$ is exactly $Av^\perp$ in component t, which is zero. □

The companion fact concerns the residual on the complement of T.

Proposition 4.3 (Direct logit contribution off T)**.** *For any $t \notin T$, the direct logit contribution of $v^\perp$ to token t is*

$$\big(\widetilde{W}_U{}^\star v^\perp\big)_t \;=\; \widetilde{W}_U{}^\star_{t,:} \cdot P_T^\perp\, v_{dec}. \tag{14}$$

This is in general nonzero, and equals the inner product of $\widetilde{W}_U{}^\star_{t,:}$ with the component of v_{dec} orthogonal to S_T.

The proof is immediate. The substantive point is that the construction zeroes the direct contribution exactly on T and does not constrain the contribution on the complement. We comment on the implications for evaluation in Section 11.

4.4 Parallel and orthogonal decomposition

The two pieces of v_{dec} are

$$v^\parallel \;:=\; P_T\, v_{\mathrm{dec}}, \qquad v^\perp \;=\; v_{\mathrm{dec}} - v^\parallel. \tag{15}$$

Both lie in $\mathbb{R}^d$, both have norms at most $\|v_{\mathrm{dec}}\|$, and they are mutually orthogonal: $\langle v^\parallel, v^\perp\rangle = 0$ by Lemma 4.1. The Pythagorean identity

$$\|v_{\mathrm{dec}}\|_2^2 \;=\; \|v^\parallel\|_2^2 + \|v^\perp\|_2^2 \tag{16}$$

quantifies how much of v_{dec} lies in the readout-aligned subspace. The ratio $\|v^\perp\|^2/\|v_{\mathrm{dec}}\|^2$ summarizes the projection geometry but does not in general track behavioral effect, and is not the test statistic of Section 7.

5 The Rank-One Variant: Logit-Difference Direction

A natural lower-dimensional choice for the projected-out subspace is a single direction: the one that, in the readout, separates T^+ from T^-. This is the direction along which raw probability mass moves from deceptive-coded tokens to honest-coded tokens.

Definition 5.1 (Logit-difference direction)**.** Let $T^+ \cup T^- = T$ be a partition of T into honest-coded and deceptive-coded subsets. The *logit-difference direction* is

$$d_{HD} \;:=\; \frac{1}{|T^+|} \sum_{t \in T^+} \widetilde{W}_{U^\star t,:} \;-\; \frac{1}{|T^-|} \sum_{t \in T^-} \widetilde{W}_{U^\star t,:} \;\in\; \mathbb{R}^d. \tag{17}$$

The vector d_{HD} points in the direction such that increasing the inner product $\langle h_n^{(L)}, d_{HD} \rangle$ increases the average logit on T^+ relative to the average logit on T^-. This is the readout-side analogue of the mass-mean honesty probe used by Marks and Tegmark [9].

Construction 5.2 (Rank-one orthogonalization)**.** Let $\hat{d}_{HD} := d_{HD}/\|d_{HD}\|_2$. Define

$$v_{HD}^{\perp} \;:=\; v_{\mathrm{dec}} \;-\; \langle \hat{d}_{HD}, v_{\mathrm{dec}} \rangle \, \hat{d}_{HD}. \tag{18}$$

Proposition 5.3 (Rank-one direct effect identity)**.** *For $v_{HD}^{\perp}$ as defined in Equation (18),*

$$\frac{1}{|T^+|} \sum_{t \in T^+} (\widetilde{W}_U{}^\star v_{HD}^{\perp})_t \;-\; \frac{1}{|T^-|} \sum_{t \in T^-} (\widetilde{W}_U{}^\star v_{HD}^{\perp})_t \;=\; 0. \tag{19}$$

Proof. The left-hand side equals $\langle d_{HD}, v_{HD}^{\perp} \rangle$ by linearity. By construction, $v_{HD}^{\perp}$ is the orthogonal projection of v_{dec} onto $\mathrm{span}\{\hat{d}_{HD}\}^{\perp} = \mathrm{span}\{d_{HD}\}^{\perp}$, so $\langle d_{HD}, v_{HD}^{\perp} \rangle = 0$. □

The rank-one variant trades zero direct effect on individual tokens in T for zero direct effect on the mean logit difference between T^+ and T^-. It leaves $d-1$ residual dimensions free, the maximal complement available to a nontrivial readout-orthogonal construction.

Remark 5.4 (Relation to Park–Choe–Veitch)**.** Under the duality of [8], d_{HD} is the unembedding-side representation $\bar{\gamma}_{HD}$ of the honesty-versus-deception concept $W = HD$. The construction in Construction 5.2 projects the steering vector orthogonal to this unembedding-side representation. The dual statement is that the steering vector is forced to live in the kernel of the Riesz map $v \mapsto \langle \bar{\gamma}_{HD}, v \rangle_C$, modulo the difference between the standard inner product on $\mathbb{R}^d$ and the causal inner product. We make the comparison precise in Section 9.

6 The Direct-Versus-Indirect Path Decomposition

6.1 Path setup

Let $\ell^\star$ be the intervention layer. Without intervention the model evolves the residual stream by the recurrence Equation (1) from $h_n^{(\ell^\star)}$ to $h_n^{(L)}$. With intervention, the model is run identically except that the residual at layer $\ell^\star$ at position n is replaced by $h_n^{(\ell^\star)} + v$ for a chosen $v \in \mathbb{R}^d$. Let $F^{(\ell^\star \to L)} : \mathbb{R}^d \to \mathbb{R}^d$ denote the residual-to-residual map from layer $\ell^\star$ to layer L at the final position, treated as a function of the perturbation at position n holding earlier positions fixed.

For a small perturbation v, a first-order expansion gives

$$h_n^{(L)}(v) - h_n^{(L)}(0) \;=\; \left(I_d + M^{(\ell^\star \to L)}\right) v \;+\; O(\|v\|^2), \tag{20}$$

where $M^{(\ell^\star \to L)} \in \mathbb{R}^{d \times d}$ is the sum of Jacobians of the attention and MLP blocks composed along layers $\ell^\star, \ell^\star + 1, \ldots, L-1$, evaluated at the unperturbed activations. The identity term I_d in Equation (20) is the direct residual path: the perturbation passes through subsequent residual additions unchanged. The matrix $M^{(\ell^\star \to L)}$ is the indirect path: every downstream attention head and feed-forward block reads the perturbed residual and writes a contribution back into the stream.

6.2 Direct and indirect logit contributions

Composing Equation (20) with the readout linearization Equation (4),

$$\begin{aligned}\boldsymbol{\ell}(h_n^{(L)}(v)) - \boldsymbol{\ell}(h_n^{(L)}(0)) &= \widetilde{W}_U{}^{\star}\left(I_d + M^{(\ell^\star \to L)}\right) v \;+\; O(\|v\|^2) \\ &= \underbrace{\widetilde{W}_U{}^{\star} v}_{\text{direct path}} \;+\; \underbrace{\widetilde{W}_U{}^{\star} M^{(\ell^\star \to L)} v}_{\text{indirect path}} + O(\|v\|^2). \end{aligned} \tag{21}$$

The direct path is what the logit lens [16] reads at the intervened layer if we read it through the same final RMSNorm. The indirect path is everything else: every attention head and feed-forward block downstream of $\ell^\star$ that reads the perturbed residual and writes a contribution back into the stream.

6.3 The projection isolates the indirect path on T

Apply the decomposition to $v = v^\perp$ and restrict to the rows indexed by T.

Theorem 6.1 (Direct/indirect decomposition at T)**.** *Let $v^\perp = P_T^\perp v_{dec}$ be the projected steering vector from Equation (12), and let $\boldsymbol{\ell}_T(\cdot)$ denote the subvector of logits indexed by T. Then*

$$\boldsymbol{\ell}_T(h_n^{(L)}(v^\perp)) - \boldsymbol{\ell}_T(h_n^{(L)}(0)) \;=\; (\widetilde{W}_U{}^{\star} M^{(\ell^\star \to L)} v^\perp)_T \;+\; O(\|v^\perp\|^2). \tag{22}$$

At first order, the change in T-indexed logits produced by injecting $v^\perp$ runs entirely through the indirect path.

Proof. Apply Equation (21) with $v = v^\perp$ and take the T-indexed subvector. The direct term $(\widetilde{W}_U{}^{\star} v^\perp)_T$ vanishes by Proposition 4.2. The remainder is the indirect term plus the second-order error. □

The companion identity for the original v_{dec} is

$$\boldsymbol{\ell}_T(h_n^{(L)}(v_{\text{dec}})) - \boldsymbol{\ell}_T(h_n^{(L)}(0)) \;=\; (\widetilde{W}_U{}^{\star} v_{\text{dec}})_T \;+\; (\widetilde{W}_U{}^{\star} M^{(\ell^\star \to L)} v_{\text{dec}})_T \;+\; O(\|v_{\text{dec}}\|^2), \tag{23}$$

which separates the two contributions explicitly. Subtracting Equation (22) from Equation (23) gives the change attributable to the projected-out part of the steering vector,

$$(\widetilde{W}_U{}^{\star} v_{\text{dec}})_T \;+\; (\widetilde{W}_U{}^{\star} M^{(\ell^\star \to L)} v^\parallel)_T \;+\; O(\|v_{\text{dec}}\|^2), \tag{24}$$

which combines the direct readout of the full v_{dec} with the indirect readout of its S_T-parallel component. Equations (22) to (24) together constitute the formal content of the experiment.

7 Test Statistic and Hypothesis Schema

The construction above gives a clean test only when paired with a way to summarize the behavioral effect. We use two summaries: a benchmark-level effect on a deception evaluation, and a per-sample logit-shift on the diagnostic token set. Both fit a single hypothesis schema.

7.1 Benchmark-level effect

Let B be a deception benchmark (Section 9 reviews the candidates). For an intervention $v \in \mathbb{R}^d$, let $\beta(v) := \text{score}_B(\mathcal{M}_v)$ denote the score on B when $\mathcal{M}$ is run with v added at layer $\ell^\star$ at every token position after the prompt, following the CAA inference convention [2]. The baseline is $\beta(0)$. The effect of an intervention is $\Delta(v) := \beta(v) - \beta(0)$.

Definition 7.1 (Depth-of-representation statistic)**.** The depth-of-representation statistic for the steering vector v_{dec} relative to the token set T is

$$\rho(v_{\text{dec}}, T) \;:=\; \frac{\Delta(v^\perp)}{\Delta(v_{\text{dec}})} \;=\; \frac{\beta(v^\perp) - \beta(0)}{\beta(v_{\text{dec}}) - \beta(0)}, \tag{25}$$

with $v^\perp = P_T^\perp v_{\text{dec}}$ as in Equation (12). The statistic is well-defined whenever $\Delta(v_{\text{dec}}) \neq 0$.

7.2 Per-sample logit-shift

For a deception-coded prompt p, let $\mu^+(v; p)$ and $\mu^-(v; p)$ denote the average log-probabilities the model assigns to tokens in T^+ and in T^- respectively at the relevant continuation position when run with intervention v. The honest-shift is

$$\eta(v; p) \;:=\; \big(\mu^+(v; p) - \mu^-(v; p)\big) \;-\; \big(\mu^+(0; p) - \mu^-(0; p)\big), \tag{26}$$

and the analogous depth statistic is

$$\rho_\eta(v_{\text{dec}}, T; p) \;:=\; \frac{\eta(v^\perp; p)}{\eta(v_{\text{dec}}; p)}. \tag{27}$$

7.3 Hypothesis schema

We state the schema as a pair of empirical hypotheses about the value of ρ across benchmarks and prompt distributions.

Hypothesis 7.2 (Shallow null)**.** The behavioral effect of v_{dec} is mediated entirely by the direct logit-attribution path on a small token set. In this case the construction predicts $\rho(v_{\text{dec}}, T) \approx 0$ for any T that exhausts the relevant readout-aligned mass, and $\rho_\eta(v_{\text{dec}}, T; p) \approx 0$ for the same T.

Hypothesis 7.3 (Deep alternative)**.** The behavioral effect of v_{dec} is mediated substantially by the indirect path through downstream attention and feed-forward layers. In this case the construction predicts $\rho(v_{\text{dec}}, T) \not\approx 0$, with ρ approaching 1 in the limit that the indirect path carries all of the behavioral effect.

The two hypotheses are not exhaustive; in practice one expects $\rho \in (0, 1)$, with the value an empirical statement about the proportion of the steering effect that lies in the readout-aligned subspace. The schema admits an informative interpretation under any empirical outcome. A small ρ across token sets, models, and benchmarks is evidence for the shallow account; a large ρ is evidence for the deep account; an intermediate ρ is a quantitative decomposition of the two contributions.

7.4 Robustness statistic

A subsidiary statistic measures how stable ρ is across choices of T. For a family of token sets $\{T_j\}_{j=1}^{J}$ (Section 11 gives a candidate family), the cross-set stability is

$$\sigma_T(v_{\text{dec}}) \;:=\; \text{Var}_j\left[\rho(v_{\text{dec}}, T_j)\right]^{1/2}. \tag{28}$$

A small σ_T is evidence that the depth statistic is a property of the steering vector and not an artefact of the token-set choice. We will report both ρ and σ_T in the experimental program.

8 Minimum-Norm Characterization

The projection in Equation (12) is a specific choice of intervention on v_{dec}. Among all interventions that satisfy the zero-direct-effect constraint on T, it is the one that perturbs v_{dec} minimally in the Euclidean sense. This is the natural analogue of the LEACE characterization [5], restricted to the token-conditional setting.

Theorem 8.1 (Minimum-norm projection)**.** *Let $A := \widetilde{W}_U{}^{\star}[T,:]$ and let $v_{dec} \in \mathbb{R}^d$ be given. Consider the optimization problem*

$$\min_{u\in\mathbb{R}^d} \;\| u - v_{dec} \|_2 \qquad \textit{subject to} \qquad Au = \mathbf{0}. \tag{29}$$

The unique solution is $u^{\star} = P_T^{\perp} v_{dec} = v_{dec} - A^{+}Av_{dec}$.

Proof. Write any feasible $u = v_{\text{dec}} - \Delta$ with $\Delta \in \mathbb{R}^d$. The constraint $Au = 0$ rewrites as $A\Delta = Av_{\text{dec}}$. Among all Δ satisfying $A\Delta = Av_{\text{dec}}$, the minimum-norm solution is the minimum-norm least-squares solution to this linear system, which by the standard pseudoinverse identity is $\Delta^{\star} = A^{+}Av_{\text{dec}}$, see [17, Sec. 5.5]. Substituting gives $u^{\star} = v_{\text{dec}} - A^{+}Av_{\text{dec}} = P_T^{\perp} v_{\text{dec}}$, and this is the unique solution because the constraint set is closed and convex and the objective is strictly convex. □

Corollary 8.2 (Mahalanobis-weighted variant)**.** *Let $\Sigma \succ 0$ be a positive-definite weighting matrix and replace the objective in Theorem 8.1 by the Mahalanobis distance $\|u - v_{dec}\|_{\Sigma} := \sqrt{(u - v_{dec})^{\top}\Sigma^{-1}(u - v_{dec})}$. The unique minimizer is*

$$u_{\Sigma}^{\star} \;=\; v_{dec} \;-\; \Sigma(A\Sigma)^{+}Av_{dec}. \tag{30}$$

Proof. Change variables to $w := \Sigma^{-1/2}(u - v_{\text{dec}})$. The objective becomes $\|w\|_2$ and the constraint becomes $A\Sigma^{1/2}w = -Av_{\text{dec}}$, namely $\widetilde{A}w = -Av_{\text{dec}}$ with $\widetilde{A} := A\Sigma^{1/2}$. The minimum-norm solution in w is $w^{\star} = -\widetilde{A}^{+}Av_{\text{dec}}$. Substituting back, $u_{\Sigma}^{\star} = v_{\text{dec}} + \Sigma^{1/2}w^{\star} = v_{\text{dec}} - \Sigma^{1/2}(A\Sigma^{1/2})^{+}Av_{\text{dec}}$. Using the identity $\Sigma^{1/2}(A\Sigma^{1/2})^{+} = \Sigma(A\Sigma)^{+}$ ([18, Cor. 1.4.2]) gives the stated form. □

The natural Σ for our setting is the covariance of residual-stream activations at layer $\ell^{\star}$ on a held-out distribution. This connects the construction to LEACE [5], in which the relevant covariance is computed on the same residual stream that the projection is applied to. We use the unweighted version for the main analysis and treat the Mahalanobis variant as a robustness check.

Remark 8.3 (LEACE comparison)**.** LEACE projects out the column space of the whitened cross-covariance $W\Sigma_{XZ}$, where Z is an externally provided concept label and X is the residual activation. The minimum-norm projection in the LEACE Mahalanobis metric is the unique affine map that makes a linear classifier of Z given X impossible. Our construction is structurally identical, with two differences: the projected-out subspace is read directly off the model's effective unembedding rather than learned from labeled data, and the constraint is zero direct logit contribution on T rather than zero linear classifier accuracy on a label. Theorem 8.1 is the LEACE theorem specialized to a T-indexed concept defined by the unembedding.

9 Relation to Prior Work

We position the construction relative to the closest precedents. Table 1 summarizes the comparison; the text expands the rows that are load-bearing for the novelty claim.

Work	Projects out	Asks
LEACE [5]	Whitened cross-covariance of label Z on X	Linear Z-decoding impossible?
INLP [6]	Iteratively trained classifier rows	Same
R-LACE [7]	Rank-k adversarially fitted subspace	Same
Arditi et al. [4]	A learned refusal direction $\hat{r}$ everywhere	Behavior absent if direction unwritable?
Park–Choe–Veitch [8]	N/A (duality theorem)	Are concept directions and steering directions dual?
Marks–Tegmark [9]	N/A (causal patch)	Does mass-mean truth direction flip outputs?
Venkatesh–Kurapath [10]	Jacobian-null perturbations	Are steering vectors identifiable?
Nadaf [11]	N/A (logit-lens decoding gap)	Do function vectors steer beyond the logit lens?
hughvd [12]	Token unembeddings W_U rows on a token set	Worked example for sentiment on Gemma-2-9b
This work	RMSNorm-corrected $\widetilde{W_U}^\star$ rows on T	Honesty/deception case on modern benchmarks

Table 1: Comparison of orthogonal-projection-based interventions and the subspace each projects out.

9.1 Arditi et al. weight orthogonalization

The closest methodological precedent is the weight-orthogonalization recipe of Arditi et al. [4]. They compute the refusal direction $\hat{r}$ by a mean-difference construction on harmful versus harmless instructions and modify every matrix that writes to the residual stream by the rank-one projection

$$W_{\text{out}} \;\mapsto\; W_{\text{out}} \;-\; \hat{r}\hat{r}^\top W_{\text{out}}. \tag{31}$$

The construction in Construction 5.2 is the dual: project the steering vector v_{dec} orthogonal to a readout-derived direction $\hat{d}_{HD}$ rather than projecting all write-side matrices orthogonal to a behavior-derived direction $\hat{r}$. Both rely on the same rank-one projector $I - \hat{u}\hat{u}^\top$; Arditi removes a direction from every matrix that writes to the residual stream, while we remove a direction from the steering vector itself.

9.2 Park–Choe–Veitch duality

In the formal apparatus of [8] the unembedding side of a concept W is a direction $\bar{\gamma}_W \in \mathbb{R}^d$. The construction in Construction 5.2, with T^+, T^- aligned to the concept's positive and negative tokens, projects v_{dec} orthogonal to $\bar{\gamma}_{HD}$ in the standard inner product. Under the causal inner product $\langle \cdot, \cdot \rangle_C$ induced by the unembedding covariance, the dual identification $\bar{\lambda}_{HD} = \Sigma_\gamma^{-1} \bar{\gamma}_{HD}$ holds. Our rank-one construction is the standard-inner-product version. The Mahalanobis variant in Corollary 8.2 with $\Sigma = \Sigma_\gamma$ is the causal-inner-product version. We will treat both in the experiments.

9.3 Venkatesh–Kurapath non-identifiability

[10] show that steering vectors are not uniquely identified: large equivalence classes of behaviorally indistinguishable interventions exist, parametrized by perturbations in the null space of the activation-to-logit Jacobian. The relevant subspace for their non-identifiability is $\ker(J)$ where $J = W_U \cdot J_{\mathcal{N}_\gamma}(\cdot) \cdot M^{(\ell^\star \to L)}$; ours is $\ker(\widetilde{W}_U{}^\star[T,:])$. The two address different questions. Theirs establishes that $v^\perp$ is in general not behaviorally equivalent to v_{dec}; ours measures how much of the discrepancy concentrates in the readout-aligned subspace as a depth statistic.

9.4 Nadaf function vectors

[11] demonstrate that function vectors can steer model behavior in cases where the logit lens cannot decode the steered output at any intermediate layer. That is direct evidence of an off-readout steering channel for the in-context-learning function-vector setting. Our construction asks whether the same gap exists for honesty-and-deception steering vectors specifically, on modern deception benchmarks, with a quantitative measurement of the gap.

9.5 Prior unembedding-orthogonal benchmark

The hughvd unembedding-steering-benchmark repository [12] proposes a related test on Gemma-2-9b with sentiment as the worked example, and develops the orthogonalized-steering construction and the evaluation harness. The construction has not been applied to honesty and deception steering with quantitative measurement on the modern deception benchmarks.

10 Predicted Outcomes Under Each Hypothesis

The experiment instantiated by Definition 7.1 admits two distinct predictions, each derivable from one of the hypotheses in Section 7.

10.1 Predictions under the shallow null

If the behavioral effect of v_{dec} is driven primarily by direct logit contribution at deception-coded tokens, then $\Delta(v^\perp)$ should be near zero whenever T exhausts the relevant token set, that is, whenever the projected-out subspace covers the readout-aligned mass of v_{dec}. The expected pattern is the following.

- $\rho(v_{\text{dec}}, T)$ small and decreases with $|T|$. As more deception-coded tokens are included in T, more of the direct path is removed and the surviving effect shrinks.
- $\rho_\eta(v_{\text{dec}}, T; p)$ near zero on prompts whose continuation uses tokens in T.
- Sharp degradation in out-of-distribution settings. If the steering vector encodes only a token-distribution shift, it should fail to generalize to languages, registers, or scenarios where the deception-coded vocabulary differs from T.
- Increasing the steering magnitude α should rescale $\Delta(v_{\text{dec}})$ but not change ρ, because the direct path is linear in α to first order and the projection commutes with scalar multiplication.

10.2 Predictions under the deep alternative

If the behavioral effect of v_{dec} rides on the indirect path through downstream layers, ρ should be substantial and stable.

- $\rho(v_{\text{dec}}, T) \in [0.5, 1]$, with stability across T choices. If the indirect path carries the bulk of the effect, projecting out the readout-aligned subspace removes only a small fraction.
- $\rho_\eta(v_{\text{dec}}, T; p)$ substantial on prompts even when the continuation uses tokens in T.
- Robust generalization. The indirect path operates on the model's downstream computations rather than on its raw vocabulary mass, so an effect that survives projection should also survive moderate distribution shift, including paraphrase and translation of the prompt.
- Stability under varying α within a first-order regime.

10.3 Intermediate outcomes

The expected real-world outcome is intermediate. The interesting empirical questions are then quantitative.

- *What is the typical value of ρ across models and benchmarks?* A consistent value of ρ around 0.5 across Llama-2-7B, Llama-2-13B, Mistral-7B, and Gemma-2-9B is a meaningful summary of the field.
- *Does ρ correlate with capability?* A monotonic relationship between ρ and model size or post-training quality is predicted by accounts under which larger models develop deeper conceptual structure.
- *Does ρ depend on the form of the steering vector?* If ρ for the LAT vector [1] differs systematically from ρ for the CAA vector [2], this is informative about which construction captures which structure.
- *Does ρ track generalization to MASK and Liars' Bench?* If steering with $v^\perp$ retains a substantial fraction of the original effect on MASK [19] and Liars' Bench [20], those interventions are mediated by something other than vocabulary mass.

11 Limitations

11.1 Choice of the token-coded set T

The selection of T is a methodological choice with non-trivial effect. Three options span the design space.

- *Curated lists.* A short manually compiled list of honest- and deceptive-coded words, optionally extended by morphological variants. Advantage: interpretability. Disadvantage: sensitive to authorial bias and to lexical idiosyncrasy.
- *Statistical extraction.* Take the top-k tokens by mean log-probability ratio between deceptive and honest contexts on a held-out corpus. Advantage: replicable. Disadvantage: T depends on the corpus.

- *Probe-derived.* Train a logistic regression to predict the honesty label from the unembedding direction and take the top-k tokens by classifier-coefficient magnitude. Advantage: directly targets the readout-aligned linear structure. Disadvantage: depends on probe regularization.

We report results across all three constructions and use the cross-set stability σ_T as a robustness measure. A finding that depends on a single T choice is weak.

11.2 RMSNorm linearization

The construction uses the effective unembedding $\widetilde{W}_U{}^\star = W_U \cdot J_{\mathcal{N}_\gamma}(z^\star)$ evaluated at the unperturbed readout point $z^\star$. The Jacobian is exact only in the small-perturbation limit; for perturbations of magnitude $\alpha \cdot \|v_{\text{dec}}\|$ comparable to $\|z^\star\|$, the higher-order RMSNorm nonlinearity introduces an error term. Empirically the regime of interest is $\alpha = O(1)$ with $\|v_{\text{dec}}\| \ll \|z^\star\|$, but we report the deviation between the predicted direct contribution and the measured one as a check.

11.3 LayerNorm and the centering null

For models using full LayerNorm rather than RMSNorm, the relevant subspace is shifted by the centering operation. The full LayerNorm Jacobian at z is

$$J_{\text{LN},\gamma}(z) \;=\; \frac{1}{\sigma(z')}\,\text{diag}(\gamma)\left(I_d - \frac{1}{d}\mathbf{1}\mathbf{1}^\top - \frac{z'(z')^\top}{d\,\sigma(z')^2}\right), \tag{32}$$

with $z' := z - d^{-1}(\mathbf{1}^\top z)\mathbf{1}$ the centered residual. The centering operator $I_d - d^{-1}\mathbf{1}\mathbf{1}^\top$ has a one-dimensional kernel along $\mathbf{1}$, so the all-ones direction is in the kernel of $J_{\text{LN},\gamma}(z)$ for every z. This is a real but small effect for our construction; the projection $P_T^\perp$ remains well-defined and the only practical change is to project against $W_U \cdot J_{\text{LN},\gamma}(z^\star)[T,:]$.

11.4 Single-layer assumption

The construction intervenes at a single layer $\ell^\star$. In practice, RepE methods sometimes intervene at several layers simultaneously [1]. The decomposition in Equation (21) generalizes by replacing $M^{(\ell^\star \to L)}$ with a sum over intervention layers, but the projection then must be against the effective unembedding at every layer simultaneously. We treat the single-layer version as primary and report multi-layer results as a robustness check.

11.5 Behavioral effect on the complement of T

By Proposition 4.3, the projected vector $v^\perp$ can shift logits on tokens not in T. If the benchmark B tests behaviors with output vocabulary that overlaps neither T^+ nor T^-, then a residual direct contribution from $v^\perp$ on those tokens may explain the surviving effect. We address this by reporting ρ both on benchmarks whose outputs lie inside T (where the direct contribution is zero by construction) and on benchmarks whose outputs lie outside T (where a residual direct contribution exists), and we report the off-T direct contribution magnitude as a side measurement.

11.6 Semantic scope of the test statistic

The statistic ρ measures the fraction of v_{dec}'s behavioral effect surviving token-conditional readout suppression. Interpretive claims about the upstream representation require independent evidence: out-of-distribution generalization, transfer across model scales, and identification of the downstream circuits that consume the indirect-path contribution.

12 Conclusion

We have developed a mathematical apparatus for separating the direct-readout and indirect-propagation contributions of representation-engineering steering vectors. The construction is a token-conditional orthogonal projection against the rows of the effective unembedding restricted to a chosen deception-coded token set. The projection is well-defined, admits a minimum-norm characterization in the style of LEACE, and the resulting test statistic ρ quantifies the proportion of the steering effect carried by each path. The companion document specifies the experimental program: model checkpoints, benchmark suite, token-set constructions, hypothesis tests, and OOD probes. The expected empirical regime is intermediate ρ; the primary questions are the value of ρ across models, vectors, and benchmarks, and whether ρ tracks capability and generalization.

References

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

[2] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

[3] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023.

[4] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024.

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

[6] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of ACL*, 2020.

[7] Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure. In *Proceedings of ICML*, 2022.

[8] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of ICML*, 2024.

[9] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv:2310.06824*, 2023.

[10] Sohan Venkatesh and Ashish Mahendran Kurapath. On the non-identifiability of steering vectors in large language models. *arXiv:2602.06801*, 2026.

[11] Mohammed Suhail B Nadaf. Steerable but not decodable: Function vectors operate beyond the logit lens. *arXiv:2604.02608*, 2026.

[12] Hugh van Deventer. Unembedding-steering benchmark. `https://github.com/hughvd/unembedding-steering-benchmark`, 2024.

[13] Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, 2019.

[14] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv:1607.06450*, 2016.

[15] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Anthropic Technical Report*, 2021.

[16] nostalgebraist. Interpreting GPT: The logit lens. *LessWrong*, 2020.

[17] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, second edition, 2012.

[18] Adi Ben-Israel and Thomas N. E. Greville. *Generalized Inverses: Theory and Applications*. Springer, second edition, 2003.

[19] Richard Ren, Arunim Agarwal, Mantas Mazeika, Cristina Menghini, Robert Vacareanu, Brad Kenstler, Mick Yang, Isabelle Barrass, Alice Gatti, Xuwang Yin, Eduardo Trevino, Matias Geralnik, Adam Khoja, Dean Lee, Summer Yue, and Dan Hendrycks. The MASK benchmark: Disentangling honesty from accuracy in AI systems. *arXiv:2503.03750*, 2025.

[20] Kieron Kretschmar, Walter Laurito, Sharan Maiya, and Samuel Marks. Liars' bench: Evaluating lie detectors for language models. *arXiv:2511.16035*, 2025.