

A Control Knob Is Not a Concept: What a Working Steering Vector Licenses, Why Its Equivalence Class Is Set by the Readout That Scores It, and the Controls That Would Upgrade Control to Representation

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Adding a vector to a language model's residual stream and watching a behaviour change has become the field's most common argument that the model represents a concept along that vector. Refusal directions, persona vectors, truth directions and instruction vectors are all reported this way. A second body of work, mostly from 2025 and 2026, finds that geometrically unrelated vectors produce the same behavioural shift, that refusal is carried by several directions and multi-dimensional cones rather than one, that random directions of activation-scale norm can break safety behaviour, and that the direction which best detects a property is often a poor direction for controlling it. In this analysis the author reviews and synthesises published results to show that steering's defenders and critics have measured different things. A steering result is an estimate indexed by six choices - the contrast set, the extraction rule, the injection site, the magnitude, the readout that scores the behaviour, and the rule that selected the vector from its candidates - and it is asked to carry six claims: control, specificity, necessity, uniqueness, content and use. The paper argues three points. First, the equivalence classes behind the non-identifiability results are set by the readout: the papers that report equivalence under a trait classifier or a binary refusal judge also report, under next-token logits or under the framing of the refusal, differences those readouts hide, so coarse scoring manufactures part of the equivalence. Second, adding a vector and removing it are not symmetric tests: addition is many-to-one, while ablation discriminates between directions, and the single-direction refusal claim was contested on the ablation side. Third, reading and writing come apart in both directions. The paper consolidates the measured values in one table and one figure, states an upgrade procedure from control to representation as an algorithm, and names five studies that would settle the open part. On the record, control is well supported, necessity is supported for a few behaviours, and uniqueness is supported for none.

Keywords: activation steering; steering vectors; representation engineering; linear representation hypothesis; identifiability; mechanistic interpretability; refusal direction; causal intervention

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

The argument is short enough to state in one line. Take pairs of inputs that differ in some property, average the difference of their activations at one layer, add a multiple of that average to the residual stream of a new input, and the model's output moves toward the property. Turner et al. (2023) made the move with a single pair of prompts; Panickssery et al. (2023) averaged over many pairs and called the result contrastive activation addition; Zou et al. (2023) framed a whole programme, representation engineering, around reading and writing population-level directions; Li et al. (2023a) shifted attention-head activations along truthfulness directions and raised an instruction-tuned LLaMA's TruthfulQA score from 32.5% to 65.1%. The most cited instance, Arditi et al. (2024), extracts one difference-in-means direction per model across 13 open chat models up to 72B parameters and shows that projecting it out disables refusal of harmful requests while adding it induces refusal of harmless ones. Persona vectors (Chen et al., 2025), function vectors (Todd et al., 2024), task vectors (Hendel et al., 2023), instruction

vectors (Stolfo et al., 2025) and a misalignment direction shared across fine-tunes (Soligo et al., 2025) followed the same extract-and-inject template.

The template carries an inference that is rarely stated. From "adding v produces behaviour B and removing v suppresses it" the literature moves to "the model represents concept C along v ". The second sentence is much stronger than the first. It says that v is the model's direction for C , not merely a direction that works; that it encodes C rather than something correlated with C ; and that the model itself uses v when it produces B without anyone intervening. Arditi et al. (2024) did not make that move themselves: their limitations section treats "refusal direction" as a functional label and allows that the direction could encode harm or danger, or nothing with a simple name. The field's later uses of their result have been less careful than its authors.

Against the template sits a growing set of results that look, at first sight, like a refutation. Venkatesh and Kurapath (2026) argue that steering vectors are non-identifiable because the Jacobian from an intermediate layer to the logits has a large null space, and report that perturbing a steering vector orthogonally, or replacing it

with a vector from a different extraction method that is often near-orthogonal or even anti-aligned to it, moves trait-classifier scores only by effect sizes at or near the threshold they treat as negligible (Cohen's d of 0.2). Wollschlaeger et al. (2025) find several independent directions and cones of up to five dimensions that mediate refusal. Joad et al. (2026) find that refusal directions estimated from eleven categories of non-compliance are often far from parallel, yet steering along any of them traces the same refusal versus over-refusal curve. Korznikov et al. (2025) show that adding random directions at magnitudes from a quarter to twice the typical activation norm makes safety-tuned models comply with harmful requests they otherwise refuse. Galeone et al. (2026) measure the angle between a direction that detects a property perfectly and the direction that best produces the corresponding behaviour, and find cosines of 0.12 to 0.20 across four models.

The two literatures do not engage with each other. Steering papers keep reporting efficacy and treating it as evidence about representation; the critical papers report brittleness, multiplicity and non-identifiability without saying what would upgrade a control result into a representational one. This paper argues that the disagreement dissolves once one asks what each result measured, and that what is left after the dissolution is a sharper and more useful question.

The argument runs as follows. Section 2 separates six claims that a steering result is used to support and connects them to the criteria for representation that philosophy of cognitive science and the probing literature have already worked out. Section 3 names the six choices on which every steering result depends. Section 4 shows that the equivalence classes behind the non-identifiability results are defined relative to the readout that scored the behaviour, and that the finer readouts reported in the same papers distinguish vectors the coarse ones do not. Section 5 shows that adding and removing a direction answer different questions, and that the single-direction refusal claim was contested on the removal side. Section 6 shows that steering vectors are usually selected by their effect, which makes their effect weak evidence for anything beyond control. Section 7 assembles the evidence that reading and writing dissociate in both directions. Section 8 explains why identifiability of a direction is purchased with a linearity assumption that is itself the claim at issue. Section 9 gives the strongest case for representational readings, which turns out to rest on invariance across extraction routes, models and languages rather than on efficacy. Section 10 consolidates the measured values in Table 1 and Figure 1. Section 11 states the upgrade procedure as Algorithm 1 and places the published record against it in Table 2. Section 12 draws out what follows for the safety uses of steering. Section 13 names five studies that would settle the open part, and Section

14 states what this analysis does not show.

The paper is a companion to two kinds of prior work. On the steering side, the closest precursors are Venkatesh and Kurupath (2026), who establish the non-identifiability argument, and Lin and Liu (2026), who audit mechanistic interpretability papers for unstated identification assumptions and conclude that validation is not identification. On the conceptual side, Harding (2023) operationalises representation in neural language models through three criteria - information, use and misrepresentation - and the probing literature (Hewitt and Liang, 2019; Ravichander et al., 2021; Elazar et al., 2021; Belinkov, 2021) worked through the reading half of the same problem several years ago. This paper does for the writing half what that literature did for the reading half: it asks what an intervention result is evidence of, given everything else that could produce it.

2. Six claims under one steering result

A paper that reports "a direction v for C " can mean any of the following, and the evidence needed rises steeply from the first to the last.

C1. Control. Adding α times v at site s shifts a behavioural readout R toward the target on distribution D . This is a claim about an intervention and an outcome. It says nothing about why the outcome occurs.

C2. Specificity. The shift produced by v is larger, or different in kind, from the shift produced by matched controls: random directions of the same norm, directions orthogonal to v , and vectors built from other contrast sets for the same property. Without C2, C1 could be a property of perturbing the residual stream at that site and magnitude at all.

C3. Necessity. Removing the component of the activations along v , by projection or mean-replacement, suppresses the behaviour on inputs where the model would otherwise show it. Necessity is a claim about the unsteered model: the behaviour, as the model normally produces it, passes through v .

C4. Uniqueness. The set of directions that satisfy C1 to C3 is, up to scale, a single ray. This is what "the direction for C " asserts and what " C is mediated by a one-dimensional subspace" asserts. It is an identifiability claim.

C5. Content. What v encodes is C , rather than a correlate of C in the contrast set: harm rather than refusal, a compliant persona rather than refusal, an answer-position preference rather than corrigibility, a surface register rather than truthfulness.

C6. Use. In unsteered operation, downstream computation reads the component along v and that reading is part of how the model produces B . Use differs from necessity. A direction can be necessary because its removal pushes activations off-distribution and breaks downstream computation generally, without being the

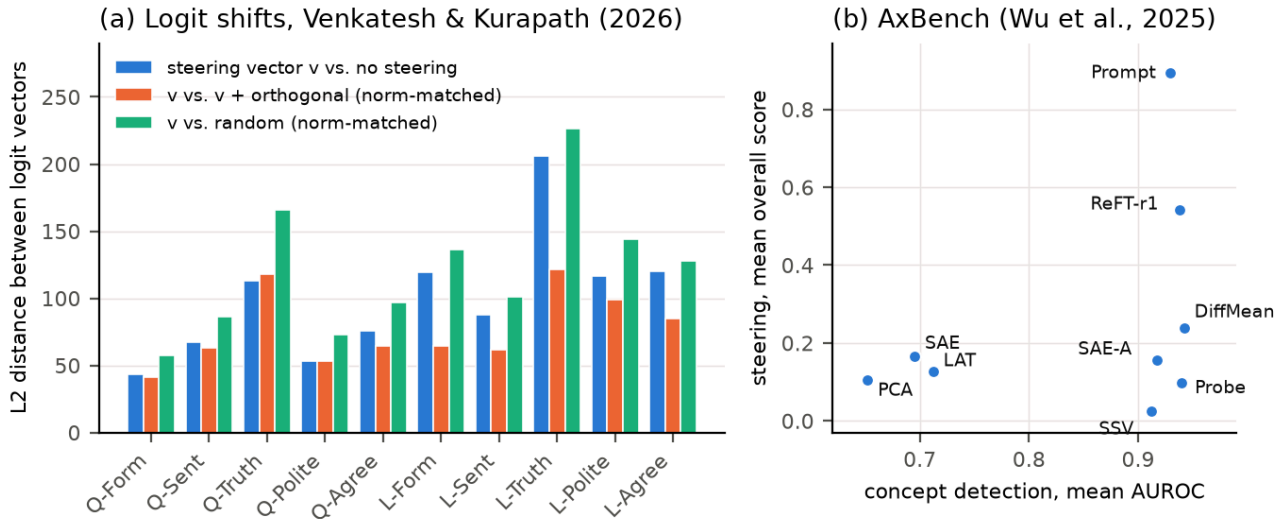


Figure 1. Published values re-presented by the author. (a) L2 distance between next-token logit vectors for five traits in Qwen2.5-3B-Instruct (Q) and Llama-3.1-8B-Instruct (L): steering vector v against no steering, v against v plus a norm-matched orthogonal direction, and v against a norm-matched random direction, from the appendix of Venkatesh and Kurapath (2026). Scored by logits, the orthogonal perturbation moves the output by an amount comparable to v itself, although trait classifiers rate unit-norm orthogonal perturbations as equivalent. (b) Mean concept-detection AUROC against mean overall steering score for the nine methods that appear in both of AxBench's main tables, from Wu et al. (2025). Detection and steering rank the methods differently. Values come from two studies with different models and readouts and are not a controlled comparison.

channel the model consults.

These are not idiosyncratic distinctions. Harding (2023) argues that a pattern of activations represents a property when it carries information about the property, when that information is used by the downstream system, and when it is in principle possible for the pattern to misrepresent the property. In this paper's terms, C5 is the information criterion made specific, C6 is the use criterion, and the possibility of misrepresentation is what C2 and C5 test when they ask whether v tracks C or something that merely co-varies with C on the contrast set. The interventionist account of causation in Woodward (2003) explains why C1 is cheap to establish and still worth something: a variable is causally relevant to an outcome when an intervention on it changes the outcome, and steering is an intervention. Causal relevance of a manipulation, however, is not the same as the claim that the manipulated quantity is the system's own variable for C. Many different manipulations can be causally relevant to the same outcome. Mueller et al. (2025) organise mechanistic interpretability around the choice of causal mediator - neurons, directions, subspaces, features - and Geiger et al. (2023) place steering alongside patching, concept erasure and circuit analysis within one theory of causal abstraction; both frameworks make explicit that an intervention result is a result about a chosen mediator, and that choosing it is the analyst's act, not the model's.

The probing literature already learned this lesson for reading. Hewitt and Liang (2019) showed that a probe's accuracy reflects the probe's capacity as well as the representation, and introduced control tasks to separate

them. Ravichander et al. (2021) showed that models encode properties their training task does not need, and encode them above chance even when the property is random noise in the data. Elazar et al. (2021) found that how well a property can be probed does not predict how much the model relies on it, and proposed removing the property and watching behaviour instead. Belinkov (2021) consolidated these cautions. The steering literature inherited the removal idea from this line - the ablation half of Arditi et al. (2024) is, in effect, amnesic probing with a behavioural readout - but not the matching caution for addition. Adding a direction is to writing what a probe is to reading: a test with its own capacity, which can succeed for reasons that have little to do with the model's own variables.

One further distinction keeps the argument honest. Steering can be a fully legitimate engineering result while failing every claim from C4 onward. Venkatesh and Kurapath (2026) make the point themselves: their equivalence results do not say that steering fails to move behaviour, only that the movement does not localise the semantic content. The question this paper asks is not whether steering works but what its working licenses.

3. A steering result is an estimate with six indices

Every reported steering vector is the output of a procedure, and the procedure has at least six free choices. Naming them makes clear what a reported vector is conditional on.

The contrast set D. Which pairs of inputs are taken to differ in the property, and in what else they differ. Tan

et al. (2024) show that, for some of the Model-Written Evaluations datasets, a large share of the variance in steerability across inputs is explained by which answer letter or token carries the positive option, a property of the template rather than of the behaviour. Braun (2026) finds that steering vectors trained on seven different prompt formats for the same behaviour point in distinct directions while producing similar average effects.

The extraction rule E. Difference of means, the first principal component of the differences, a logistic-regression weight vector, a gradient-optimised direction, a sparse-autoencoder latent or a learned low-rank intervention. Im and Li (2025) prove that the mean of differences is the optimal steering vector for moving negative examples toward positive ones under their objective and find that it outperforms PCA and classifier-based directions on most tasks they test; Singh et al. (2024) derive optimal affine steering functions under least-squares constraints. These are optimality results for an objective, not identification results for a concept, and different objectives give different vectors.

The site s. Layer and token position. Arditi et al. (2024) compute a candidate at every post-instruction position and every layer before choosing one; Wang et al. (2026) find that the layer at which an emotion is best read out by a probe and the layer at which steering is most effective differ in three omni-modal models, by seven layers in Qwen2.5-Omni-7B.

The magnitude alpha. Steering is not scale-free. Korznikov et al. (2025) scale unit vectors to between a quarter and twice the mean activation norm at the layer, and their random-direction effects appear across a range of such multiples; Venkatesh and Kurapath (2026) report main-text equivalence tests with random unit-norm perturbations added to an unnormalised mean-difference vector, so the perturbation's size relative to the vector is not stated, and run a separate norm-matched test in an appendix. Results at different magnitudes are results about different interventions.

The readout R. A trait classifier, an LLM judge, a refusal string match, the probability of the behaviour-matching multiple-choice letter, a next-token logit vector, or a full distribution over open-ended generations. Pres et al. (2024) show that contrastive activation addition for corrigibility succeeds when scored by multiple-choice letter probability and fails to produce the behaviour in open-ended generation. Section 4 argues that R is the index that sets the size of the equivalence class.

The selection rule S. How the reported vector was chosen from its candidates. Arditi et al. (2024) select, from the $|I|$ times L candidates, the one that best bypasses refusal when ablated, subject to inducing refusal when added and to a bound on the divergence it causes on harmless inputs. Wu et al. (2025) select the steering factor per concept by the judge score it achieves. Selection on the outcome is ordinary

engineering; Section 6 argues that it changes what the outcome can be evidence for.

The six indices make a reported vector a sample from a procedure, not a property of the model. The question every representational claim must answer is which of the indices the claim depends on and which it survives.

4. Equivalence is set by the readout

The strongest form of the deflationary argument is non-identifiability: many vectors produce the same behaviour, so behaviour cannot pick out the model's vector. The argument is correct as far as it goes. What it leaves implicit is that "the same behaviour" is always behaviour as some readout measures it, and the size of the equivalence class depends on how fine that readout is.

4.1 The null-space argument and what its tests measured

Venkatesh and Kurapath (2026) linearise the map from a layer's activations to the output logits and observe that any vector in the null space of that Jacobian can be added to a steering vector without changing the linearised output. They estimate the dimension of a related subspace from the covariance of activations and report effective ranks of 57 to 63 of 4096 dimensions in Llama-3.1-8B-Instruct and 116 to 139 of 2048 in Qwen2.5-3B-Instruct, leaving 86% to 94% of the space with near-zero variance. Their own limitations section notes that this activation-variance subspace has not been shown to coincide with the Jacobian null space the proof requires.

Their behavioural tests use five traits - formality, politeness, sentiment, truthfulness and agreeableness - scored by pre-trained classifiers. Adding random unit vectors orthogonal to the steering vector gives mean Cohen's d of 0.119 (Llama) and 0.131 (Qwen) between the original and perturbed conditions over ten seeds, below the 0.2 threshold they use for negligible effects. Comparing a mean-difference vector with the first principal component of the same differences, whose cosine with it ranges from -0.54 to +0.32, gives $|d|$ below 0.27 in every trait and model. The criterion is an effect-size threshold, not a significance test, and not every condition meets it: in a test across prompt distributions the in-distribution values for truthfulness and agreeableness reach d of about 0.7, which the authors attribute to "higher classifier variance for these cognitively complex traits rather than a genuine breakdown of equivalence" (Venkatesh and Kurapath, 2026).

The same paper also reports a finer readout, and the finer readout tells a different story. In its appendix the authors compute the L2 distance between next-token logit vectors under the steering vector alone, under the steering vector plus a norm-matched orthogonal

direction, and under a norm-matched random direction. For Qwen, the orthogonal perturbation moves the logits by 40.9 to 118.2 units against 43.7 to 113.1 units for the steering vector itself relative to no steering; on truthfulness the perturbation moves the logits further (118.24) than the steering vector does (113.06). For Llama the orthogonal perturbation's shift is 54% to 84% of the steering vector's own. In every condition it is 27% to 53% smaller than the shift caused by a norm-matched random direction (Figure 1a). The authors read this as evidence that orthogonal perturbations are less disruptive than random ones, which it is. It is also evidence that, scored by logits rather than by a trait classifier, a norm-matched orthogonal perturbation is not behaviourally inert: it changes the next-token logits by 54% to 105% of the amount the steering vector itself does (the ratios are the author's, computed from the published table). The classifier-level equivalence is reported for unit-norm perturbations and the logit-level comparison for norm-matched ones, so the two tables do not measure the same perturbation. An appendix asserts that orthogonal perturbations "leave semantic classifier scores nearly unchanged even when their norm is comparable to the steering vector itself" (Venkatesh and Kurupath, 2026), but the paper reports no classifier measurement under the norm-matched condition. What can be said is that the paper's own data show no equivalence at the logit level for norm-matched orthogonal directions.

The point is not that the non-identifiability result is wrong. Under a readout that maps outputs to five trait scores, a large class of vectors is equivalent, and that is a real and useful finding about what trait-level evaluation can distinguish. The point is that the class is a property of the pair (model, readout), not of the model alone. A claim of the form "v is the model's direction for C" cannot be refuted or established by any readout coarser than the distinctions the claim makes.

4.2 One knob, many directions, different refusals

Joad et al. (2026) reach the same structure from the other side. They estimate a refusal direction from each of eleven splits of four refusal and non-compliance benchmarks and work with Gemma-2-9B-IT, Llama-3-8B-Instruct and Qwen-7B-Chat. On Gemma-2-9B-IT, pairwise cosines between the directions are typically 0.4 to 0.6, with some pairs close to orthogonal, and they cluster by category: safety and content-policy refusals sit together, capability-limitation and underspecification refusals elsewhere. Directions estimated from repeated 32-versus-32 samples of the same category agree at cosines typically of 0.95 or higher, so the spread across categories is not estimation noise. Yet when each direction is added with increasing strength, all of them trace nearly the same curve of refusal on harmful prompts against over-refusal on benign ones. Scored by a binary refusal judge, the directions are one knob.

Scored more finely, they are not. The same authors classify each steered refusal by its framing - capability limitation, safety or ethics, underspecification, generic - and find that the direction used changes how the model refuses even when it does not change whether it does. A sparse-autoencoder decomposition shows a shared core of refusal latents plus latents specific to style and domain. Their own summary is that linear interventions collapse a structured internal representation into uniform behavioural control. That is the readout argument in its cleanest form: under one readout the class of refusal directions is one-dimensional; under a readout that records refusal type, it is not.

4.3 Different prompts, different vectors, similar averages

Braun (2026) trains steering vectors for 36 Model-Written Evaluations behaviours on Llama-2-7B-Chat at layer 13 using seven prompt types. Every prompt type gives a net positive shift in the behaviour-matching logit on average, none consistently beats the others, and the vectors from different prompt types do not align. Averaged over datasets, between 29% and 43% of evaluation samples move in the opposite direction from the one intended, and on individual datasets the fraction ranges from 3% to 50%. The averaged readout makes the vectors look interchangeable; the per-sample readout shows that each is unreliable in its own way. Tan et al. (2024), on 40 datasets with Llama-2-7b-Chat and Qwen-1.5-14b-Chat, find several datasets on which close to half of inputs are anti-steerable, and attribute part of the variance to steerability biases toward a particular answer letter or token.

4.4 A random-direction control is a readout control too

Riegler and Torpmann-Hagen (2026) make the same point about sparse-autoencoder features. The usual protocol labels a feature from its top-activating contexts and confirms the label by steering that feature at one typical magnitude. They argue that this inspects one cell of a grid that crosses the steering condition - a single feature, a joint feature set, or a matched random direction - with the coefficient, and that other cells change the label. Features labelled as an AI self-disclaimer switch, under steering, to a contemplative voice in one model and a collective first-person voice in another; three near-orthogonal features that are individually substitutable turn out to be jointly necessary. When the residual stream is distorted by the same amount, steering along the feature directions harms performance on unrelated control tasks while random directions of equal magnitude leave it intact, and the confidence intervals of the two conditions do not overlap in any of the three models tested. A random-direction control at matched geometry, scored on unrelated tasks, separates directions that a single-cell readout does not.

4.5 What the section establishes

Taken together, these results support a precise statement. For a readout R , define the equivalence class of v as the set of vectors whose steered outputs R cannot distinguish from those of v . The class is large when R is a trait classifier, a binary judge or an averaged score; it shrinks under logits, refusal framing, per-sample effects and off-target task damage. Non-identifiability results are therefore results about R as much as about the model, and so are uniqueness claims. Neither side of the dispute has reported the class under the finest readout available - the full next-token distribution on a broad input set - and that is the measurement the dispute needs.

5. Adding is not removing

The second observation concerns the direction of the intervention. Arditi et al. (2024) made two claims at once: that ablating the direction disables refusal, and that adding it induces refusal. The literature since then has treated these as one result. They are not, and the evidence has accumulated asymmetrically.

5.1 Addition is many-to-one

The evidence that many directions suffice to produce a behaviour is strong and comes from several directions. Joad et al. (2026) show that any of eleven refusal directions induces refusal along the same trade-off. Wollschlager et al. (2025) find that refusal can be induced and bypassed through multiple directions and through cones of up to five dimensions in every model they test, and that larger models in the Qwen2.5 family support higher-dimensional cones. Korznikov et al. (2025) find that random directions, not chosen for any property, raise harmful compliance from 0% to between 1% and 13% depending on model and prompt, and that averaging twenty random vectors that each jailbreak the same single prompt yields an attack that transfers to unseen requests. Subramani et al. (2022) showed early that a frozen decoder can be steered to reproduce a specific target sentence nearly perfectly (above 99 BLEU) by an added vector found for that sentence. If a vector exists for an arbitrary target sentence, the existence of a vector for a target behaviour is weak evidence that the behaviour has a dedicated representation.

The sufficiency side also admits a specific mechanism for illusion. Makelov et al. (2023) show that patching a subspace can make a model behave as if a feature had changed by activating a dormant parallel pathway through a different subspace that, in normal operation, is causally disconnected from the output. The intervention succeeds, and the subspace it was applied to is not where the model keeps the feature. Wu et al. (2024) reply that the criterion Makelov et al. use would classify some intuitively correct explanations as illusions too, and that the illusions observed in practice reflect training and evaluation choices; the exchange leaves standing that sufficiency of an intervention, taken alone, does not

locate a variable. Hase et al. (2023) found the editing-side version of the same gap: the layer that causal tracing identifies as storing a fact gives no insight into which layer is best to edit in order to override it, while the layer edited is a far better predictor of editing success. Where an intervention works and where the information lives are separate questions.

5.2 Removal discriminates

Ablation behaves differently. In Joad et al. (2026), projecting out a single category's refusal direction flips only part of the harmful prompts the model originally refused, and how much depends sharply on the direction and model: from 0.0% to 73.5% across categories and models in their main table on the unified test pool (the appendix reports 91.8% for the over-refusal direction in Qwen-7B-Chat). The bottom row of that table, which the caption labels a round-robin direction sampled across all eleven splits, reads 87.1% (Qwen), 69.5% (Llama) and 81.1% (Gemma), but the per-recipe tables in their Appendix J place these values elsewhere: 87.1% belongs to a direction built from a single curated pool of WildGuardMix prompts, and 69.5% and 81.1% are the best of five seeds of a filtered recipe, which the authors call Crisp, that draws its harmful prompts from all the per-split pools; its five-seed mean is 51.4% (Qwen), 65.5% (Llama) and 78.9% (Gemma). Directions pooled round-robin from three prompts per category, 33 prompts in all, flip 35.0% (Qwen), 37.9% (Llama) and 45.8% (Gemma) on average over five seeds, and 63.8%, 41.0% and 73.6% for the best seed; a 55-prompt round-robin pool flips less, 27.0% to 43.3% on average. Under addition the directions were one knob; under removal they are distinguishable, and in only 2 of the 33 category-model pairs does a single category direction flip a majority of refusals. Wollschlager et al. (2025) report that directions orthogonal in the Euclidean sense still interfere under ablation - ablating one changes how strongly the other is represented downstream - and introduce representational independence, a criterion that requires directions not to affect each other under intervention across layers. Ablating more representationally independent directions raises attack success monotonically in Gemma 2 2B. Piras et al. (2026) derive multiple refusal directions with self-organising maps and find that ablating several outperforms the single direction and specialised jailbreak algorithms. Pan et al. (2025) find in Llama 3 8B a dominant refusal direction accompanied by smaller orthogonal directions for features such as hypothetical narrative and role-play, which promote or suppress the dominant one.

Removal also has its own artefacts, and they are different ones. Marshall et al. (2024) show that projecting out a refusal direction produces incoherent outputs in at least one model, an RWKV v5 recurrent network, and that an affine edit which also accounts for the mean activation controls refusal more precisely across ten models

including Llama 3 70B. Necessity results can thus be inflated by off-distribution damage: removing a direction may break generation broadly rather than removing the variable the model consults. That is why C3 and C6 are separate claims.

5.3 Where the single-direction claim was actually contested

Read with this asymmetry in mind, the contradiction between Arditi et al. (2024) and its successors becomes specific. On the addition side, the successors confirm that a single direction suffices to induce refusal; they add that many others do too. On the removal side, they show that ablating one direction is not the only way to bypass refusal, that ablating several does better, and that the direction which ablates well is plausibly better described as the common core of a family than as the refusal variable, although in Joad et al. (2026) how much refusal a pooled direction removes depends on how the pool is built: a direction pooled round-robin across the family's eleven categories flips fewer than half of refusals on average in each of the three main-text models (78.9% in Qwen-1.8B-Chat), while a filtered pool drawn from the same categories flips 51.4% to 78.9% on average. What failed is C4, and it failed on removal evidence. C1 survived, and so did a weaker version of C3: Arditi et al. (2024) find a selected direction whose removal disables most refusal in most of the models they test, a single direction built from one curated pool flips 87.1% of refusals in Qwen-7B-Chat (Joad et al., 2026), and Wang et al. (2025a) find that a direction extracted in one safety-aligned language transfers to others with near-complete effectiveness across 14 languages.

The asymmetry has a simple explanation in terms of Section 4. Addition is scored by whether the output moves toward the target, a readout that many perturbations satisfy. Removal is scored by whether the behaviour, produced by the model's own computation, survives; only perturbations that intersect the route the model actually uses can block it. Removal is a finer test of C6 than addition is, because it asks the model's computation to do the discriminating.

6. Selected because it works

The third observation concerns selection. A steering vector reported in a paper is rarely the only vector the authors computed. It is the one that worked.

6.1 Outcome selection

Arditi et al. (2024) compute a difference-in-means candidate for each post-instruction token position and each layer and keep the candidate that best bypasses refusal when ablated, provided it also induces refusal when added and keeps the KL divergence it causes on harmless prompts when ablated below 0.1, and provided it lies before the last fifth of the layers. Their claim that refusal is mediated by a single direction is therefore

partly a claim about the best member of a large candidate set, chosen by the criteria the claim then reports. This is not a flaw in an existence proof, which is how the authors describe their paper. It does mean that the success of the selected vector cannot also count as independent evidence that it is the model's direction: with $|I|$ times L candidates, part of the selected vector's success may reflect selection alone, and how much is not reported. Wu et al. (2025) select the steering factor per concept by judge score, though on one half of their instructions, scoring it on the other half; Korznikov et al. (2025) select coefficients by sweeping. Wollschlager et al. (2025) note that in many models, especially in the Qwen 2.5 family, most difference-in-means candidates have side-effects too large for them to count as a successful attack. Selection is everywhere, and a vector's selected success has to be discounted by the size of the pool it was selected from.

6.2 The contrast set predicts steerability

If steering success reflected a property of the concept, it should be roughly constant across contrast sets for the same concept and differ across concepts. The evidence points the other way. Braun (2026) finds that steerability is predicted by geometric properties of the contrast set's activations: the more the individual activation differences agree in direction with their mean, and the better separated positive and negative activations are along that mean, the more reliable the steering. Tan et al. (2024) find that steerability is largely a property of the dataset, correlated across two models, and that in-distribution and out-of-distribution steerability correlate at Spearman 0.891 for Llama-2-7b-Chat and 0.694 for Qwen-1.5-14b-Chat, with outliers. Da Silva et al. (2025) test DoLa, function vectors and task vectors on up to 36 models from 14 families between 1.5B and 70B parameters and find that many models show no improvement, and some degrade, under methods that were reported to work on a handful of models.

These results do not show that steering vectors are noise. They show that steerability is a property of the triple (concept, contrast set, model), and that the contrast-set term is large. A contrast set encodes whatever its pairs differ in. Farquhar et al. (2023) prove for contrast-consistent search, the unsupervised method of Burns et al. (2023), that arbitrary features, not only knowledge, satisfy the consistency structure it relies on, and show experimentally that the method finds whichever feature of the activations is most salient. Supervised difference-in-means inherits the same exposure in a milder form: it recovers the most salient direction along which the contrast set's pairs differ, which is C only if the pairs differ in nothing else that the model encodes prominently. Bolukbasi et al. (2021) traced an earlier interpretability illusion to the same source: directions that seem to encode one simple concept on one corpus encode something more complex once tested on others.

6.3 What selection does to the claims

Outcome selection and contrast-set dependence bear on different claims. Selection inflates C1 and C3 for the reported vector relative to a typical vector from the same procedure; the correction is to report the distribution over candidates and the success rate of a held-out selection. Contrast-set dependence threatens C5 directly: if the vector is the most salient difference in D, its content is whatever D varies, and only a second contrast set that holds the suspected confound fixed can separate C from it. Pres et al. (2024) add a third point: a vector that succeeds on the readout used for selection - multiple-choice letter probability - can fail on the readout that matters downstream, open-ended generation.

7. Reading and writing come apart

A representational claim implies that the direction which carries information about C and the direction through which the model acts on C are related. The published record shows that they can come apart in both directions.

7.1 Readable but not writable

Wu et al. (2025) built AxBench to compare concept detection and steering across methods on the same concepts in Gemma-2-2B and 9B. For detection, difference-in-means and linear probes reach mean AUROC of 0.942 and 0.940. For steering, scored by a judge as the harmonic mean of concept presence, instruction following and fluency, the same two methods score 0.239 and 0.098, against 0.894 for prompting and 0.543 for a rank-one representation fine-tune. The probe direction is among the best detectors and among the worst steerers; sparse-autoencoder latents improve detection when selected with labels but steer no better (Figure 1b). Wang et al. (2026) find that the probe-best layer for emotion scatters over most of the network's depth across three omni-modal models while the steer-best layer stays between 0.48 and 0.74 of normalised depth. On Qwen2.5-Omni-7B, single-layer injection at the probe-best layer is indistinguishable from a random direction; moving that single-layer injection to a mid-band layer strengthens the effect about sevenfold, and injecting across several mid-to-late layers gives an effect about 26 times that at the probe-best layer. Xun (2026) reports, across the Pythia suite from 160M to 12B and eight checkpoints, that a linear probe reads a target variable with AUROC of at least 0.990 from the first checkpoint at every scale while steering along the probe's direction is null-equivalent in 43 of 48 model-checkpoint cells. Galeone et al. (2026) find that Gemma 2 2B separates fake from real entities with perfect linear separability from layer 5, yet a direction that detects them perfectly sits at cosine 0.12 to the direction that produces refusal of fake entities, a second perfect detector built by difference of means sits at -0.06, and

the first cosine stays between 0.12 and 0.20 across four models from three families. The same paper reports the opposite case for output format, where "detection and control collapse onto a single axis: one and the same direction both classifies the behavior and, when added, flips it" (Galeone et al., 2026); reading and writing can coincide, and whether they do depends on the behaviour.

These are the writing-side analogues of Ravichander et al. (2021) and Elazar et al. (2021). Information about C in a direction, even information the model will later use, does not make that direction a handle on the behaviour; and the handle need not be where the information is most readable. Arad et al. (2025) draw the same distinction inside sparse autoencoders: features that capture patterns in the input and features with an interpretable effect on the output rarely score highly on both, and filtering for output effect improves steering by a factor of two to three.

7.2 Writable but not read

The opposite dissociation is the one that matters most for representational claims, because it shows control without information. Random directions that break safety behaviour (Korznikov et al., 2025) are writable handles that were not chosen to encode anything. Subramani et al. (2022) find vectors for arbitrary target sentences. Makelov et al. (2023) find subspaces whose patching changes the output through a dormant pathway the model does not use. Sutter et al. (2025) push the point to its limit for causal abstraction: if the map between a model's activations and a hypothesised variable may be arbitrarily nonlinear, any network can be aligned to any algorithm, and they report 100% interchange-intervention accuracy on indirect object identification in randomly initialised language models. Control through a sufficiently flexible intervention is available whether or not the model represents anything.

Zhong and Li (2026) supply a case where the writable handle and the model's own route are in series. In Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct, steering along a compliant-persona direction drops Llama's refusal rate from 97% to 2%, and the non-refusals split into 42% bypass and 56% degenerate output, so more than half of the steered outputs are degenerate rather than compliant; adding the refusal direction back brings refusal partly back when done late in the network and not when done early; and removing the persona direction over a window of late layers returns refusal to baseline, which removing a random direction fails to do. Refusal, on this evidence, is computed in one place and gated at a later expression stage by persona. A steering result on either direction would have been reported as a direction "for" refusal. Reblitz-Richardson (2026) adds that, in OLMo-3, widening the patched slice of a moral-judgment subspace transfers more and more of moral judgment's response but levels off for refusal at the level

of a single harm direction, with about three-quarters of refusal's causal input lying outside the moral subspace; refusal reads a narrow harm percept rather than the moral content the model also has.

7.3 What the dissociations establish

The dissociations show that C1 and C5 are dissociable in practice, not only in principle: a direction can carry C without being a handle on behaviour, and can be a handle without carrying C. The upgrade from C1 to a representational claim therefore needs evidence that bears on information and use directly. The tools exist: a probe with control tasks for information (Hewitt and Liang, 2019), removal with a measured downstream effect for use (Elazar et al., 2021), and the misrepresentation test Harding (2023) proposes, in which an input whose property the model gets wrong should push the direction the wrong way.

8. The linear premise does the identifying

A direction is identifiable only relative to a model of how the network encodes information. Without that model, the question "which direction represents C" has no answer, and with it, the answer is only as good as the model.

8.1 The case for linearity

There are serious reasons to expect linear structure. Park et al. (2024) give two formalisations of linear representation, one in the output-embedding space and one in the input-representation space, prove that they connect to probing and steering respectively, and identify a non-Euclidean causal inner product under which the notions unify; in Llama-2 they construct probes and steering vectors from counterfactual pairs. Jiang et al. (2024) argue that the next-token objective with softmax cross-entropy and the implicit bias of gradient descent together favour linear representations in a latent-variable model of concept dynamics. Marconato et al. (2025) show that, under suitable conditions, many linear properties hold either in all next-token predictors that induce the same distribution or in none of them. Roeder et al. (2020) show that a large family of discriminative models is identifiable in function space up to a linear transformation. Elhage et al. (2022) explain how a network can store more sparse features than it has dimensions by superposition, which makes linear directions natural carriers. Park et al. (2026) argue that, for representations that parameterise a softmax, the natural geometry is information geometry, and derive a dual steering method from linear probes that modifies the target concept optimally while minimising off-target change.

These results are the strongest support for reading steering representationally, and they share a structure. Each one says that if the model's concepts are linear in a stated sense, and if the intervention is built in a stated

way - from counterfactual pairs, in the right inner product - then probing and steering recover the same object. None of them says that the model's concepts are linear in that sense. Marconato et al. (2025) say that linear properties are shared across distribution-equivalent models when they hold; Roeder et al. (2020) identify up to a linear map, which leaves which direction in that map is "the" concept to further constraints.

8.2 The case against global linearity

Several results show representations that one direction cannot capture. Engels et al. (2025) find irreducibly multi-dimensional features in GPT-2 and Mistral 7B, among them circular representations of days of the week and months, and give intervention evidence in Mistral 7B and Llama 3 8B that these circles are the fundamental unit of computation for modular arithmetic over them. Csordas et al. (2024) find that gated recurrent networks trained to repeat sequences can represent the token at each position by an order of magnitude rather than a direction - the smallest networks find only this solution, while larger ones have linear representations - a counterexample to the strong linear representation hypothesis. Li et al. (2023b) originally found that a nonlinear probe was needed to read the Othello board state from a sequence model; Nanda et al. (2023) then found a linear representation once the board was coded relative to the current player, which shows that linearity can depend on choosing the right variable as well as the right direction. Hernandez et al. (2023) find that for a subset of relations a single linear map from subject to attribute approximates the model's computation well, but also many relations the model knows that are not linearly decodable that way. Bigoulaeva et al. (2026) find instruction representations that are linearly separable yet interact non-linearly in their causal effects, and describe them as selectors among later-layer pathways.

On the steering side, Raval et al. (2026) measure the ratio of geodesic to Euclidean distances between activations and report concept-dependent distortions, and find that steering in a polynomial kernel-PCA feature space outperforms linear PCA steering most where the distortion is largest. Gao et al. (2026) propose that steering acts through a central axis plus a surrounding plane whose sensitive sectors are sample-specific and cannot be identified from difference vectors, which they offer as a principled source of steering's fluctuation even with well-aligned directions. Braun (2026) concludes that steering is unreliable when the behaviour's representation is not well approximated by the linear steering direction.

8.3 The dilemma

The two halves combine into a dilemma that Sutter et al. (2025) state for causal abstraction and that applies to steering without change. If the analyst constrains interventions to be linear and additive, a steering vector

is identifiable up to the equivalence classes of Section 4, but the identification is conditional on a linearity premise that several results show fails for some features. If the analyst relaxes linearity, the class of interventions that reproduce the behaviour becomes so large that the analysis can align a randomly initialised network with a task it cannot perform. Meloux et al. (2025) find the same multiplicity by exhaustive enumeration in small networks where every candidate can be checked: several circuits replicate a behaviour, a circuit admits several interpretations, and one algorithm aligns with several subspaces. Classical results make the same point for unsupervised recovery: nonlinear independent component analysis has infinitely many solutions without further constraints (Hyvarinen and Pajunen, 1999). Rajendran et al. (2024) give concepts a formal definition and show that they can be provably recovered when the data are diverse enough, which is an identification result with stated assumptions rather than an empirical finding about a given model.

The consequence for the steering debate, offered here as the author's conjecture rather than a tested result, is that a working linear steering vector can be a local linear approximation to a nonlinear representation, valid on the inputs and magnitudes tested. That would explain at once why steering often works, why it fails on a large fraction of individual inputs, why vectors from different contrast sets differ while working similarly on average, and why random directions sometimes work: near a given activation, many directions have a component along the tangent of whatever the model uses. On this reading C1 is real, C4 is ill-posed until the representational format is fixed, and the linear premise has been doing the identifying all along. Lin and Liu (2026) ask mechanistic interpretability papers to state such premises as identification assumptions; for steering, the first premise to state is that the property is encoded linearly at the injection site over the input range tested.

9. The strongest case for representational readings

The deflationary results do not show that no steering vector picks out a representation. Several lines of evidence support representational readings, and what they have in common is instructive.

Invariance across extraction routes. Marks and Tegmark (2024) find linear structure in language-model representations of simple true/false statements by visualisation, probe transfer across datasets and surgical intervention that makes the model treat false statements as true and vice versa, and report that difference-in-means directions generalise as well as other probes while being more causally implicated in outputs. When directions from independent routes - reading by probe, writing by intervention - agree, the agreement is evidence that neither route alone provides.

Necessity on natural tasks. Tigges et al. (2023) find a sentiment direction across several models and show, in Pythia-2.8b on the Stanford Sentiment Treebank reviews that it classifies correctly and that contain a comma, that ablating it removes 76% of above-chance zero-shot classification accuracy, of which 36 points come from ablating the summarised sentiment direction at comma positions alone. This is a removal result on the model's own task with a fine readout, which is evidence for C3 and C6.

Invariance across languages and models. Wang et al. (2025a) find that a refusal direction extracted from one safety-aligned language bypasses refusal in others with near-complete effectiveness across 14 languages, and attribute this to parallel refusal vectors in the embedding space. Soligo et al. (2025) find that a misalignment direction extracted from one emergently misaligned fine-tune of Qwen2.5-14B-Instruct ablates misaligned behaviour in fine-tunes with higher-rank adapters and different datasets. Wang et al. (2025b) find, by comparing sparse-autoencoder features before and after fine-tuning, a toxic-persona feature that most strongly controls emergent misalignment and predicts whether a model will show it. A direction that survives a change of language, of fine-tuning run or of dataset has passed a test of the C5 kind that one contrast set cannot pass.

Prediction of the model's own changes. Chen et al. (2025) report that shifts along persona vectors after fine-tuning correlate strongly with the trait changes the fine-tuning produces, both intended and unintended, and that the vectors flag training samples that will produce the shift. A direction that predicts how the model's behaviour will move under training it was not built from is evidence of use, not only of control. Blank et al. (2026) report that subliminal learning - a student acquiring a teacher's trait from semantically unrelated outputs - is mediated by a steering vector that approximates the teacher's system prompt, and that system prompts not well approximated by a steering vector are not transmitted; they also show that adding a semantic vector can have model-specific effects with no semantic content.

Structure in the vector itself. Todd et al. (2024) find that a small number of attention heads carry a compact representation of a demonstrated task, robust to changes of context, which triggers the task in zero-shot settings and composes to some extent under addition. Nanda et al. (2023) control an Othello model's behaviour with vector arithmetic on a linear board representation. Function vectors and the Othello board representation come with evidence about where the vector comes from in the model's own computation, which is evidence of use.

What these results share is that none of them rests on efficacy. Each adds an invariance - across extraction routes, positions, languages, fine-tuning runs, training

dynamics or tasks - or a removal result on the model's own task with a fine readout. That is the pattern the upgrade procedure in Section 11 formalises. The strongest representational evidence in the steering literature is strong because it does what the efficacy evidence does not: it shows that the direction is not an artefact of the six indices of Section 3.

10. The evidence, consolidated

Table 1 brings together, one cited result to a row, the published numbers that the preceding sections lean on. Three patterns stand out. First, the rows that vary the extraction rule or the contrast set while holding a coarse readout fixed (rows 1, 2, 4 and 6) report equivalence or similar averages, and the rows that score comparable conditions with a finer readout (rows 3 and 7) report differences. Second, the rows that report addition (rows 4, 9 and 17) find many sufficient directions, while the rows that report removal (rows 5, 8, 10 and 11) find that single directions are partially necessary, that combinations do better, and that removal has artefacts of its own. Third, the reading-versus-writing rows (12 to 15) show dissociation in four settings with different models, modalities and methods, and row 18 shows a refusal handle that sits downstream of a persona direction. The rows that support representational readings (19 to 21) report removal on a natural task or invariance across languages and fine-tunes, not efficacy alone; row 22 is the kind of efficacy result the template starts from.

Two cautions apply to Table 1 and Figure 1. The rows come from different models, layers and readouts, and the table invites comparison within rows, not across them. And several rows are 2026 preprints from single groups; Section 14 returns to what that does to the conclusions.

11. Upgrading control to representation

11.1 Evidence required per claim

For every claim of Section 2, Table 2 gives the evidence that licenses it, the alternative that evidence excludes, and the state of the published record as Sections 4 to 9 have read it. The assignments are this paper's.

11.2 The procedure

Algorithm 1 states the decision structure. Its organising rule is that a claim is licensed only after the test that rules out the alternative it competes with has been run under a readout at least as fine as the distinction the claim makes.

Three features of the procedure deserve comment. First, step 3 is where most published claims stop, because the controls it asks for are rarely reported; when they are, as in Korznikov et al. (2025), Riegler and Torpmann-Hagen (2026) and Wang et al. (2026), they change the reading. Second, step 5 replaces the question "is this the direction?" with the question "how large is the class of

directions that do the same thing, under the finest readout available?". A subspace claim is often the strongest claim the evidence supports, and it is a respectable one: Wollschlager et al. (2025) and Joad et al. (2026) both end with subspace or core-plus-periphery descriptions of refusal. Third, step 7 is the only step that does not involve steering at all. Use is established by what the direction predicts about the unsteered model, including when the model gets the property wrong, which is the misrepresentation criterion of Harding (2023).

11.3 Three published results, re-read

The refusal direction. Arditi et al. (2024) establish C1 in 13 models and a strong form of C3: projecting out one selected direction disables most refusal. Selection (step 2) was on the outcome over a large candidate pool. Specificity (step 3) against norm-matched random directions is not reported, and Korznikov et al. (2025) show that random directions alone produce some harmful compliance - 1% to 13%, much less than the bypass Arditi et al. report for their selected direction, though the two studies' setups differ. Uniqueness (step 5) fails: Wollschlager et al. (2025) and Joad et al. (2026) find multiple directions and cones. Content (step 6) is where the most informative recent evidence lies. In an appendix analysis, Joad et al. (2026) contrast four prompt buckets - harmful prompts the model refuses, harmful prompts it complies with, benign prompts it refuses and benign prompts it complies with. The standard harmful-refused minus benign-complied direction is recovered at mean cosine 0.96 from 32-versus-32 subsamples, but the direction separating benign refusals from benign compliance is nearly orthogonal to it (mean cosine about 0.10), and the authors conclude that the standard direction is specific to the harmful-versus-benign contrast rather than a generic refusal axis. Varying refusal among harmful prompts gives a direction moderately aligned with the standard one (cosine 0.68 to a held-out oracle); varying it among benign prompts does not. The contrast set that defines "the refusal direction" varies harm and refusal together, and once refusal is varied with harmlessness held fixed, a largely different direction appears. This is the C5 failure Arditi et al. anticipated when they allowed that the direction might encode harm. The reading the evidence supports is that there is a harm-keyed direction whose removal disables most refusal of harmful requests, embedded in a larger refusal subspace that also contains category- and style-specific directions. In Joad et al. (2026) that shared core shows up both in sparse-autoencoder latents and in pooled directions, and how much refusal a pooled direction removes depends on how the pool is built: round-robin pools flip 35.0% to 45.8% of refusals on average in the three main-text models, a filtered pool across the categories 51.4% to 78.9%. Reblitz-Richardson (2026) arrives at a compatible picture from patching.

Table 1. Published measurements behind the argument, one cited result per row. The last column gives the claim of Section 2 to which the number is most relevant. Every value is taken from the source named on its row; the selection and grouping are the author's.

#	Source	Setting	Varied	Value	Claim
1	Venkatesh and Kurapath (2026)	Llama-3.1-8B, Qwen2.5-3B; 5 traits	unit orthogonal add	trait-classifier d 0.119 / 0.131	C4
2	Venkatesh and Kurapath (2026)	same	mean-diff vs PCA, cos -0.54 to 0.32	all d below 0.27	C4
3	Venkatesh and Kurapath (2026)	same, App. B	norm-matched orthogonal, logits	54-105% of v's own shift (author's ratio)	C2, C4
4	Joad et al. (2026)	3 chat models, 11 splits	category direction, add	cos 0.4-0.6; one trade-off	C4
5	Joad et al. (2026)	same	category direction, ablate	0.0-73.5% in main table (App.: 91.8% max); 33-prompt round-robin pool 35.0-45.8%, filtered cross-category pool 51.4-78.9% (5-seed means); best recipe 69.5-87.1%	C3, C4
6	Braun (2026)	Llama-2-7B-Chat, 36 sets	7 prompt types	distinct vectors, similar means	C4, C5
7	Braun (2026)	same	per-sample readout	29-43% anti-steerable (3-50%)	C1
8	Wollschläger et al. (2025)	Gemma 2, Qwen2.5, Llama 3	cone dimension	cones up to 5 dims	C4
9	Korznikov et al. (2025)	Llama3.1, Qwen2.5, Falcon3	random direction	harmful compliance 0% to 1-13%	C2
10	Piras et al. (2026)	safety-aligned LMs	SOM directions, ablate	beats single direction	C3, C4
11	Marshall et al. (2024)	10 models to 70B	projection vs affine	projection incoherent in an RWKV v5 RNN	C3, C6
12	Wu et al. (2025) AxBench	Gemma-2 2B / 9B	probe vs steer	AUROC 0.940; steer 0.098	C1 vs C5
13	Wang et al. (2026)	Qwen2.5-Omni-7B	probe-best vs mid-band injection	about 7x weaker at probe-best (single layer); about 26x vs multi-layer mid-band	C1 vs C5
14	Xun (2026)	Pythia 160M-12B	probe direction, steer	AUROC 0.990 + ; null in 43/48	C1 vs C5
15	Galeone et al. (2026)	4 models, 1B-9B	detect vs control direction	cos 0.12-0.20	C1 vs C5
16	Tan et al. (2024)	Llama-2-7b, Qwen-1.5-14b	ID vs OOD	rho 0.891 / 0.694	C1
17	Subramani et al. (2022)	pretrained decoders	vector per sentence	above 99 BLEU	C2
18	Zhong and Li (2026)	Llama-3.1-8B, Qwen2.5-7B	persona direction	refusal 97% to 2%	C5, C6
19	Tigges et al. (2023)	Pythia-2.8b, SST subset with commas	ablate sentiment dir.	76% of above-chance acc. lost	C3, C6
20	Wang et al. (2025a)	14 languages	cross-lingual transfer	near-complete bypass	C5
21	Soligo et al. (2025)	Qwen2.5-14B fine-tunes	cross-fine-tune ablation	ablates misalignment	C5
22	Li et al. (2023a)	Alpaca, TruthfulQA	ITI head shift	32.5% to 65.1%	C1
23	Sutter et al. (2025)	random-init LMs, IOI	nonlinear alignment map	100% IIA	C4

Table 2. The six claims, the evidence each needs, and the status of the published record.

Claim	Evidence that licenses it	Rules out	Status in the record
C1 Control	per-sample effect of adding v, downstream readout	nothing about why	well supported; per-sample reliability poor
C2 Specificity	norm-matched random, orthogonal, second-D controls	generic perturbation	rarely run; separates under fine readouts
C3 Necessity	ablation with off-target coherence check	sufficiency illusions	partial for refusal and sentiment
C4 Uniqueness	one ray under the finest readout	multiplicity	no behaviour; contradicted for refusal
C5 Content	second contrast set holding confounds fixed; transfer	contrast-set confound	partial: languages, fine-tunes, harm vs refusal
C6 Use	predicts unsteered behaviour, errors, training shifts	off-distribution damage	partial: persona, function vectors, sentiment

Persona vectors. Chen et al. (2025) establish C1 for several traits and report evidence of the C6 kind: shifts along the vectors during fine-tuning track the trait changes fine-tuning produces. That is a stronger kind of evidence than efficacy, because the vector was not built

from the fine-tuning data. Uniqueness is not tested in that paper, and Zhong and Li (2026) show that a compliant-persona direction gates refusal downstream, which means persona and refusal directions are not independent handles. The licensed claim is that a

Algorithm 1. Upgrading a steering result from control toward representation

```

Input: vector  $v$  at site  $s$ ; magnitude range  $A$ ; contrast set  $D$ ; extraction rule  $E$ ;
       selection rule  $S$ ; readouts  $R_{\text{task}}$  (task score) and  $R_{\text{fine}}$  (next-token distribution)
1  CONTROL: add  $a*v$  ( $a$  in  $A$ ) on held-out inputs; report per-sample effects under  $R_{\text{task}}$ 
   and the fraction moved the wrong way. If the mean effect is null: stop, no claim.
2  SELECTION: report the pool  $v$  was chosen from; re-select on one split and score on
   another. Report  $C1$  for the held-out choice, not for the best-of-pool.
3  SPECIFICITY: repeat step 1 with norm-matched random directions, norm-matched
   directions orthogonal to  $v$ , and a vector from a second contrast set  $D2$ .
   Score every condition under  $R_{\text{task}}$  and  $R_{\text{fine}}$ .
   If  $v$  is not separable from random under  $R_{\text{fine}}$ : license  $C1$  only; stop.
4  NECESSITY: project out  $v$  (and mean-replace) on inputs where the unsteered model
   shows the behaviour; measure off-target coherence at the same time.
   If the behaviour survives: license  $C1$  and  $C2$  only.
   If coherence collapses: necessity is confounded with damage; license  $C1$  and  $C2$ .
5  UNIQUENESS: estimate  $Q = \{u : R_{\text{fine}}(\text{model} + u) \text{ matches } R_{\text{fine}}(\text{model} + v)\}$ 
   by orthogonal search, alternative extraction rules and cone search.
   If  $Q$  is one ray up to scale: license  $C4$ .
   Else report  $\dim(Q)$ ; state claims about the subspace  $Q$ , not about  $v$ .
6  CONTENT: build  $D3$  holding the suspected confounds of  $D$  fixed (answer position,
   register, harm versus refusal, persona). Re-extract. If the new direction lies in  $Q$ 
   and  $v$  transfers across  $D3$ , languages or fine-tunes: license  $C5$  for  $Q$ .
7  USE: test whether the projection onto  $Q$  predicts unsteered behaviour, the model's
   errors about the property, and behaviour changes under training. If so: license  $C6$ .
Output: the strongest claim whose test passed, with readout, site, magnitude range and
       contrast sets stated alongside it.

```

persona subspace exists whose occupancy predicts trait expression and its changes, not that each trait has one direction.

Truth directions. Marks and Tegmark (2024) combine three routes - visualisation, probe transfer across datasets and intervention - and report that difference-in-means directions are the most causally implicated. This reaches step 6 for content across their curated datasets. Levinstein and Herrmann (2024) report that earlier lie-detection probes fail to generalise in basic ways, and Farquhar et al. (2023) show that an unsupervised truth-finding method can lock onto whatever feature is most salient, so content claims beyond curated statement sets are open. The misrepresentation test of step 7 - whether the direction follows the model's own mistaken judgments rather than ground truth - is the decisive one, and the record does not contain it in a form that settles the question.

12. What follows for safety uses of steering

The distinction between control and representation is not academic for safety work, because different safety uses need different rungs of Algorithm 1.

Attacks and red-teaming need only control. Removing a refusal direction to disable safety behaviour works whether or not the direction is the model's refusal variable, and Yu et al. (2025) build an adversarial-training defence on exactly this: input-level jailbreaks share a mechanism that ablates a refusal feature, refusal feature ablation approximates the worst-case perturbation, and simulating it during training improves robustness across three models at much lower cost than conventional adversarial training. A defence that trains against $C1$ and $C3$ effects needs no $C4$ claim. Coalson et al. (2026) take the multiplicity results as a design principle: since a jailbreak that suppresses one dominant

refusal feature can collapse alignment, they iteratively ablate learned refusal directions during training so that the model rebuilds refusal along new, causally independent directions, and report robustness gains across four jailbreak attacks. Their method treats non-uniqueness as something to engineer rather than a defect to explain away.

Precision claims need specificity. Steering is often presented as a precise, interpretable and therefore safer alternative to fine-tuning. Korznikov et al. (2025) report that random directions and benign sparse-autoencoder features both raise harmful compliance, with SAE features 1 to 4 percentage points above random directions under identical conditions. Cao et al. (2026) report that activation steering can induce broad emergent misalignment, with steered outputs more coherent and semantically relevant than those of fine-tuned counterparts. van der Weij et al. (2024) find that combining steering vectors for different behaviours into one vector is largely unsuccessful, while injecting separate vectors at different places works better. Lee et al. (2025) make steering conditional on a separate activation-based check of the input, which is control engineering on top of $C1$ and does not rely on $C4$. None of these results says that steering cannot be made safe. They say that the precision of a steering intervention is an empirical property of the intervention under stated controls, not a consequence of the direction's being "the" concept.

Monitoring needs content and use. Using a direction as a monitor - reading persona drift, deception or evaluation awareness off activations - needs $C5$ and $C6$ together with robustness of the reading, and it needs them under distribution shift. The reading-versus-writing dissociations of Section 7 cut both ways here: a direction can be a good monitor without being a good handle, as

AxBench's probes are, and a good handle without being a good monitor. The author's earlier analysis of refusal as a context-conditioned policy rather than a rate (Mahendrakar, 2026) argued that a binary refusal readout hides the distinctions a refusal policy has to make; Section 4.2 shows the same readout hiding the distinctions among refusal directions.

13. What would settle the open part

Five studies would move the open questions to answered ones. None requires a new method; each requires reporting a comparison that the published papers already have the ingredients for.

Study 1: a readout ladder for equivalence. For a fixed steering vector and a fixed set of norm-matched perturbations (orthogonal, random, alternative extraction rules, alternative contrast sets), report equivalence under a ladder of readouts: a trait classifier, an LLM judge, multiple-choice probability, top-k next-token overlap, the full next-token KL divergence, and a divergence over sampled continuations. The prediction from Section 4 is that the equivalence class shrinks monotonically down the ladder. If it does not - if vectors equivalent under a trait classifier are also equivalent in full next-token distribution at matched norm - the non-identifiability result is a model property and not a readout property, and the deflationary reading wins outright.

Study 2: an addition-removal matrix. For K directions reported for the same behaviour, report a K by K table of effects: adding direction i, removing direction j, with an off-target coherence measure in each cell. Section 5 predicts that the addition rows are nearly uniform and the removal columns are not. The matrix would say directly whether necessity concentrates on a core and how large the core is.

Study 3: held-out selection. Select the steering vector on one split by the usual criteria, then score it on a second split, and report the success rate as a function of the size of the candidate pool. This turns the outcome-selection concern of Section 6 into a number and separates "a direction exists" from "this procedure finds it". AxBench already does this for the steering factor: "We split the instructions into two equal sets - one for selecting the best factor and the other for evaluation" (Wu et al., 2025). That is a partial instance of step 2 of Algorithm 1; what remains is to apply the same split to the choice of the vector itself and to report it against the size of the pool.

Study 4: confound-controlled contrast sets at scale. Extend the four-bucket design of Joad et al. (2026) - varying harm and refusal independently - to other behaviours: truthfulness against confidence, sycophancy against agreement, persona against topic. For each, report whether the direction from the confounded contrast set lies in the subspace found when the

confound is held fixed. This is the direct test of C5.

Study 5: a misrepresentation test. Choose inputs on which the model is reliably wrong about the property - statements it treats as true that are false, harmless requests it classifies as harmful - and ask whether the projection onto the direction follows the model's judgment or the ground truth, and whether steering along it moves the model's judgment on those inputs. A direction that tracks the model's errors is the model's variable; one that tracks the truth on its errors is a probe's variable that the model does not consult. Galeone et al. (2026), who compare the directions that detect and that control the handling of fake entities, supply one ingredient of this design but not the test itself.

The five studies together would also say whether the subspace descriptions that are emerging for refusal - a shared core plus category- and style-specific directions (Joad et al., 2026), cones with representationally independent axes (Wollschlager et al., 2025), a harm-keyed slice embedded in a wider moral subspace (Reblitz-Richardson, 2026) - describe the same object.

14. What this analysis does not show

The analysis has five limits that a reader should weigh.

First, it re-reads published results; every number in it was reported by a cited source, and the consolidation, the six-claim decomposition, the readout argument and Algorithm 1 are the author's synthesis. Where the author computed a ratio from published values - the 54% to 105% range in Table 1, row 3 - the computation is identified as such.

Second, the readout argument in Section 4.1 compares two tables in Venkatesh and Kurapath (2026) that do not use the same perturbation: the classifier-level tests use random unit-norm orthogonal vectors, and the logit-level test uses norm-matched ones. The conclusion drawn is accordingly limited: the paper's own logit-level data show no equivalence for norm-matched orthogonal perturbations, and its classifier-level equivalence is for perturbations whose size relative to the steering vector is not reported. The paper asserts classifier equivalence at comparable norm but reports no classifier measurement for the norm-matched case, and nobody has reported both readouts on the same perturbation.

Third, much of the critical evidence is recent and from single groups: a master's thesis (Braun, 2026) and several 2026 preprints (Venkatesh and Kurapath; Galeone et al.; Xun; Wang et al.; Riegler and Torpmann-Hagen; Reblitz-Richardson), along with a workshop paper (Zhong and Li, 2026). Joad et al. (2026), which has been accepted at EMNLP 2026, and Wollschlager et al. (2025) are the strongest critical results by the care of their controls. If several of the single-group results fail to replicate, the reading-writing dissociation in Section 7 weakens more than the multiplicity argument in Sections 4 and 5 does,

because the latter rests on the two stronger papers.

Fourth, the scope is additive and projective interventions on the residual stream or attention-head outputs of language models. Sparse-autoencoder latents enter only where they are used as steering directions, and weight edits, fine-tuning-based interventions and nonlinear steering methods are discussed only as they bear on the linearity premise. The analysis says nothing about whether language models have concepts in any richer sense; it is about what a particular kind of experiment can show.

Fifth, the analysis does not argue that steering is unreliable as engineering. It argues that the evidence for steering as engineering and the evidence for steering as interpretation are different evidence, and that the second has been borrowed from the first.

15. Conclusion

A steering vector that works is a control knob: an intervention, at a site and a magnitude, that moves a behaviour as a given readout scores it. The literature has treated working knobs as concepts, and the critical literature has shown that knobs are cheap - many directions turn the same dial, random ones turn it a little, and the direction that reads a property best is often not the one that moves it. Both literatures are right about what they measured. The equivalence that makes knobs look interchangeable is set by the readout that scores them, and finer readouts in the same papers tell many of them apart. Adding a direction tests sufficiency, which many directions pass; removing it tests necessity, which few pass alone; and the single-direction claim for refusal failed on removal and on content, not on control. What survives as evidence for representation is invariance - across extraction routes, contrast sets, languages, fine-tuning runs and the model's own errors - and that evidence is available, in pieces, for a handful of behaviours. The rule that follows is to state a steering result at the rung of Algorithm 1 that its controls reach: control is well supported, necessity is supported for a few behaviours, content for fewer, and uniqueness for none.

References

- Arad, D., Mueller, A., & Belinkov, Y. (2025). SAEs Are Good for Steering – If You Select the Right Features. arXiv:2505.20063. doi:10.48550/arXiv.2505.20063. [PREPRINT]
- Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in Language Models Is Mediated by a Single Direction. arXiv:2406.11717. doi:10.48550/arXiv.2406.11717. [PREPRINT]
- Belinkov, Y. (2021). Probing Classifiers: Promises, Shortcomings, and Advances. Computational Linguistics. arXiv:2102.12452. doi:10.48550/arXiv.2102.12452.
- Bigoulaeva, I., Rohweder, J., Dutta, S., & Gurevych, I. (2026). Patches of Nonlinearity: Instruction Vectors in Large Language Models. ACL 2026. arXiv:2602.07930. doi:10.48550/arXiv.2602.07930.
- Blank, C., Bhatia, A., Rajamanoharan, S., Conmy, A., & Nanda, N. (2026). Subliminal Learning Is Steering Vector Distillation. arXiv:2606.00995. doi:10.48550/arXiv.2606.00995. [PREPRINT]
- Bolukbasi, T., Pearce, A., Yuan, A., Coenen, A., Reif, E., Viegas, F., & Wattenberg, M. (2021). An Interpretability Illusion for BERT. arXiv:2104.07143. doi:10.48550/arXiv.2104.07143. [PREPRINT]
- Braun, J. (2026). Understanding Unreliability of Steering Vectors in Language Models: Geometric Predictors and the Limits of Linear Approximations. Master's thesis, University of Tuebingen. arXiv:2602.17881. doi:10.48550/arXiv.2602.17881.
- Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2023). Discovering Latent Knowledge in Language Models Without Supervision. ICLR 2023. arXiv:2212.03827. doi:10.48550/arXiv.2212.03827.
- Cao, Q., Lou, J., Liu, M., Feng, W., Li, D., Ng, S.-K., & Luu, A. T. (2026). Activation Steering Induces Emergent Misalignment: A More Comprehensive Evaluation. arXiv:2606.08682. doi:10.48550/arXiv.2606.08682. [PREPRINT]
- Chen, R., Arditi, A., Sleight, H., Evans, O., & Lindsey, J. (2025). Persona Vectors: Monitoring and Controlling Character Traits in Language Models. arXiv:2507.21509. doi:10.48550/arXiv.2507.21509. [PREPRINT]
- Coalson, Z., Sohler, B., Gabriel, A., & Hong, S. (2026). Fail-Closed Alignment for Large Language Models. arXiv:2602.16977. doi:10.48550/arXiv.2602.16977. [PREPRINT]
- Csordas, R., Potts, C., Manning, C. D., & Geiger, A. (2024). Recurrent Neural Networks Learn to Store and Generate Sequences using Non-Linear Representations. arXiv:2408.10920. doi:10.48550/arXiv.2408.10920. [PREPRINT]
- Da Silva, P. Q., Sethuraman, H., Rajagopal, D., Hajishirzi, H., & Kumar, S. (2025). Steering off Course: Reliability Challenges in Steering Language Models. arXiv:2504.04635. doi:10.48550/arXiv.2504.04635. [PREPRINT]
- Elazar, Y., Ravfogel, S., Jacovi, A., & Goldberg, Y. (2021). Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. Transactions of the Association for Computational Linguistics. arXiv:2006.00995. doi:10.48550/arXiv.2006.00995.
- Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., et al. (2022). Toy Models of Superposition. arXiv:2209.10652. doi:10.48550/arXiv.2209.10652. [PREPRINT]
- Engels, J., Michaud, E. J., Liao, I., Gurnee, W., & Tegmark, M. (2025). Not All Language Model Features Are One-Dimensionally Linear. ICLR 2025. arXiv:2405.14860. doi:10.48550/arXiv.2405.14860.
- Farquhar, S., Varma, V., Kenton, Z., Gasteiger, J., Mikulik, V., & Shah, R. (2023). Challenges with unsupervised LLM knowledge discovery. arXiv:2312.10029. doi:10.48550/arXiv.2312.10029. [PREPRINT]
- Galeone, C., Ettorre, A., Park, M., Ettorre, G., & Ligorio, D. (2026). Perfect Detection, Failed Control: The Geometry of Knowing vs. Steering in Language Models. arXiv:2606.24952. doi:10.48550/arXiv.2606.24952. [PREPRINT]
- Gao, L., Zhang, J., Liu, W., Ji, F., Wang, C., Song, Z., et al. (2026). The Cylindrical Representation Hypothesis for Language Model Steering. ICML 2026. arXiv:2605.01844. doi:10.48550/arXiv.2605.01844.
- Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., et al. (2023). Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability. arXiv:2301.04709. doi:10.48550/arXiv.2301.04709. [PREPRINT]
- Harding, J. (2023). Operationalising Representation in Natural Language Processing. The British Journal for the Philosophy of Science. doi:10.1086/728685. arXiv:2306.08193.
- Hase, P., Bansal, M., Kim, B., & Ghandeharioun, A. (2023). Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models. NeurIPS 2023. arXiv:2301.04213. doi:10.48550/arXiv.2301.04213.

- Hendel, R., Geva, M., & Globerson, A. (2023). In-Context Learning Creates Task Vectors. Findings of EMNLP 2023. arXiv:2310.15916. doi:10.48550/arXiv.2310.15916.
- Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., & Bau, D. (2023). Linearity of Relation Decoding in Transformer Language Models. arXiv:2308.09124. doi:10.48550/arXiv.2308.09124. [PREPRINT]
- Hewitt, J., & Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. EMNLP 2019. arXiv:1909.03368. doi:10.48550/arXiv.1909.03368.
- Hyvarinen, A., & Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. Neural Networks. doi:10.1016/S0893-6080(98)00140-3.
- Im, S., & Li, S. (2025). A Unified Understanding and Evaluation of Steering Methods. arXiv:2502.02716. doi:10.48550/arXiv.2502.02716. [PREPRINT]
- Jiang, Y., Rajendran, G., Ravikumar, P., Aragam, B., & Veitch, V. (2024). On the Origins of Linear Representations in Large Language Models. arXiv:2403.03867. doi:10.48550/arXiv.2403.03867. [PREPRINT]
- Joad, F., Hawasly, M., Boughorbel, S., Durrani, N., & Sencar, H. T. (2026). There Is More to Refusal in Large Language Models than a Single Direction. EMNLP 2026. arXiv:2602.02132. doi:10.48550/arXiv.2602.02132.
- Korznikov, A., Galichin, A., Dontsov, A., Rogov, O. Y., Oseledets, I., & Tutubalina, E. (2025). The Rogue Scalpel: Activation Steering Compromises LLM Safety. arXiv:2509.22067. doi:10.48550/arXiv.2509.22067. [PREPRINT]
- Lee, B. W., Padhi, I., Natesan Ramamurthy, K., Miehl, E., Dognin, P., Nagireddy, M., & Dhurandhar, A. (2025). Programming Refusal with Conditional Activation Steering. ICLR 2025. arXiv:2409.05907. doi:10.48550/arXiv.2409.05907.
- Levinstein, B. A., & Herrmann, D. A. (2024). Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks. Philosophical Studies. doi:10.1007/s11098-023-02094-3. arXiv:2307.00175.
- Li, K., Patel, O., Viegas, F., Pfister, H., & Wattenberg, M. (2023a). Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. NeurIPS 2023. arXiv:2306.03341. doi:10.48550/arXiv.2306.03341.
- Li, K., Hopkins, A. K., Bau, D., Viegas, F., Pfister, H., & Wattenberg, M. (2023b). Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. ICLR 2023. arXiv:2210.13382. doi:10.48550/arXiv.2210.13382.
- Lin, Z., & Liu, F. (2026). Position: Mechanistic Interpretability Must Disclose Identification Assumptions for Causal Claims. arXiv:2605.08012. doi:10.48550/arXiv.2605.08012. [PREPRINT]
- Mahendrakar, P. (2026). Refusal Is Not a Rate: What a Context-Conditioned Refusal Policy Would Have to Specify, Why the Benchmarks That Score Context Assume It Is Verified, and the Provenance Term No Instrument Prices. Life of Research. doi:10.5281/zenodo.22984335.
- Makelov, A., Lange, G., & Nanda, N. (2023). Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. arXiv:2311.17030. doi:10.48550/arXiv.2311.17030. [PREPRINT]
- Marconato, E., Lachapelle, S., Weichwald, S., & Gresele, L. (2025). All or None: Identifiable Linear Properties of Next-token Predictors in Language Modeling. AISTATS 2025. arXiv:2410.23501. doi:10.48550/arXiv.2410.23501.
- Marks, S., & Tegmark, M. (2024). The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. COLM 2024. arXiv:2310.06824. doi:10.48550/arXiv.2310.06824.
- Marshall, T., Scherlis, A., & Belrose, N. (2024). Refusal in LLMs is an Affine Function. arXiv:2411.09003. doi:10.48550/arXiv.2411.09003. [PREPRINT]
- Meloux, M., Maniu, S., Portet, F., & Peyrard, M. (2025). Everything, Everywhere, All at Once: Is Mechanistic Interpretability Identifiable? ICLR 2025. arXiv:2502.20914. doi:10.48550/arXiv.2502.20914.
- Mueller, A., Brinkmann, J., Li, M., Marks, S., Pal, K., Prakash, N., et al. (2025). The Quest for the Right Mediator: Surveying Mechanistic Interpretability Through the Lens of Causal Mediation Analysis. arXiv:2408.01416. doi:10.48550/arXiv.2408.01416. [PREPRINT]
- Nanda, N., Lee, A., & Wattenberg, M. (2023). Emergent Linear Representations in World Models of Self-Supervised Sequence Models. arXiv:2309.00941. doi:10.48550/arXiv.2309.00941. [PREPRINT]
- Pan, W., Liu, Z., Chen, Q., Zhou, X., Yu, H., & Jia, X. (2025). The Hidden Dimensions of LLM Alignment: A Multi-Dimensional Analysis of Orthogonal Safety Directions. ICML 2025. arXiv:2502.09674. doi:10.48550/arXiv.2502.09674.
- Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. M. (2023). Steering Llama 2 via Contrastive Activation Addition. arXiv:2312.06681. doi:10.48550/arXiv.2312.06681. [PREPRINT]
- Park, K., Choe, Y. J., & Veitch, V. (2024). The Linear Representation Hypothesis and the Geometry of Large Language Models. ICML 2024. arXiv:2311.03658. doi:10.48550/arXiv.2311.03658.
- Park, K., Nief, T., Choe, Y. J., & Veitch, V. (2026). The Information Geometry of Softmax: Probing and Steering. ICML 2026. arXiv:2602.15293. doi:10.48550/arXiv.2602.15293.
- Piras, G., Mura, R., Brau, F., Oneto, L., Roli, F., & Biggio, B. (2026). SOM Directions are Better than One: Multi-Directional Refusal Suppression in Language Models. AAAI 2026. arXiv:2511.08379. doi:10.48550/arXiv.2511.08379.
- Pres, I., Ruis, L., Lubana, E. S., & Krueger, D. (2024). Towards Reliable Evaluation of Behavior Steering Interventions in LLMs. NeurIPS 2024 Workshop on Foundation Model Interventions. arXiv:2410.17245. doi:10.48550/arXiv.2410.17245.
- Rajendran, G., Buchholz, S., Aragam, B., Schoelkopf, B., & Ravikumar, P. (2024). Learning Interpretable Concepts: Unifying Causal Representation Learning and Foundation Models. NeurIPS 2024. arXiv:2402.09236. doi:10.48550/arXiv.2402.09236.
- Raval, S., Song, H. J., Wu, L., Harrasse, A., Phillips, J. M., Barez, F., & Abdullah, A. (2026). Curveball Steering: The Right Direction To Steer Isn't Always Linear. arXiv:2603.09313. doi:10.48550/arXiv.2603.09313. [PREPRINT]
- Ravichander, A., Belinkov, Y., & Hovy, E. (2021). Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? EACL 2021. arXiv:2005.00719. doi:10.48550/arXiv.2005.00719.
- Reblitz-Richardson, O. (2026). Refusal Reads Only a Slice of What the Model Knows: Harm-Keyed Routing and Its Exceptions Across Model Families. arXiv:2609.14759. doi:10.48550/arXiv.2609.14759. [PREPRINT]
- Riegler, M. A., & Torpmann-Hagen, B. S. F. (2026). Steering grids for sparse-autoencoder features: when a top-context label names an activation regime rather than a causal axis. arXiv:2605.03160. doi:10.48550/arXiv.2605.03160. [PREPRINT]
- Roeder, G., Metz, L., & Kingma, D. P. (2020). On Linear Identifiability of Learned Representations. arXiv:2007.00810. doi:10.48550/arXiv.2007.00810. [PREPRINT]
- Singh, S., Ravfogel, S., Herzig, J., Aharoni, R., Cotterell, R., & Kumaraguru, P. (2024). Representation Surgery: Theory and Practice of Affine Steering. ICML 2024. arXiv:2402.09631. doi:10.48550/arXiv.2402.09631.
- Soligo, A., Turner, E., Rajamanoharan, S., & Nanda, N. (2025). Convergent Linear Representations of Emergent Misalignment. arXiv:2506.11618. doi:10.48550/arXiv.2506.11618. [PREPRINT]
- Stolfo, A., Balachandran, V., Yousefi, S., Horvitz, E., & Nushi, B. (2025). Improving Instruction-Following in Language Models through Activation Steering. ICLR 2025. arXiv:2410.12877. doi:10.48550/arXiv.2410.12877.

- Subramani, N., Suresh, N., & Peters, M. E. (2022). Extracting Latent Steering Vectors from Pretrained Language Models. *Findings of ACL 2022*. arXiv:2205.05124. doi:10.48550/arXiv.2205.05124.
- Sutter, D., Minder, J., Hofmann, T., & Pimentel, T. (2025). The Non-Linear Representation Dilemma: Is Causal Abstraction Enough for Mechanistic Interpretability? *NeurIPS 2025*. arXiv:2507.08802. doi:10.48550/arXiv.2507.08802.
- Tan, D., Chanin, D., Lynch, A., Kanoulas, D., Paige, B., Garriga-Alonso, A., & Kirk, R. (2024). Analyzing the Generalization and Reliability of Steering Vectors. arXiv:2407.12404. doi:10.48550/arXiv.2407.12404. [PREPRINT]
- Tigges, C., Hollinsworth, O. J., Geiger, A., & Nanda, N. (2023). Linear Representations of Sentiment in Large Language Models. arXiv:2310.15154. doi:10.48550/arXiv.2310.15154. [PREPRINT]
- Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, B. C., & Bau, D. (2024). Function Vectors in Large Language Models. *ICLR 2024*. arXiv:2310.15213. doi:10.48550/arXiv.2310.15213.
- Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., & MacDiarmid, M. (2023). Steering Language Models With Activation Engineering. arXiv:2308.10248. doi:10.48550/arXiv.2308.10248. [PREPRINT]
- van der Weij, T., Poesio, M., & Schoots, N. (2024). Extending Activation Steering to Broad Skills and Multiple Behaviours. arXiv:2403.05767. doi:10.48550/arXiv.2403.05767. [PREPRINT]
- Venkatesh, S., & Kurupath, A. M. (2026). On the Non-Identifiability of Steering Vectors in Large Language Models. arXiv:2602.06801. doi:10.48550/arXiv.2602.06801. [PREPRINT]
- Wang, X., Wang, M., Liu, Y., Schuetze, H., & Plank, B. (2025a). Refusal Direction is Universal Across Safety-Aligned Languages. arXiv:2505.17306. doi:10.48550/arXiv.2505.17306. [PREPRINT]
- Wang, M., Dupre la Tour, T., Watkins, O., Makelov, A., Chi, R. A., Miserendino, S., et al. (2025b). Persona Features Control Emergent Misalignment. arXiv:2506.19823. doi:10.48550/arXiv.2506.19823. [PREPRINT]
- Wang, Y., Dang, J., Wang, B., Wu, Y., Zhang, W., Peng, H., Liu, J., Hu, B., Tian, Q., & Chua, T.-S. (2026). Read-Best Is Not Steer-Best: A Probing--Steering Layer Dissociation in Omni-Modal Large Language Models. arXiv:2609.22135. doi:10.48550/arXiv.2609.22135. [PREPRINT]
- Wollschlaeger, T., Elstner, J., Geisler, S., Cohen-Addad, V., Gunnemann, S., & Gasteiger, J. (2025). The Geometry of Refusal in Large Language Models: Concept Cones and Representational Independence. arXiv:2502.17420. doi:10.48550/arXiv.2502.17420. [PREPRINT]
- Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press. doi:10.1093/0195155270.001.0001.
- Wu, Z., Geiger, A., Huang, J., Arora, A., Icard, T., Potts, C., & Goodman, N. D. (2024). A Reply to Makelov et al. (2023)'s "Interpretability Illusion" Arguments. arXiv:2401.12631. doi:10.48550/arXiv.2401.12631. [PREPRINT]
- Wu, Z., Arora, A., Geiger, A., Wang, Z., Huang, J., Jurafsky, D., Manning, C. D., & Potts, C. (2025). AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. arXiv:2501.17148. doi:10.48550/arXiv.2501.17148. [PREPRINT]
- Xun, X. (2026). Lagged Coupling: Internal Representations Become Readable Before They Become Causal. arXiv:2609.01048. doi:10.48550/arXiv.2609.01048. [PREPRINT]
- Yu, L., Do, V., Hambardzumyan, K., & Cancedda, N. (2025). Robust LLM safeguarding via refusal feature adversarial training. arXiv:2409.20089. doi:10.48550/arXiv.2409.20089. [PREPRINT]
- Zhong, V., & Li, Q. (2026). Refusal Lives Downstream of Persona in Chat Models. *ICML 2026 Mechanistic Interpretability Workshop*. arXiv:2606.26161. doi:10.48550/arXiv.2606.26161.
- Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405. doi:10.48550/arXiv.2310.01405. [PREPRINT]