EVIDENCE CONSISTENT WITH FEATURE-SPECIFIC ERROR CORRECTION IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the features of large language models (LLMs) is a central goal of interpretability. LLMs are commonly assumed to use superposition to represent more features than they have dimensions. They may not only represent features in superposition but also perform computation in superposition. Theory predicts that computing in superposition requires error correction that privileges feature directions over generic ones, but this prediction has not been tested empirically. We propose an empirical test of error correction in LLMs based on activation perturbations. Perturbing residual-stream activations, we find that they are robust to perturbations, forming activation plateaus consistent with error correction. They are also less robust along candidate feature directions ("pure" directions, constructed from contrastive prompt pairs) than along mixtures of two such directions, indicating that the pure directions are privileged. We quantify this privilegedness by modeling the perturbation effect as a function of the L^p-norm of its decomposition into feature components. For $p = 2$ the response is a quadratic form with at most as many nonzero eigenvalues as the residual-stream dimension, which cannot privilege the many feature directions superposition requires. $p > 2$ lifts this constraint and is consistent with feature-specific error correction. We find $p > 2$ for contrastive, MELBO, and SAE-decoder directions, and $p \approx 2$ for random and PCA directions (controls). These results replicate across Gemma-2-9B, Qwen3-1.7B, Llama-3.1-8B, Mistral-7B-v0.3, Aya-Expanse-8B, and Yi-1.5-9B. We further validate our method on a toy model of error correction with known ground-truth features, recovering $p > 2$ for true feature directions, degrading toward 2 as we rotate away from them.[†]

1 INTRODUCTION

Representations in large language models (LLMs) are poorly understood. It is commonly assumed that LLMs make use of superposition (Elhage et al., 2022) to represent more concepts than they have dimensions available, and potentially to compute in superposition (CiS) (Hänni et al., 2024; Adler & Shavit, 2024; Olah et al., 2025). However, we have only indirect evidence for superposition, chiefly the success of sparse autoencoders (Cunningham et al., 2023; Gao et al., 2024; Templeton et al., 2026) at extracting interpretable directions, and only theoretical evidence for CiS (Hänni et al., 2024).

An empirical prediction of CiS is that neural networks must correct interference noise while preserving feature signal (Hänni et al., 2024). This requires networks to be less sensitive to perturbations along non-feature directions than along feature directions. We name this property feature-specific error correction (FSEC). While we cannot rule FSEC out or in without ground-truth feature directions, we can still ask whether FSEC-like behaviour occurs for any directions: if the model's error correction privileges certain directions, those directions are candidate features, and we can detect them by measuring sensitivity, requiring only a generic input to perturb rather than feature-specific labeled data.

We provide empirical evidence consistent with FSEC, showing that the robustness of LLM activations to perturbations privileges certain candidate feature directions over others. Concretely, we perturb residual-stream activations at early layers and measure the downstream response as a function of

[†]Code: https://anonymous.4open.science/r/fsec-anonymous-5C9D/

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

perturbation direction and magnitude. We construct candidate feature directions via contrastive means for a variety of concepts, including languages, programming languages, gender, sentiment, registers, and verb tenses. Across Gemma-2-9B (Team et al., 2024), Qwen3-1.7B (Yang et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), Aya-Expanse-8B (Dang et al., 2024), and Yi-1.5-9B (Young et al., 2024), we find that contrastive feature directions elicit a stronger downstream response than mixtures thereof, consistent with FSEC that privileges feature directions while suppressing interference along non-feature directions. We formalize this by modeling the LLMs' response to perturbation as an L^p norm of the perturbation's decomposition into candidate feature directions. The $p = 2$ case reduces to a basis-invariant quadratic form, meaning no choice of basis is privileged. $p > 2$ breaks this invariance, indicating that the candidate pure feature directions are more sensitive than their mixtures, as predicted by FSEC. We measure $p \approx 2.3$ for contrastive directions across models. MELBO (Mack & Turner, 2024b) and SAE directions, which also aim to recover model features, likewise yield $p > 2$, albeit with smaller values. PCA and random directions do not, consistent with the interpretation that $p > 2$ reflects alignment with the model's features. We validate this methodology in a toy model of error correction with ground-truth features, confirming that p degrades toward 2 as directions are misaligned with the true features.

Our contributions are:

1. We propose feature-specific error correction (FSEC) as a test of computation in superposition: we model the LLM's response to a perturbation as a function of the L^p-norm of the perturbation's decomposition into candidate feature directions, with FSEC predicting $p > 2$.
2. We find evidence consistent with FSEC for three types of candidate feature directions—contrastive, SAE-decoder, and MELBO—each with $p > 2$, and show the contrastive result replicates across six LLMs from different families. PCA and random directions yield $p \approx 2$, consistent with them not being privileged.
3. We show that FSEC occurs in a toy model of error correction (Vaintrob, 2026).

2 Related Work

Activation plateaus. Prior work has established that in-distribution activations of LLMs are resistant to perturbations (Heimersheim & Mendel, 2024; Janiak et al., 2024; Shinkle & Heimersheim, 2025), a phenomenon known as *activation plateaus*. We introduce a novel measurement of the activation plateau boundary geometry, and identify its connection to FSEC.

Direction-dependent sensitivity for feature finding. Prior work has exploited the fact that LLMs have direction-dependent sensitivity to perform unsupervised optimizations for directions maximizing downstream response, resulting in interpretable steering vectors, including MELBO (Mack & Turner, 2024b;a). We make use of the same phenomenon to empirically probe error correction, and apply our analysis to directions obtained in this way (Section 4.1). Unlike this line of work, our analysis adds the novel study of the sensitivity geometry and its connection to error correction.

Error correction for computation in superposition. Hänni et al. (2024) argue theoretically that computation in superposition requires error correction that privileges feature directions. We provide empirical evidence consistent with superposition and error correction occurring in LLMs.

3 Methodology

We probe error correction in LLMs by perturbing residual stream activations and measuring the downstream response. In-distribution activations are robust to perturbations below a threshold magnitude, a phenomenon known as *activation plateaus* (Heimersheim & Mendel, 2024; Shinkle & Heimersheim, 2025). This robustness is direction-dependent, i.e., the model is more sensitive to perturbations along some directions than others. Two lines of evidence suggest feature directions in particular are privileged: empirically, prior work recovers interpretable directions by optimizing for sensitivity, the inverse of robustness (Mack & Turner, 2024b;a); and theoretically, FSEC predicts that feature directions are privileged. We test this by comparing the downstream response along candidate feature directions to the response along non-feature baselines.

In all experiments, we perturb the residual stream at an early layer ℓ (default $\ell = 2$) and measure the downstream response at the second-to-last layer, maximizing the number of intervening layers. This is because activation plateaus are known to be more pronounced the greater the distance between perturbation and measurement (Shinkle & Heimersheim, 2025). We avoid the last residual stream layer, which is known to behave atypically. The downstream response is computed by patching the perturbed activation $\mathbf{a}(\alpha)$ back into the model (Meng et al., 2022; Heimersheim & Nanda, 2024), performing a forward pass, and taking the L^2 distance between the perturbed and unperturbed residual streams at the measurement layer. We show in Section 4.2 that our results are robust to varying the perturbation and measurement layers and the perturbation token position, and hold also when measuring cosine distance or KL-divergence in the logits.

To distinguish directional effects from those of magnitude, we follow prior perturbation analyses of activation plateaus (Heimersheim & Mendel, 2024) and perturb by rotating the activation vector $\mathbf{a}$ towards the perturbation direction $\mathbf{d}$ while keeping the activation's norm constant. We refer to this as a *norm-matched perturbation*. Concretely, a perturbation of angle α of $\mathbf{a}$ toward $\mathbf{d}$ is:

$$\mathbf{a}(\alpha) = \cos(\alpha)\mathbf{a} + \sin(\alpha)\,\|\mathbf{a}\|\,\frac{\mathbf{d}_\perp}{\|\mathbf{d}_\perp\|}, \tag{1}$$

where $\mathbf{d}_\perp$ is the component of $\mathbf{d}$ that is orthogonal to $\mathbf{a}$.

We quantify the model's sensitivity along a given direction as the *plateau-breaking angle*, the smallest perturbation angle for which the downstream response exceeds a threshold T. We set T per direction-pair, at a level that is guaranteed to be crossed by both single-axis sweeps (see Appendix B for why a global threshold fails). We aggregate the downstream response across a fixed set of $N = 30$ inputs (*anchors*) $\{\mathbf{x}_n\}_{n=1}^N$—last-token residual-stream activations of 5-token FineWeb-Edu prompts (Penedo et al., 2024)—by taking the median over anchors. Writing $L^2(\mathbf{x}_n, \alpha; \mathbf{d})$ for the response of anchor $\mathbf{x}_n$ when perturbed by angle α along direction $\mathbf{d}$, the single-axis response curve is

$$L^2(\alpha; \mathbf{d}) = \underset{n=1,\ldots,N}{\text{median}}\, L^2(\mathbf{x}_n, \alpha; \mathbf{d}). \tag{2}$$

$\max_\alpha L^2(\alpha; \mathbf{d}_i)$ is then the largest median response attained when sweeping along $\mathbf{d}_i$ alone. All sweeps cover $\alpha \in [0°, 60°]$ in $2°$ steps, so these maxima are taken over that range. For a pair $(\mathbf{d}_1, \mathbf{d}_2)$ we set the threshold

$$T = f \cdot \min\big(\max_\alpha L^2(\alpha; \mathbf{d}_1),\ \max_\alpha L^2(\alpha; \mathbf{d}_2)\big), \tag{3}$$

a fraction f of the smaller of the two single-axis maxima. Taking the smaller maximum ensures both axes reach T, so the single-axis plateau-breaking angles α_1, α_2 that calibrate the fit are always defined. Throughout we use $f = 0.5$. We show in Section 4.2 that our results are robust to varying T within half-to-double its nominal value.

We perform perturbations along six types of directions: contrastive (Panickssery et al., 2023; Turner et al., 2023), MELBO (Mack & Turner, 2024b), SAE latents (Lieberum et al., 2024), PCA directions, random, and random-difference directions. The first three are all candidate feature directions, while PCA, random, and random-difference directions function as non-feature baselines. PCA directions are computed by performing PCA on the residual-stream activations arising from a randomly selected sample of 10000 5-token-long FineWeb-Edu inputs (Penedo et al., 2024). Random directions are sampled isotropically on the unit sphere, $\hat{\mathbf{d}}_{\text{rand}} \sim \text{Unif}(\mathcal{S}^{d-1})$.

Contrastive directions are constructed as the difference of mean activations between two sets of $P \geq 30$ matched prompt pairs $\mathcal{P}^+, \mathcal{P}^-$ differing along a single concept (e.g., for gender, $\mathcal{P}^+ \ni$ "He ran home" is paired with $\mathcal{P}^- \ni$ "She ran home") (Panickssery et al., 2023; Turner et al., 2023). The prompts are LLM generated and human verified, and are available in our code release. Letting $\mathbf{a}(s) \in \mathbb{R}^d$ denote the activation at the perturbation layer ℓ for prompt s,

$$\mathbf{d}_{\text{contrast}} = \frac{1}{P}\sum_{i=1}^{P}\big(\mathbf{a}(s_i^+) - \mathbf{a}(s_i^-)\big), \tag{4}$$

$$\hat{\mathbf{d}}_{\text{contrast}} = \mathbf{d}_{\text{contrast}}/\|\mathbf{d}_{\text{contrast}}\|_2. \tag{5}$$

Random-difference directions are constructed exactly as in Equation 4, but applied to $P = 30$ randomly paired activations (s_i, s_i') drawn from the same 10000-input FineWeb-Edu sample used

for PCA, giving a control that is matched to contrastive directions in everything except semantic content. Unlike contrastive directions, the paired prompts share no concept, so we do not expect them to be features. Unlike isotropic random directions, random-difference directions are differences of activations, and hence inherit their covariance. This makes it our most stringent non-feature baseline: it differs from a contrastive direction only in the absence of concept-coherent pairing. We construct 40 such directions.

MELBO directions (Mack & Turner, 2024b) are constructed by optimising a unit-norm direction $\hat{\mathbf{d}}$ at the perturbation layer to maximise the L^2 distance between perturbed and unperturbed activations at a downstream layer. Unlike contrastive directions, this requires no labelled prompt pairs as the directions are found unsupervised, and have been shown to produce interpretable steering when applied at sufficient magnitude.

SAE latents are decoder columns of a sparse autoencoder trained on the model's residual stream activations (Bricken et al., 2023; Cunningham et al., 2023). Each column is a candidate feature direction under the SAE's decomposition. We select the top-33 most active SAE latents in the same sample of 10000 FineWeb-Edu inputs used for PCA directions. We do this for Gemma-2-9B only and use Gemma Scope (Lieberum et al., 2024) (the width-16k residual-stream SAE at layer 2).

We also perform perturbations toward combinations of two directions $\mathbf{d}_1, \mathbf{d}_2$. We first restrict them to the tangent space at $\hat{\mathbf{a}} = \mathbf{a}/\|\mathbf{a}\|$ via the projector $I - \hat{\mathbf{a}}\hat{\mathbf{a}}^\top$, and orthonormalize (Gram–Schmidt) to obtain unit vectors $\hat{\mathbf{d}}_1^\perp, \hat{\mathbf{d}}_2^\perp$ satisfying $\hat{\mathbf{a}} \cdot \hat{\mathbf{d}}_i^\perp = 0$ and $\hat{\mathbf{d}}_1^\perp \cdot \hat{\mathbf{d}}_2^\perp = 0$. We then form the unit tangent direction

$$\hat{\mathbf{d}}(\varphi) \;=\; \cos\varphi\, \hat{\mathbf{d}}_1^\perp + \sin\varphi\, \hat{\mathbf{d}}_2^\perp, \qquad \varphi \in [0, \pi/2], \tag{6}$$

which interpolates between $\hat{\mathbf{d}}_1^\perp$ at $\varphi = 0$ and $\hat{\mathbf{d}}_2^\perp$ at $\varphi = \pi/2$. Perturbing $\mathbf{a}$ by angle α toward $\hat{\mathbf{d}}(\varphi)$ then proceeds exactly as in the single-direction case (Equation 1),

$$\mathbf{a}(\alpha, \varphi) \;=\; \cos\alpha\, \mathbf{a} + \sin\alpha\, \|\mathbf{a}\|\, \hat{\mathbf{d}}(\varphi). \tag{7}$$

For details of how we implement this sweep in practice, see Appendix C.

We observe behavior (see Section 4) that appears consistent with the downstream response depending on the perturbation's projections onto privileged directions, raised to a common power p. To formalize this, we model the downstream response R to a perturbation vector $\mathbf{v}$ as

$$R(\mathbf{v}) = F\left(\sum_i w_i \left|\hat{\mathbf{f}}_i^\top \mathbf{v}\right|^p\right), \tag{8}$$

where $\{\hat{\mathbf{f}}_i\}$ are unit directions, the weights $w_i \geq 0$ are free parameters capturing each direction's sensitivity, and F is a monotone scalar function. This is a modelling assumption rather than a prediction of FSEC.

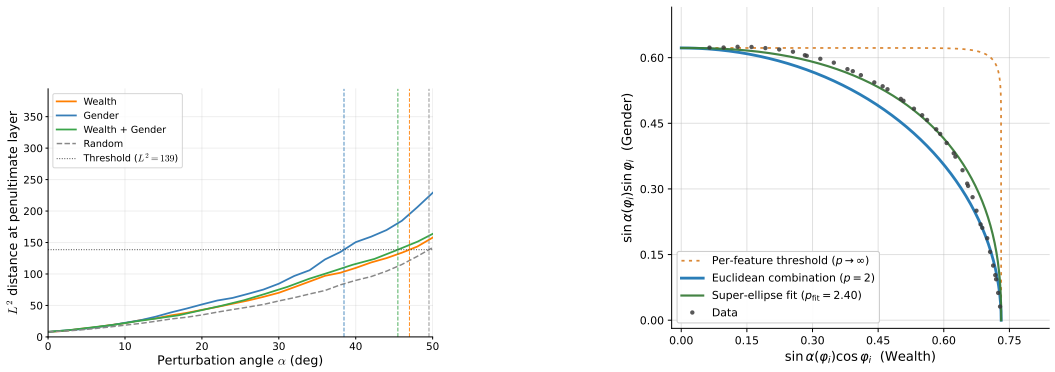
When we perturb along two directions simultaneously and vary their relative weighting via the mixing angle φ, the plateau-breaking angle becomes a function of φ. Taking $\hat{\mathbf{f}}_1 = \hat{\mathbf{d}}_1^\perp$ and $\hat{\mathbf{f}}_2 = \hat{\mathbf{d}}_2^\perp$, we derive in Appendix A that under this response model the plateau-breaking angles $\alpha(\varphi)$ satisfy

$$\left(\frac{\sin\alpha(\varphi)\cos\varphi}{\sin\alpha_1}\right)^p + \left(\frac{\sin\alpha(\varphi)\sin\varphi}{\sin\alpha_2}\right)^p = 1, \tag{9}$$

where the weights w_1, w_2 have been calibrated from the single-axis sweeps: α_1, α_2 are the plateau-breaking angles measured along $\hat{\mathbf{d}}_1^\perp$ and $\hat{\mathbf{d}}_2^\perp$ individually. That is, the plateau-breaking angles define a superellipse of exponent p in the normalised coordinates $(\sin\alpha(\varphi)\cos\varphi/\sin\alpha_1,\ \sin\alpha(\varphi)\sin\varphi/\sin\alpha_2)$ (see Figure 1b). We estimate p by least-squares fitting of Equation 9 over the measured $(\sin\alpha(\varphi)\cos\varphi,\ \sin\alpha(\varphi)\sin\varphi)$ points. We quantify the quality of every such fit by the *radial fit residual* ρ, the mean fractional radial distance of the measured points from the fitted curve ($\rho = 0$ is a perfect fit); Appendix G gives the explicit formula and the distribution of ρ across all conditions.

4 Results

We begin by illustrating our measurement procedure on a concrete example. Figure 1a shows the



(a) (b)

Figure 1: **Measuring plateau-breaking angles (a) and the resulting iso-plateau boundary (b).** **(a)** Downstream response as a function of perturbation angle for two contrastive directions (Wealth, Gender) and an equal combination of both, at Gemma-2-9B layer 2, illustrating how plateau-breaking angles are extracted. The plateau-breaking angle is the angle at which the downstream L^2 distance first exceeds the threshold (here $T \approx 139$). The grey dashed "Random" curve shows the median response across perturbations along 10 isotropic random unit directions. All curves are medians over 30 FineWeb-Edu anchor prompts; we omit uncertainty bands because the absolute L^2 scale varies substantially from anchor to anchor in a way that is shared across all curves and largely cancels in within-anchor comparisons. The within-anchor gap between each feature direction and the random baseline is nonetheless robust: at every angle, the feature curve exceeds the random baseline for most anchors ($\geq 63\%$), and the median feature-minus-random difference is significantly positive (its 95% bootstrap confidence interval over anchors excludes zero). **(b)** Plateau-breaking angles for the Wealth $\times$ Gender pair at Gemma-2-9B layer 2 (per-pair threshold $T \approx 139$). The superellipse exponent is fit in the normalised coordinates $(\sin\alpha(\varphi)\cos\varphi/\sin\alpha_1,\ \sin\alpha(\varphi)\sin\varphi/\sin\alpha_2)$ of Section 3. The boundary is well fit by a superellipse of exponent $p_{\text{fit}} = 2.40$ (radial fit residual 1.2%); $p > 2$ indicates these directions are privileged.

downstream L^2 response in Gemma-2-9B when perturbing along the Wealth contrastive direction, the Gender contrastive direction, and an equal combination of the two. The vertical dashed lines mark the plateau-breaking angle for each of the directions. Repeating this measurement across a range of mixing angles φ yields a plateau-breaking angle for each, which we plot in Figure 1b. The resulting boundary is well fit by a superellipse with exponent $p_{\text{fit}} = 2.40$. See Appendix E for supporting evidence that combining contrastive directions in this way is meaningful.

4.1 Superellipse exponents across direction types

We now systematically characterise these patterns across directions, direction types, and models. The superellipse exponent p records whether a mixture of two directions breaks the plateau sooner, later, or at the same angle as the pure directions. If the model responds only to a perturbation's overall size (its L^2 norm), splitting it between two directions changes nothing, and the plateau-breaking angles trace an ellipse ($p = 2$). If the model is instead specially sensitive to the individual directions, an even mixture—aligning only partially with each—must be pushed further before the plateau breaks: the boundary bulges outward, approaching a square ($p \to \infty$) in the limit where each direction carries a fully independent threshold. The exponent $p > 2$ measures how far toward this "independent thresholds" regime the model sits. Each axis is normalised by its own plateau-breaking angle, so only the privileging of the pure directions over their mixtures—not their individual sensitivities—affects p. We verify this directly: stratifying pairs by the mean of their two single-axis plateau-breaking angles, contrastive pairs have higher p than PCA pairs at every level of sensitivity, with PCA pairs sitting at $p \approx 1.97$ throughout (Appendix H). This selectivity is what error correction requires: staying responsive to individual features while suppressing generic mixtures of them. We give the formal statement—that $p = 2$ cannot privilege feature directions under superposition whereas $p > 2$ can—in Section 6.1.

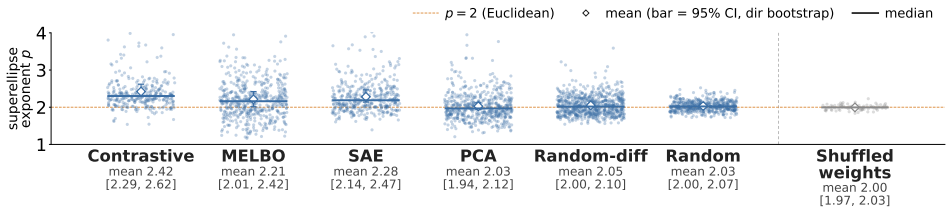


Figure 2: **Superellipse exponents by direction type.** Each dot is one fitted superellipse exponent p for a pair of directions. The black horizontal lines are the per-column medians. The dashed orange line corresponds to the $p = 2$ isotropic reference. The white markers are the per-column means, with error bars corresponding to the 95% confidence interval on the mean, computed by direction bootstrapping. Candidate feature directions (contrastive, MELBO, top-activating SAE latents) sit consistently above $p = 2$. The PCA, random, and random-difference baselines cluster at $p \approx 2$. The grey column right of the dashed divider is the shuffled-weights control (described in Section 4.1), for which we get $p \approx 2$. The superellipse fits are good: no per-condition median radial fit residual exceeds 2%.

We repeat the analysis across six direction types. For contrastive directions, we form 318 pairs from 33 directions (14 binary semantic concepts, 10 natural languages, 9 programming languages), keeping only pairs with intra-pair cosine similarity below 0.1. The filter is needed because our two-direction analysis orthogonalises the pair (Section 3): if the two directions have significant overlap, orthogonalising one against the other distorts it substantially, so the perturbation no longer probes the labelled feature. The conservative 0.1 threshold keeps this distortion small. The full list and pairwise overlap are in Appendix D. For MELBO, we use 507 pairs obtained via the procedure described in Section 3, after the same cosine-similarity filter. For SAE latents, we use top-activating decoder columns from Gemma Scope's residual stream layer 2 width-16k SAE, giving 395 pairs after the same filter. PCA and random directions serve as non-feature baselines, as do random-difference directions, for which we form 780 pairs from 40 directions and keep the 710 with intra-pair cosine similarity below 0.1. Finally, as a methodological control, we rerun our entire procedure—contrastive direction extraction, anchor computation, sweeps, and fits—on copies of Gemma-2-9B with randomly permuted entries of every transformer-block weight matrix (5 seeds).

We observe that contrastive directions (mean 2.42, 95% CI [2.29, 2.62], median 2.30), MELBO directions (mean 2.21, 95% CI [2.01, 2.42], median 2.16), and top-activating SAE latents (mean 2.28, 95% CI [2.14, 2.47], median 2.19) sit consistently above $p = 2$. This indicates that the model is more sensitive to perturbations along these directions than along their mixtures, consistent with them being aligned with the features and with FSEC. We emphasize that our result is not merely that these directions are sensitive—as prior work on activation plateaus and MELBO has shown—but that they are more sensitive than mixtures of them, which is precisely what $p > 2$ measures. The same is not true for any of our baselines: top PCA directions (mean 2.03, 95% CI [1.94, 2.12], median 1.97), random-difference directions (mean 2.05, 95% CI [2.00, 2.10], median 2.02), and random directions (mean 2.03, 95% CI [2.00, 2.07], median 2.02) all cluster at $p \approx 2$.

We compare the exponents of the contrastive directions against those of the random-difference baseline with a one-sided Mann–Whitney U test: the difference is significant at the 10^{-47} level, and a randomly drawn contrastive pair exceeds a randomly drawn baseline pair 76% of the time. This test assumes independent samples, which our pair fits are not, since two pairs can share a direction. We therefore also estimate the difference in medians with a direction-level cluster bootstrap—resampling directions with replacement and rebuilding all pairs from the resampled set—which puts the contrastive median 0.29 above the baseline median (95% CI $[0.18, 0.41]$). This is expected for the directions not aligned with the features under FSEC, as neither direction in the pair should be privileged over its mixtures. The shuffled-weights measurement rerun yields $p \approx 2$ (mean 2.00, 95% CI [1.97, 2.03]), confirming that $p > 2$ arises only for trained weights.

4.2 Ablations

We now show that our results hold across a variety of settings. To do so, we vary our setup along the following axes:

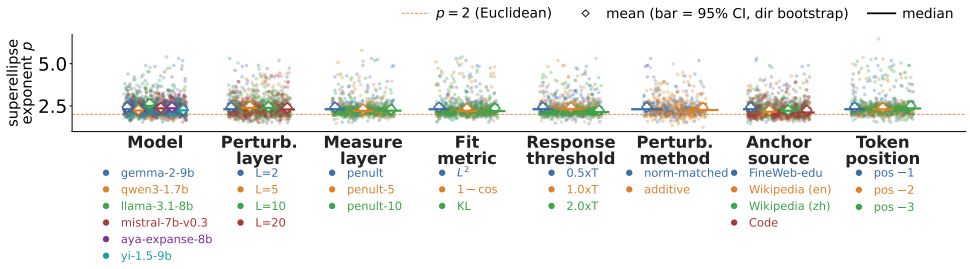


Figure 3: **Robustness of $p > 2$ for contrastive directions across setup choices.** Each dot is one fitted superellipse exponent p for a pair of contrastive directions; the colored marker shows the per-setting mean with 95% CI, and the horizontal dash shows the median. The dashed orange line is the $p = 2$ isotropic reference. Columns vary the model, perturbation layer, measurement layer, response metric, response threshold, perturbation method, activation source, and perturbation token position. Fitted exponents remain above $p = 2$ across all variations. The superellipse fits are good: no per-condition median radial fit residual exceeds 2%.

- **Perturbation method:** additive rather than norm-matched.
- **Perturbation layer:** $\ell \in \{5, 10, 20\}$ in addition to the default $\ell = 2$.
- **Measurement layer:** layers 30 and 35 in addition to the default second-to-last (layer 40 for Gemma-2-9B).
- **Response metric:** cosine distance at second-to-last residual stream layer and KL-divergence in logit space in addition to L^2 distance.
- **Plateau-breaking threshold:** 50% and 200% of the per-pair threshold T.
- **Activation source:** wikipedia-en, wikipedia-zh, and the-stack-smol (default FineWeb-Edu).
- **Perturbation token position:** second- and third-to-last token (default: last token).
- **Model:** Qwen3-1.7B, Llama-3.1-8B, Mistral-7B-v0.3, Aya-Expanse-8B, and Yi-1.5-9B.

All our results hold across every ablation with the full results shown in Figure 3. For contrastive feature directions, the means are all in the [2.24, 2.63] interval, and the medians in [2.11, 2.50], all consistent with contrastive feature directions being privileged. We also check the estimator itself by checking its ability to recover a known exponent from synthetic contours in Appendix J.

5 Toy model validation

To validate our method, we test whether it detects FSEC in a toy model with known ground-truth features. This lets us (i) verify that our methodology yields high p and good superellipse fits on known ground-truth features, and (ii) measure how p degrades as perturbation directions are rotated away from the *true* features, which the toy model exposes.

We use the two-layer denoising network of Vaintrob (2026) as our toy model. It denoises two-hot d-dimensional inputs corrupted with Gaussian noise using $H < d$ neurons, so the d features are denoised (i.e., computed) in superposition over fewer than d dimensions. The input is expressed in the feature basis, giving us direct access to the ground-truth feature directions.

We apply our two-direction perturbation analysis as before, perturbing additively along pairs of ground-truth inactive feature directions. To measure the effect of misalignment, we rotate the perturbation directions away from the true feature directions toward random directions by varying amounts. We sweep three feature-to-neuron ratios ($4\times$, $8\times$, $16\times$) with $H = 1024$ neurons. Full details of the toy model setup are in Appendix K.

Figure 4 shows the fitted exponent p as a function of the perturbation directions' alignment with the true feature axes. For ground-truth feature directions, we obtain median p in the range 3–3.4, well above 2, increasing with the feature-to-neuron ratio (highest for $16\times$). As directions are rotated away from the true features, p degrades monotonically toward 2, which is recovered for random directions. We apply a similar rotation procedure to an LLM—where we lack ground-truth features

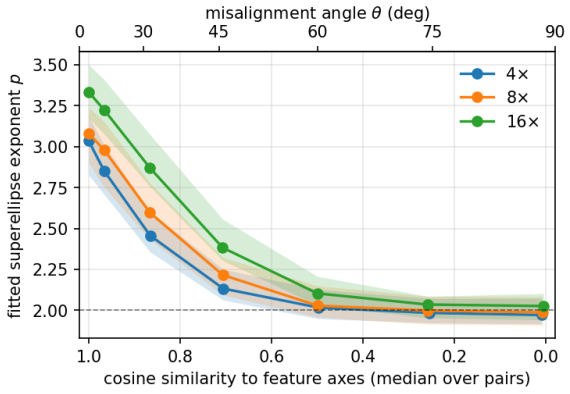


Figure 4: **Superellipse exponents versus alignment with feature directions.** Each dot is the median of 80 fitted superellipse exponents; the shaded region shows the interquartile range. The x-axis shows the median realized overlap $|\langle \mathbf{u}, \mathbf{e}_j \rangle|$ between the perturbation direction $\mathbf{u}$ and the feature axis $\mathbf{e}_j$ over the 80 pairs: the leftmost points correspond to ground-truth feature directions, the rightmost to isotropic random unit vectors. The three lines correspond to feature-to-neuron ratios of $4\times$, $8\times$, and $16\times$, all with $H = 1024$ neurons. Ground-truth feature directions yield median p in the range 3–3.4; p degrades monotonically toward 2 as directions are rotated away from the true features.

and instead start from the candidate contrastive directions—and obtain equivalent results: p decreases monotonically from ≈ 2.4 at the contrastive directions to ≈ 2.0 at full rotation (see Appendix F for details). This is consistent with our prediction that FSEC produces $p > 2$ and that imperfect feature alignment biases p downward. Due to the significant differences between this toy model and LLMs (two layers, binary features, $\tanh^3$ activation function), we do not place much weight on the specific values obtained; we view this primarily as validation of the methodology and the directional privileging prediction.

6 Discussion

6.1 Why $p > 2$ is necessary for error correction

When neural networks use superposition, features are embedded non-orthogonally, so each active feature produces a small spurious "interference" activation along the others. To compute in superposition despite this interference noise, a network must implement error correction (Hänni et al., 2024), and such a mechanism must treat feature directions differently from non-feature directions. To test whether this is the case, we measure the exponent p in our response model (Equation 8).

The exponent p controls how selectively the network can respond to individual feature directions. When $p = 2$, the response reduces to a quadratic form,

$$R(\hat{\mathbf{d}}) = F\left(\hat{\mathbf{d}}^\top \mathbf{A}\, \hat{\mathbf{d}}\right), \tag{10}$$

where $\mathbf{A} = \sum_i w_i \hat{\mathbf{f}}_i \hat{\mathbf{f}}_i^\top$ is a positive semidefinite matrix with at most d_{model} nonzero eigenvalues. Under superposition, we expect far more features than dimensions, yet the local maxima of a quadratic form on the unit sphere are exactly the eigenvectors of $\mathbf{A}$, so at most d_{model} directions can be privileged above all others. Even without superposition, $p = 2$ does not allow all feature directions to be more responsive than all non-feature directions: if the eigenvalues of $\mathbf{A}$ are not all equal, convexity implies that some non-feature directions (linear combinations of high-eigenvalue features) will be more responsive than the lowest-eigenvalue features; if all eigenvalues are equal, the response is isotropic and no direction is privileged at all. In either case, $p = 2$ is insufficient for error correction that privileges feature directions. For simplicity, we focused here on the case where the quadratic form has no cross terms with respect to the feature directions. We generalize this in Appendix I and rule out quadratics with cross terms as an explanation of our results.

At the other extreme, $p \to \infty$ makes the response sensitive only to the single largest feature projection, perfectly suppressing interference from all other directions. This extreme is likely not how LLMs function either, so we expect the true exponent to lie between the two: $2 < p < \infty$.

Empirically, we find $p \approx 2.3$ for contrastive directions and $p \approx 2.2$ for MELBO and SAE directions, consistently above $p = 2$ and consistent with FSEC. Since our candidate feature directions are imperfect approximations of the model's true features, the measured p is likely a lower bound on the exponent that would be obtained with ground-truth feature directions: imperfect alignment would dilute the directional sensitivity, biasing p toward 2. Our toy model analysis (Section 5) supports this, showing that p degrades monotonically toward 2 as directions are rotated away from the true features. In fact, moderate misalignment in the form of perturbation directions with ~ 0.8 cosine similarity to the features reduces the median fitted exponent from 3–3.4 to 2.3–2.6, so our measured $p \approx 2.3$ in LLMs is consistent with significantly higher p for true feature directions.

6.2 Alternative interpretations

Our finding that $p > 2$ for candidate feature directions is consistent with FSEC. However, there exist alternative interpretations of this result and of activation plateaus more broadly.

Activation plateaus might not be signatures of error correction at all. They could instead reflect learned robustness, perhaps related to broad basins of attraction in the activation space of LLMs. This could hold whether or not LLMs implement error correction: error correction could be present but plateaus the wrong empirical signature, or absent, with plateaus reflecting a different phenomenon. However, our main result does not depend on interpreting plateaus as error correction per se: the finding that $p > 2$ for candidate feature directions but $p \approx 2$ for baselines remains, and still indicates that the network treats these directions preferentially.

It is also possible that LLMs implement error correction in a way that is not directly direction-aware, without features being linear directions consistent across the activation space. For instance, the in-distribution activations of LLMs may lie on a lower-dimensional manifold (Goodfellow et al., 2016; Cheng et al., 2023), and activation plateaus could reflect a mechanism by which off-manifold perturbations are mapped back on-manifold. This would be broadly consistent with our observations but would admit a different interpretation than the one we have put forward.

7 Conclusion

We have proposed perturbation analysis of residual-stream activations as an empirical test of whether feature directions are privileged and LLMs implement error correction. We find that contrastive, MELBO, and top-activating SAE directions are privileged, consistent with superposition and error correction. Our results hold across six model families and a wide range of settings.

Our results consistently select $p \approx 2.3$ for our best candidate feature directions. As discussed in Section 6, this is likely a lower bound: imperfect alignment between our candidates and the model's true features dilutes directional sensitivity, biasing p toward 2. This provides a quantitative target for toy models of error correction: a good toy model should reproduce $p > 2$ under comparable conditions. The toy model of Section 5 passes this test, yielding p of 3–3.4 for ground-truth features.

Future work. This suggests using the exponent directly as an unsupervised feature-finding objective. In preliminary experiments, $p > 2$ appears necessary but not sufficient for a direction to be a feature: maximizing the exponent alone surfaces high-p directions that do not otherwise behave like features. The toy model, with known ground-truth features, offers a controlled test of this idea. Extending it to standard activation functions and deeper architectures would test how closely toy models can match LLM exponents.

AI use statement

We used generative AI tools to generate the contrastive prompt pairs of Section 3 and to generate the majority of the code used in our experiments. Both the prompt pairs and the code were verified for correctness by one of the authors. Generative AI tools were used for feedback and editing help on the

manuscript. We did not otherwise use generative AI tools in this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

Code required to reproduce all experiments is available at https://anonymous.4open.science/r/fsec-anonymous-5C9D/. Section 3 describes the perturbation procedure we implement and how we construct all the types of directions we investigate. Appendix C describes how we implement two-direction sweeps , Appendix B the choice of threshold, and Appendix D includes all contrastive directions. The superellipse equation reported in the main text is derived in Appendix A. All LLMs we study in this work are publicly available. The SAE is in the public Gemma Scope release (Lieberum et al., 2024), and the prompts and activation samples used for PCA and random-difference baselines are drawn from FineWeb-Edu (Penedo et al., 2024). Details for the toy model experiments are given in Appendix K.

References

Micah Adler and Nir Shavit. On the complexity of neural computation in superposition. *arXiv preprint arXiv:2409.15318*, 2024.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Emily Cheng, Corentin Kervadec, and Marco Baroni. Bridging information-theoretic and geometric compression in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12397–12420, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*, 2024.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press Cambridge, MA, USA, 2016.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Kaarel Hänni, Jake Mendel, Dmitry Vaintrob, and Lawrence Chan. Mathematical models of computation in superposition. *arXiv preprint arXiv:2408.05451*, 2024.

Stefan Heimersheim and Jake Mendel. [interim research report] Activation plateaus & sensitive directions in GPT2. AI Alignment Forum, July 2024. URL https://www.alignmentforum.org/posts/LajDyGyiyX8DNNsuF/interim-research-report-activation-plateaus-and-sensitive-1. Work produced at Apollo Research.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Jett Janiak, Jacek Karwowski, Chatrik Singh Mangat, Giorgi Giglemiani, Nora Petrova, and Stefan Heimersheim. Characterizing stable regions in the residual stream of llms. *arXiv preprint arXiv:2409.17113*, 2024.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024.

Andrew Mack and Alex Turner. Deep causal transcoding: A framework for mechanistically eliciting latent behaviors in language models, 12 2024a. URL https://www.lesswrong.com/posts/fSRg5qs9TPbNy3sm5/deep-causal-transcoding-a-framework-for-mechanistically. LessWrong.

Andrew Mack and Alex Turner. Mechanistically eliciting latent behaviors in language models, 4 2024b. URL https://www.lesswrong.com/posts/ioPnHKFyy4Cw2Gr2x/mechanistically-eliciting-latent-behaviors-in-language-1. LessWrong.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Chris Olah, Nicholas L. Turner, and Tom Conerly. A toy model of interference weights. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/interference-weights/index.html.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Matthew Shinkle and Stefan Heimersheim. Activation plateaus: Where and how they emerge. LessWrong, October 2025. URL https://www.lesswrong.com/posts/WMfSbt7AAcJdHzysB/activation-plateaus-where-and-how-they-emerge. Accessed: 2026-04-27.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *arXiv preprint arXiv:2605.29358*, 2026.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Dmitry Vaintrob. A tale of three theories: Sparsity, frustration, and statistical field theory. LessWrong, January 2026. URL https://www.lesswrong.com/posts/siu22scEfuKxpSgfK/a-tale-of-three-theories-sparsity-frustration-and. Accessed: 2026-05-06.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

A Derivation of the superellipse equation

We derive Equation 9 from the response model of Equation 8. Taking $\hat{\mathbf{f}}_1 = \hat{\mathbf{d}}_1^{\perp}$ and $\hat{\mathbf{f}}_2 = \hat{\mathbf{d}}_2^{\perp}$, the inner sum of Equation 8 collapses to two terms provided the remaining $\hat{\mathbf{f}}_j$ contribute negligibly along $\hat{\mathbf{d}}(\varphi)$. This is expected under superposition, where features are packed nearly orthogonally (Elhage et al., 2022), so the projections of $\hat{\mathbf{d}}(\varphi)$ onto other directions $\hat{\mathbf{f}}_j$ are small. For the norm-matched perturbation of Equation 7, the displacement is $\mathbf{v} = \mathbf{a}(\alpha, \varphi) - \mathbf{a} = \sin\alpha \, \|\mathbf{a}\| \, \hat{\mathbf{d}}(\varphi) - (1 - \cos\alpha)\, \mathbf{a}$. The term $-(1 - \cos\alpha)\, \mathbf{a}$ lies along $\mathbf{a}$ and is therefore orthogonal to each $\hat{\mathbf{f}}_i = \hat{\mathbf{d}}_i^{\perp}$, so it drops out of every projection, $\hat{\mathbf{f}}_i^{\top} \mathbf{v} = \sin\alpha \, \|\mathbf{a}\| \, (\hat{\mathbf{f}}_i^{\top} \hat{\mathbf{d}}(\varphi))$; only the tangential component contributes, so

$$R = F\big(\sin^p \alpha \, \|\mathbf{a}\|^p \big[w_1 \cos^p \varphi + w_2 \sin^p \varphi\big]\big) \, . \tag{11}$$

The plateau breaks when F's argument crosses a fixed level ξ^* (the value of the argument at which the response reaches the threshold T), giving the implicit equation for the plateau-breaking angle $\alpha(\varphi)$:

$$\sin^p \alpha(\varphi) \, \|\mathbf{a}\|^p \big[w_1 \cos^p \varphi + w_2 \sin^p \varphi\big] = \xi^* . \tag{12}$$

The single-axis sweeps (at $\varphi = 0$ and $\varphi = \pi/2$) calibrate the weights at this same threshold: $\varphi = 0$ gives $w_1 \, \|\mathbf{a}\|^p = \xi^* / \sin^p \alpha_1$, and $\varphi = \pi/2$ gives $w_2 \, \|\mathbf{a}\|^p = \xi^* / \sin^p \alpha_2$, where α_1, α_2 are the plateau-breaking angles measured along $\hat{\mathbf{d}}_1^{\perp}$ and $\hat{\mathbf{d}}_2^{\perp}$ individually. Substituting and dividing through by ξ^* eliminates the unknown threshold and the activation norm, yielding Equation 9.

We note that the cancellation of the term along $\mathbf{a}$ is exact only for $\hat{\mathbf{f}}_1$ and $\hat{\mathbf{f}}_2$, which are orthogonal to $\mathbf{a}$ by construction. Each remaining feature $\hat{\mathbf{f}}_j$ is not, and picks up a contribution $-(1 - \cos\alpha)\, \hat{\mathbf{f}}_j^{\top} \mathbf{a}$, which the two-term approximation above neglects. While this could in principle bias our fits, the additive-perturbation ablation of Section 4.2 controls for this, as an additive perturbation has no term along $\mathbf{a}$ at all. Over the same 318 contrastive pairs, both methods result in similar median fitted exponents well above 2, namely 2.30 for norm-matched and 2.26 for additive.

B Threshold choice

We initially tried a single global threshold derived from random directions—a fraction f of the median, over a set of isotropic random unit directions, of the single-axis plateau height each one reaches—but abandoned it. Direction sensitivities span a wide range (up to a factor of $\sim$70 between the least and most responsive directions for the KL-divergence response metric), so a global threshold is either never reached by the least responsive directions or crossed only at very small angles by the most responsive ones, leaving most superellipse fits ill-defined. Setting T per pair from each pair's own single-axis maxima guarantees both axes cross and resolves this. We record this here for ease of reproducibility.

C Implementation of the two-direction perturbation sweeps

In this appendix, we describe our implementation of the two-direction norm-matched perturbations. For a pair $(\hat{\mathbf{d}}_1^{\perp}, \hat{\mathbf{d}}_2^{\perp})$ we evaluate the downstream response on a 31×31 grid over the per-direction rotation angles $(\alpha_1, \alpha_2) \in [0°, 60°]^2$ (2° spacing), where the perturbation rotates $\mathbf{a}$ by angle $\sqrt{\alpha_1^2 + \alpha_2^2}$

toward the tangent direction with components (α_1, α_2) along $(\hat{\mathbf{d}}_1^{\perp}, \hat{\mathbf{d}}_2^{\perp})$. This is the parameterization of Equation 7 with $\alpha = \sqrt{\alpha_1^2 + \alpha_2^2}$ and $\varphi = \arctan(\alpha_2/\alpha_1)$, so the grid's first row and column are the two single-axis sweeps, and the maxima entering the threshold T are taken over $\alpha \in [0°, 60°]$. Each grid value is the median over the 30 anchors (Equation 2). The plateau boundary is the threshold-T level set of this median grid, extracted by linear interpolation between grid points; rather than being fixed in advance, the mixing angles φ entering each fit are those realized along the extracted contour. The superellipse exponent is then fit to the contour points in the normalised coordinates $(\sin\alpha(\varphi)\cos\varphi/\sin\alpha_1,\ \sin\alpha(\varphi)\sin\varphi/\sin\alpha_2)$ of Section 3.

D CONTRASTIVE DIRECTION INVENTORY

We extract 33 contrastive directions across three families. Each direction is the difference of mean last-token activations at the perturbation layer between a positive prompt set and a matched negative set, normalised to a unit vector.

The 14 **binary semantic** directions are Age, Certainty, Era, Gender, Health, Honesty, Literary, Number, Person, Refusal, Sentiment, Status, Tense, and Wealth. Each uses 30 matched prompt pairs except Literary, which contrasts the slang and literary registers across 50 prompt templates. Prompts are LLM-generated and human-verified.

The 10 **natural-language** directions cover Arabic, Chinese, Dutch, French, German, Italian, Japanese, Portuguese, Russian, and Spanish. Each pair is an English prompt and its translation in the target language; we use 36 matched prompts per direction, drawn from a single multilingual prompt pool so that all 10 directions share the same English source side.

The 9 **programming-language** directions cover C++, Go, Haskell, Java, JavaScript, Lisp, Python, Rust, and TypeScript. Each pair is an English prompt and an equivalent code snippet in the target language; 30 matched prompts per direction.

Forming all $\binom{33}{2} = 528$ unordered pairs, we keep the 318 with raw overlap $|\langle \mathbf{d}_i, \mathbf{d}_j \rangle| < 0.1$ for our two-direction analysis (Figure 5). The filter is needed because our two-direction analysis orthogonalises each pair via Gram–Schmidt (Section 3): if two directions overlap significantly, removing the shared component leaves only a small remainder, which renormalisation to unit length then amplifies, so the resulting direction reflects this remainder rather than the labelled feature. The 0.1 threshold is conservative—at this overlap the orthogonalised direction remains more than 99% aligned with the original, and the distortion grows with overlap.

E COMPOSITIONAL STEERING

To verify that combining contrastive directions as in Section 4 is meaningful, we show in Figure 6 that steering along these combinations produces interpretable compositional behaviour changes. Concretely, steering along the Wealth direction (toward its low-wealth pole) produces poverty-themed completions from a neutral input, steering along the Gender direction produces feminine-coded completions, and steering along a combination of the two produces completions that are both poverty-themed and feminine-coded. This supports the reasonableness of the combination-perturbation setup, though we note that interpretable steering alone does not establish that a direction corresponds to a model feature.

F LLM MISALIGNMENT SWEEP

Complementing the toy-model rotation sweep of Section 5, we test in an LLM whether p decays monotonically to $p = 2$ as we rotate away from candidate feature directions. We take a pair of contrastive feature directions and we rotate each of the directions toward a fixed random orthogonal direction by an angle θ, re-orthonormalize the pair, and refit, so that $\theta = 0$ recovers the original contrastive pair and $\theta = \pi/2$ yields a random direction. We sweep $\theta \in [0, \pi/2]$ over a subsample of 40 overlap-filtered pairs (subsampled only to keep the computational cost manageable, as the full sweep is more expensive than the single-pair fits of Section 4.1). For each pair we average over four independent rotation realizations: within each, every direction is tilted toward its own fixed

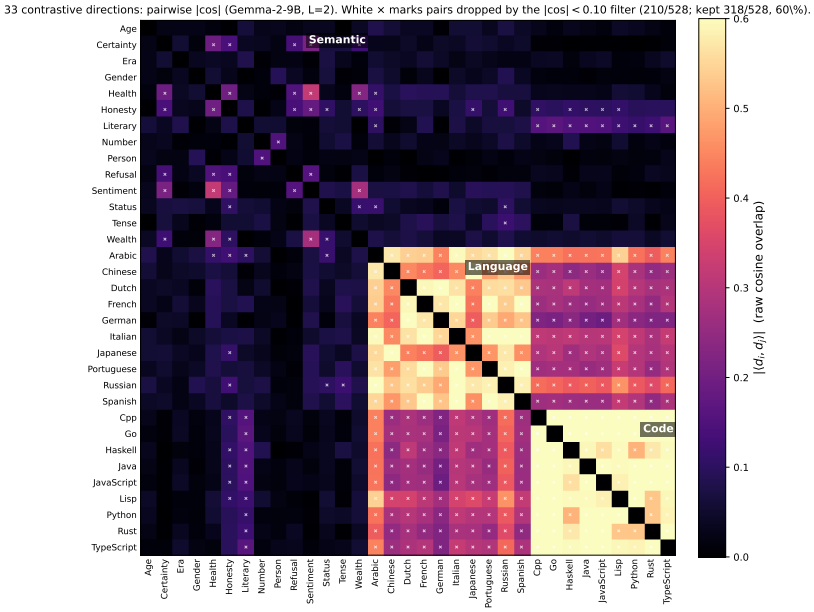


Figure 5: **Pairwise cosine overlap of the 33 contrastive directions.** Cells show $|\langle \mathbf{d}_i, \mathbf{d}_j \rangle|$ on Gemma-2-9B at layer 2. Directions are ordered and labelled by family (semantic, natural language, programming language). White $\times$ marks the 210 pairs dropped by the $|\cos| < 0.1$ filter applied throughout the paper; the remaining 318 pairs are the ones plotted as Contrastive in the beeswarms. Most filtered pairs sit inside the natural-language block.

Prompt:	The other day I met someone who
Unsteered:	was interested in the idea of moving to Austin, TX. He wanted a bit more detail about what makes it such an attractive place for professionals and students from out-of-town. What are some reasons we love living here? Here's my…
+ Poverty 30°:	was so poor that he had no money to buy food. He asked me for some bread and water, but all the while my heart bled with him as it could not give anything more than a little bit of pity…
+ Female 20°:	was wearing a pair of earrings that caught my eye. The girl explained they were her engagement ring and wedding bands all in one, so she didn't have to take off the earring set when taking them on…
+ Poverty + Female 30°:	was homeless. She told me she had nowhere to go and no money in her pocket, so she slept out on the street every night because of poverty…

Figure 6: **Compositional steering at Gemma-2-9B layer 2.** Sample completions for the prompt "The other day I met someone who" under no steering, +Poverty alone, +Female alone, and the Poverty + Female composite. Steering uses the contrastive Wealth and Gender directions (Appendix D); each row shows a single pole—the low-wealth (poverty) pole of Wealth and the feminine pole of Gender. Highlighted spans: orange for poverty and blue for feminine words.

random orthogonal target (Figure 7). The fitted exponent decreases monotonically from $p \approx 2.4$ at the contrastive directions to $p \approx 2.0$ once they are fully rotated to random, indicating that p tracks alignment with the candidate feature directions.

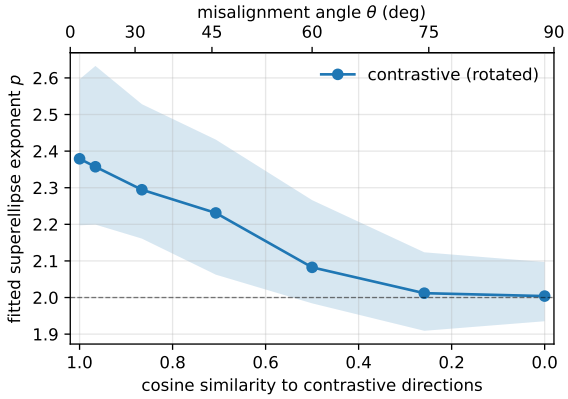


Figure 7: **Superellipse exponent versus rotation away from contrastive directions.** Each point is the median fitted p over the 40 overlap-filtered contrastive pairs and four independent rotation realizations per pair; the shaded band shows the interquartile range. The dashed line marks the $p = 2$ isotropic reference. At $\cos\theta = 1$ the perturbation directions are the contrastive directions; the median there ($p \approx 2.4$) is slightly above the full contrastive-set median (2.30, Figure 2) as it is computed over a 40-pair subsample. At $\cos\theta = 0$ they have been rotated fully onto a random orthogonal direction. p degrades monotonically from ≈ 2.4 toward 2 as the directions are rotated away from the contrastive features.

G Distribution of radial fit residuals

For every fitted pair we record the radial fit residual ρ introduced in Section 3—the same quantity that we quote in the main text—

$$\rho = \frac{1}{M}\sum_{k=1}^{M}\left|(|x_k|^p + |y_k|^p)^{1/p} - 1\right|,$$

i.e. the average fractional radial distance of the M contour points from the fitted superellipse, with $\rho = 0$ corresponding to a perfect fit. The number of contour points M varies per fit (it is the cardinality of the iso-threshold level set on the 2D L^2 grid); the per-fit median is 33 (IQR 29–37).

Each pair contributes one dot in Figure 8.

Every sub-group median radial fit residual is below 2%.

H Sensitivity does not account for the measured values of p

p is invariant to how sensitive the two directions being tested are because each axis is normalized by its own plateau-breaking angle, which removes the weights w_1, w_2 of Equation 8. However, this explanation only holds if our response model is correct. In this appendix, we check empirically whether sensitivity could explain the values of p we measure without assuming the model.

We measure sensitivity by the single-axis plateau-breaking angle α, the angle at which a perturbation along one direction alone first crosses the plateau-breaking threshold. Smaller α means a more sensitive direction.

Table 1 shows that sensitivity does not order the direction families the way p does. PCA directions are more sensitive than contrastive directions ($40.8°$ against $42.3°$) and have the lowest p of any directions. MELBO directions are the most sensitive of all and give a lower p than contrastive directions. Hence, how sensitive a family of directions is does not predict its p.

We further stratify all pairs of directions by the mean of their two single-axis plateau-breaking angles and compare the p of direction families within each stratum (Table 2). Within a stratum the sensitivity is approximately constant and only the direction family varies. Contrastive pairs exceed PCA pairs in every stratum, by 0.21 to 1.13, with PCA pairs remaining at $p \approx 1.97$ across the entire range.

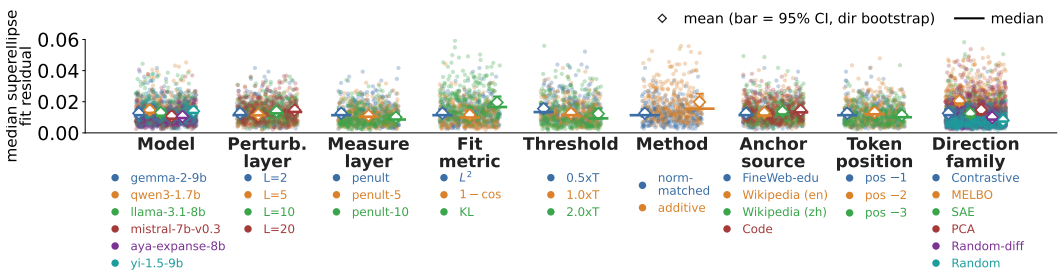


Figure 8: **Distribution of radial fit residuals across all conditions.** Each dot is one pair's radial fit residual ρ. The eight left-most columns mirror the ablation axes of Figure 3 (Model, Perturbation layer, Measurement layer, Response metric, Response threshold, Perturbation method, Token position, Activation source). The right-most column gives the radial fit residual distribution per direction family (Contrastive, MELBO, SAE, PCA, Random-diff, Random). Per-sub-group mean (white diamond) $\pm$ 95% direction-bootstrap CI and median (horizontal bar) are colour-matched to the sub-group's dots. The y-axis is clipped at 6% to keep the bulk of the distribution legible; this clips 17 of the 11,762 plotted dots ($\approx$0.14%).

Family	n	med. p	med. α	IQR of α	med. $\lvert\alpha_1 - \alpha_2\rvert$	$r_s(p, \alpha)$
Contrastive	318	2.30	42.3°	[37.7, 47.0]	9.1°	−0.60 [−0.67, −0.51]
MELBO	507	2.16	36.6°	[31.7, 43.0]	8.2°	−0.37 [−0.44, −0.29]
SAE	395	2.19	42.1°	[25.5, 45.1]	20.5°	−0.34 [−0.42, −0.25]
PCA	528	1.97	40.8°	[36.4, 43.5]	5.1°	−0.03 [−0.11, +0.06]
Random-diff	710	2.02	45.0°	[42.3, 46.9]	3.4°	−0.16 [−0.23, −0.08]
Random	528	2.02	44.0°	[42.5, 45.3]	1.9°	−0.17 [−0.25, −0.08]

Table 1: Sensitivity by direction family, Gemma-2-9B layer 2. α statistics are over individual axes, with $\lvert\alpha_1 - \alpha_2\rvert$ being the difference in sensitivity within a direction pair. r_s is Spearman's rank correlation between p and the pair's mean α, with a 95% direction-bootstrap interval.

This difference is therefore not explained by sensitivity. The same conclusion holds when pairs are stratified by the smaller or by the larger of their two angles instead of the mean.

$\bar{\alpha}$	Contrastive	MELBO	SAE	PCA	Random-diff	Random
$< 35°$	3.10 (38)	2.31 (185)	2.31 (170)	1.97 (25)	—	—
35–38°	2.46 (87)	2.17 (106)	2.34 (72)	1.97 (111)	2.25 (10)	—
38–41°	2.24 (33)	2.06 (108)	2.07 (57)	1.98 (199)	2.09 (57)	2.17 (13)
41–44°	2.22 (90)	2.06 (66)	2.04 (65)	1.96 (171)	2.04 (198)	2.03 (270)
$> 44°$	2.18 (70)	1.86 (42)	1.96 (31)	1.97 (22)	1.99 (445)	2.00 (245)

Table 2: Median p (number of pairs) by stratum of $\bar{\alpha}$, the mean of a pair's two single-axis plateau-breaking angles. Cells with fewer than 10 pairs are omitted.

Table 1 also shows that the candidate-feature families have a larger within-pair asymmetry $\lvert\alpha_1 - \alpha_2\rvert$ than the baselines, so we repeat the stratification on that quantity (Table 3). Contrastive pairs again exceed every baseline in every stratum. The gap narrows among the most symmetric pairs, to 0.11–0.20 against 0.28–0.33 over all pairs, but it does not close, so the ordering of the families is not produced by asymmetry either.

Within the three candidate-feature direction families p does vary with α ($r_s = -0.60, -0.37, -0.34$), and with $\lvert\alpha_1 - \alpha_2\rvert$ ($r_s = 0.46, 0.19, 0.16$, against 0.09 for PCA). This admits two interpretations between which we cannot adjudicate. One, that the better-aligned members of a family are both more sensitive and more privileged, so the covariation is a property of the true features. Two, that sensitivity itself raises the fitted exponent, so the covariation is a property of the measurement rather

$\lvert\alpha_1 - \alpha_2\rvert$	Contrastive	MELBO	SAE	PCA	Random-diff	Random
$< 3°$	2.13 (39)	2.04 (108)	2.07 (40)	1.93 (170)	1.99 (326)	2.02 (379)
3–6°	2.15 (47)	2.10 (86)	2.03 (43)	1.93 (137)	2.04 (223)	2.02 (134)
6–10°	2.31 (78)	2.16 (104)	2.06 (48)	1.99 (126)	2.02 (138)	1.95 (15)
10–20°	2.72 (63)	2.20 (176)	2.38 (60)	2.05 (95)	2.19 (23)	—
$> 20°$	2.42 (91)	2.61 (33)	2.24 (204)	—	—	—

Table 3: Median p (number of pairs) by stratum of the within-pair sensitivity asymmetry $|\alpha_1 - \alpha_2|$. Cells with fewer than 10 pairs are omitted.

than the features. However, the between-family comparison does not depend on which is right: PCA directions are in every stratum of Table 2 and show no variation in p across them, so whatever produces the covariation within the candidate-feature direction families does not produce $p > 2$ on its own.

I Ruling out quadratic responses with cross terms

The argument of Section 6.1 shows that a quadratic response without cross terms yields $p = 2$. For a general quadratic response $R(\mathbf{v}) = F(\mathbf{v}^\top \mathbf{A}\mathbf{v})$ the plateau-breaking condition reads $\sin^2\alpha \, \|\mathbf{a}\|^2 \left[A_{11}\cos^2\varphi + 2A_{12}\cos\varphi\sin\varphi + A_{22}\sin^2\varphi\right] = \xi^*$. The single-axis intercepts fix A_{11} and A_{22}, so in the normalized coordinates of Equation 9 the boundary becomes the tilted ellipse

$$x^2 + 2\delta\, xy + y^2 = 1, \qquad \delta = \frac{\hat{\mathbf{d}}_1^{\perp\top}\mathbf{A}\,\hat{\mathbf{d}}_2^{\perp}}{\sqrt{(\hat{\mathbf{d}}_1^{\perp\top}\mathbf{A}\,\hat{\mathbf{d}}_1^{\perp})(\hat{\mathbf{d}}_2^{\perp\top}\mathbf{A}\,\hat{\mathbf{d}}_2^{\perp})}}. \tag{13}$$

This tilted ellipse can closely mimic a superellipse with $p > 2$. Choosing p such that the boundaries of the two possible curves coincide at $x = y$, and noting that by definition they already share the axes intercepts, gives $p = 2\ln 2/\ln(2 + 2\delta)$. This implies that a cross term of $\delta \approx -0.09$ corresponds to $p \approx 2.3$, and the corresponding curves differ by less than 1% radially, below our per-pair radial fit residuals. Therefore, a fitted exponent above 2 on a single pair does not exclude a quadratic response.

However, we can bring the two hypotheses apart by comparing the fits across many pairs. For every pair of directions on Gemma-2-9B, layer 2, we extract the iso-response contour exactly as in the main analysis and fit both one-parameter families, the superellipse in p and the tilted ellipse in δ. We compare the fits by the radial fit residual of Section 3, computed for each fitted curve, and test the paired per-pair differences with a two-sided sign test.

Family	n	med. p	med. δ	resid. p	resid. δ	ellipse wins	sign test
Contrastive	318	2.30	-0.080	1.14%	1.28%	103/318	3×10^{-10}
MELBO	507	2.16	-0.050	1.88%	2.09%	169/507	5×10^{-14}
SAE	395	2.19	-0.057	1.04%	1.16%	158/395	8×10^{-5}
PCA	528	1.97	$+0.006$	1.20%	1.20%	275/528	0.36
Random-diff	710	2.02	-0.008	0.85%	0.85%	404/710	3×10^{-4}
Random	528	2.02	-0.007	0.73%	0.72%	271/528	0.57

Table 4: Superellipse versus tilted-ellipse fits on identical contour points. The two "resid." columns give the radial fit residual of each fitted family. "Ellipse wins" counts direction pairs where the tilted ellipse achieves the lower radial fit residual.

The superellipse fits better than the tilted ellipse for all three candidate-direction families. For the baselines the two fits are statistically indistinguishable (PCA, random) or differ by a negligible margin in the opposite direction: on random-difference pairs the tilted ellipse wins slightly more often (57%), but the median residuals of the two fits coincide to two decimals and the fitted cross term is essentially zero (median $\delta = -0.008$). This is expected, as the two curve families coincide at $p = 2$, $\delta = 0$, so for baseline pairs the comparison resolves only noise. The baseline direction pairs thus serve as a consistency check for the comparison. We note that the fact that for one of the baselines there's a

small preference for the tilted ellipse model indicates that the comparison is not biased toward the superellipse.

We conclude that the measured boundaries are better described as superellipse-shaped than as tilted-ellipse-shaped, which disfavours any response of the form $F(\mathbf{v}^\top \mathbf{A}\mathbf{v})$ as an account of the candidate-direction results. The two curve shapes remain close on any single pair, so this is a population-level conclusion, resting on the comparison between the candidate direction families and the baselines across many pairs.

J Robustness of the exponent estimator

In this appendix, we present two checks on the robustness of our estimator of the superellipse exponent p.

First, we fit synthetic contours with a known exponent. We build a response grid on the same 31×31 angle grid used in our sweeps that satisfies Equation 9 exactly for a chosen p and chosen axis intercepts, and pass it through the same contour extraction, normalisation, and fit. The recovered exponent matches the true one to within 0.004 for p between 1.8 and 3.0, and is unchanged as the within-pair asymmetry $|\alpha_1 - \alpha_2|$ varies from $0°$ to $30°$. The estimator is therefore unbiased over the range we measure, and asymmetry does not act on p through the fit.

Second, we reverse the order of aggregation. Instead of fitting one exponent to the anchor-median response of Equation 2, we fit an exponent to each anchor's own response and average over anchors. Over the same 318 contrastive pairs this gives mean $p = 2.41$ (95% CI $[2.37, 2.45]$) against 2.42 ($[2.36, 2.48]$) for the reported procedure, with a slightly tighter interval. Both intervals here are pair-level bootstraps, which compare the two aggregation orders on equal footing; they are tighter than the direction-level cluster-bootstrap interval of Section 4.1 ($[2.29, 2.62]$), which additionally accounts for pairs sharing directions. The within-pair spread across anchors is large ($\sigma \approx 0.54$), so a single anchor's exponent is noisy, but the two aggregation orders agree on the population.

K Toy model details

K.1 Architecture

The toy model follows Vaintrob (2026). It is a two-layer network with tied weights: an encoder $E \in \mathbb{R}^{H \times d}$ and decoder $D = E^\top$, with a $\tanh^3$ activation function applied element-wise to the hidden layer. The entries of E are drawn i.i.d. from $\{0, +1, -1\}$ with probabilities $\{1 - q, q/2, q/2\}$, where q is a sparsity parameter controlling the fraction of non-zero entries. The network receives a noisy two-hot input $x = x_{\text{clean}} + \epsilon$, where x_{clean} has exactly $s = 2$ active coordinates set to 1 and the remaining set to 0, and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ with $\sigma^2 = 0.03$. The network operates in superposition, with the number of features d exceeding the number of neurons H.

K.2 Optimization

We do not train the encoder and decoder via gradient descent. Instead, we fix the random encoder structure and optimize two scalar parameters: an input gain c_{in} and an output scale c_{out}, so that the network computes $y = c_{\text{out}} \cdot \tanh^3(c_{\text{in}} \cdot Ex)$. We minimize a per-coordinate weighted MSE loss:

$$\mathcal{L} = \frac{1}{B \cdot d} \left[\lambda \sum_{i \in \text{on}} (c_{\text{out}} f_i - 1)^2 + \sum_{i \in \text{off}} (c_{\text{out}} f_i)^2 \right], \tag{14}$$

where $f = D \tanh^3(c_{\text{in}} \cdot Ex)$, "on" denotes coordinates where $x_{\text{clean}} = 1$, and "off" denotes coordinates where $x_{\text{clean}} = 0$. The weighting λ upweights active-coordinate error terms to discourage the trivial solution of outputting zero, which otherwise dominates for very sparse inputs. Given c_{in}, the optimal c_{out} has a closed-form solution:

$$c^*_{\text{out}} = \frac{\lambda \cdot \langle f, x_{\text{clean}} \rangle_{\text{on}}}{\lambda \cdot \|f\|^2_{\text{on}} + \|f\|^2_{\text{off}}}. \tag{15}$$

We sweep c_{in} over a 121-point grid on $[0.05, 8.0]$ and select the value minimizing $\mathcal{L}$.

The sparsity parameter q was swept over $\{0.025, 0.05, 0.075, 0.1, 0.125, 0.15\}$ for each (d, H) configuration, selecting q^* by minimum training loss. The selected values were $q^* = 0.15$ for $4\times$, $q^* = 0.10$ for $8\times$, and $q^* = 0.075$ for $16\times$. The weighting parameter was set to $\lambda \in \{500, 1000, 2000\}$ for $d \in \{4096, 8192, 16384\}$ respectively, scaled roughly as $\lambda \approx d/(s \cdot \mathrm{SNR}^2)$ with $s = 2$ and empirical SNR ≈ 2.

K.3 Perturbation procedure

The active set is fixed to $S = \{0, 1\}$ throughout, so all perturbations are performed around the same clean input $x^* = e_0 + e_1$. We sample 80 pairs (j, k) uniformly without replacement from the inactive coordinates $\{2, 3, \ldots, d-1\}$. For each pair, we perturb along the two ground-truth feature directions e_j and e_k. To measure the effect of misalignment, we rotate e_j and e_k toward random directions by varying amounts. Specifically, for each pair we sample an isotropic random unit pair (w_u, w_v), mutually orthogonal but otherwise unconstrained, and interpolate between the feature direction and the random direction at each misalignment level. The same (w_u, w_v) pair is used across all misalignment levels for a given feature pair.

K.4 Superellipse fitting

For each direction pair and misalignment level, we sweep 60 mixing angles φ uniformly on $[0, \pi/2]$. For each φ, we perform a 2500-point linear scan of the perturbation magnitude α on $[0, 25]$ and identify the smallest α at which $\|f(x^* + \alpha \cdot \hat{d}(\varphi)) - f(x^*)\|_2 > \tau$. The threshold is set to $\tau = 0.5 \cdot \max_{\alpha \le 25} \|f(x^* + \alpha \cdot e_j) - f(x^*)\|_2$ for the first inactive feature $j = 2$, computed per network. The superellipse exponent is estimated by minimizing $\sum_\varphi \left((x_\varphi / r_0)^n + (y_\varphi / r_{\pi/2})^n - 1\right)^2$ over $n \in [0.5, 60]$, where $(x_\varphi, y_\varphi) = \alpha(\varphi) \cdot (\cos\varphi, \sin\varphi)$. The perturbation sweep is deterministic (no noise is added at perturbation time). All random seeds are set to 42. We show in Figure 9 an example of the superelliptical boundaries we obtain for the toy model in the style of Figure 1b in the main text. Figure 10 shows the toy-model analogue of Figure 1a: the L^2 response as a function of perturbation magnitude along two feature axes, their equal-weight combination, and a random-direction baseline. The feature axes break the plateau first ($\alpha \approx 4.1$–4.3) and random directions last ($\alpha = 5.93$); the combination breaks at $\alpha = 4.70$, later than either axis, whereas $p = 2$ would predict it breaks at the same magnitude as the axes.

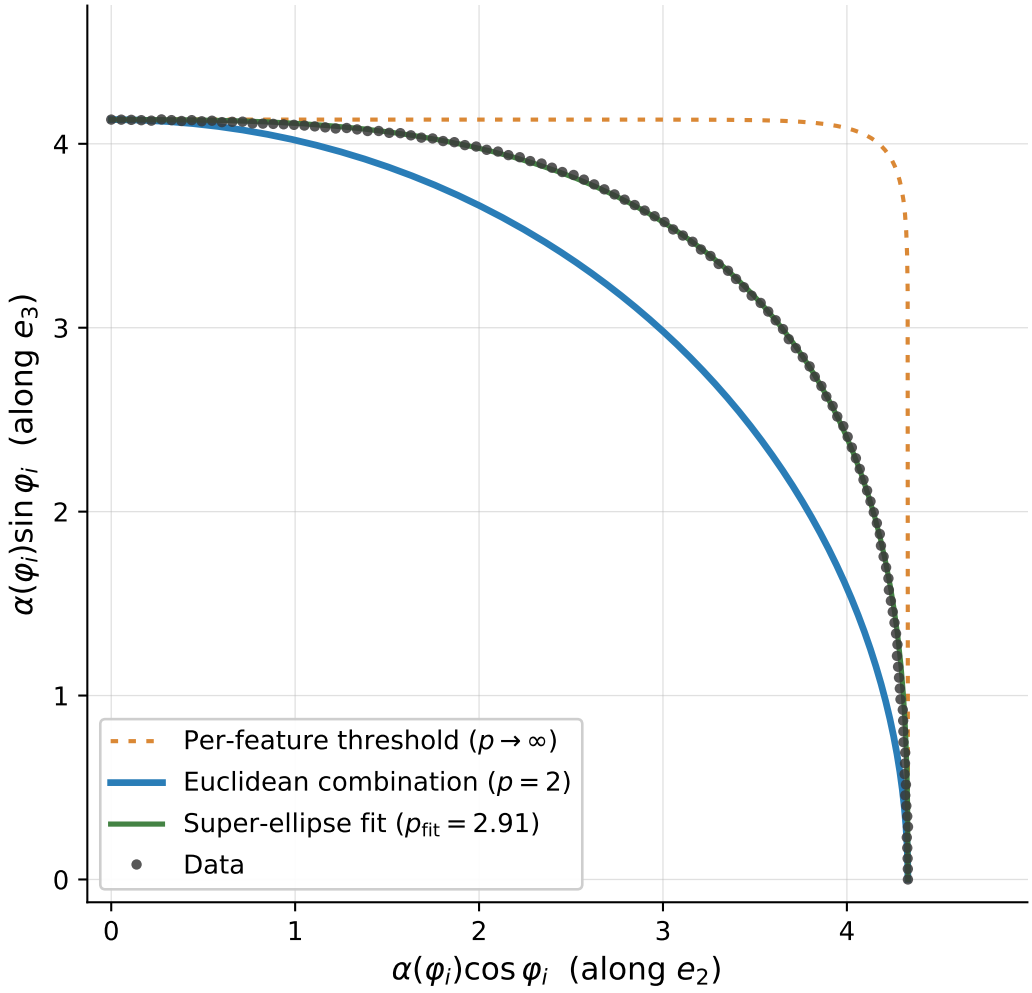


Figure 9: **Iso-plateau boundary.** Plateau-breaking perturbation magnitudes for perturbations aligned with combinations of two feature directions in the toy model with $d = 8192$. The fitted superellipse exponent is 2.91, with a radial fit residual (Section 3) of 0.13%. This single pair sits slightly below the 3–3.4 range quoted in Section 5, which reports medians over 80 pairs.

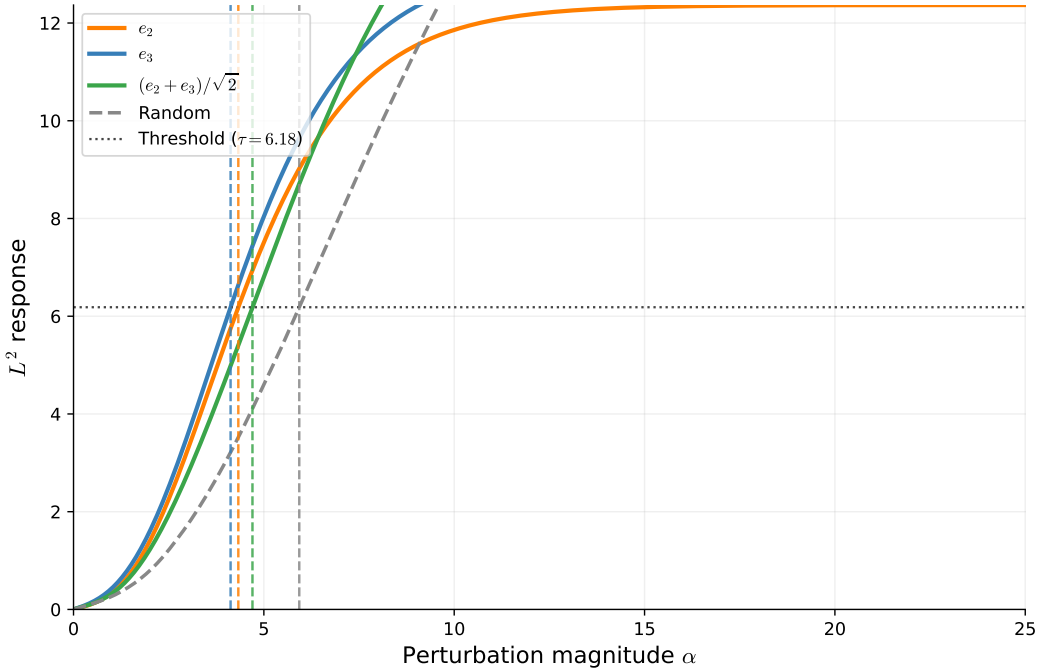


Figure 10: **Toy-model response-versus-magnitude sweep.** L^2 response of the toy model ($d = 8192$, $H = 1024$) to additive perturbations of magnitude α along two inactive feature axes $\mathbf{e}_2$ and $\mathbf{e}_3$, their equal-weight combination, and a random-direction baseline (median over 10 isotropic unit directions), with the plateau-breaking threshold τ (dotted) and each curve's plateau-breaking magnitude (dashed verticals). As in Figure 1a, feature axes break the plateau first, their combination later, and random directions last.