The Spectral Geometry of Misalignment

Aryan Gupta
aryan.cs.app@gmail.com

June 2026

Abstract

Fine-tuning a language model is a perturbation of its weights. We ask whether the *spectrum* of that perturbation reveals that the fine-tune installed a misaligned objective, without any labeled examples of the model behaving badly. We model the weight increment $\Delta W = W_{\text{ft}} - W_{\text{base}}$ of a single layer as a diffuse random bulk plus a low-rank deterministic signal, and apply the Baik–Ben Arous–Péché (BBP) phase transition for spiked covariance matrices. First, a planted signal direction of strength θ produces an eigenvalue that detaches from the Marchenko–Pastur bulk if and only if $\theta > \sqrt{\gamma}$, where γ is the aspect ratio of the layer. Second, and centrally, at *fixed* signal energy τ a perturbation is spectrally detectable if and only if its rank is below a critical value $r_\star = \tau/\sqrt{\gamma}$: a concentrated update crosses the threshold while an energy-matched diffuse update of higher rank stays invisible. Rank, not magnitude, is the discriminator. Third, above threshold the leading sample eigenvector aligns with the planted direction with computable overlap, so the test also yields a label-free estimate of the perturbation direction. We pair these with a calibrated Tracy–Widom null, so the decision rule has controlled false-positive rate. The theorems establish a conditional: *given* that a misaligned fine-tune deposits a low-rank signal and an energy-matched benign fine-tune spreads its budget over higher rank, the spectral test separates them. Whether real misaligned fine-tunes are in fact low-rank at matched energy, while benign ones are diffuse, is an empirical claim, and measuring the benign-side rank is the experiment on which everything rests. We give the construction, the estimator, the confounds that can counterfeit a spike, and where the assumptions of the theory fail for transformer activations.

1 Introduction

A practitioner is handed a fine-tuned language model and asked a hard question: does it carry a hidden, misaligned objective? The strongest available white-box answers come from supervised probes. Trained on labeled examples of a model lying or defecting, a linear probe reads deception from the residual stream at very high accuracy [30, 32]. The difficulty is that the supervision is exactly what one lacks. If one already possessed reliable labels for when a model pursues a hidden goal, the detection problem would be largely solved. The binding constraint in practice is access to such labels, and to the relevant distribution; the recent auditing literature makes this concrete, finding that the decisive resource in surfacing a hidden objective was access to the training data rather than the sophistication of any single interpretability method [31].

This motivates a different question. Fine-tuning is a transformation of the weights. Rather than asking what the model says, we ask what the fine-tune *did* to the weight matrices, and whether that leaves a structural trace that a misaligned update shares and a benign one does not. The object we examine is the weight increment $\Delta W = W_{\text{ft}} - W_{\text{base}}$, and the tool is the spectrum of ΔW. The hope is a test that needs no behavioral labels: a label-free, model-level verdict.

This hope has precedent, and the precedents constrain what we can claim. Heavy-tailed self-regularization reads a label-free, data-free fingerprint of model quality directly off weight spectra, and its taxonomy already includes a "bulk plus spikes" regime [12]. Spectral outlier analysis detects data poisoning from the singular structure of learned representations [29]. Random-matrix tools have very recently been turned on transformer internals for hallucination and out-of-distribution detection [28], and a closely related line ties weight spectra, activation covariance, and fine-tuning together on real language models [14]. None of these targets misalignment, and none isolates the structural question we pose. But they mean our contribution cannot be "use spectra to inspect a model"; that ground is taken. The contribution has to be sharper.

Contributions. We make the following.

1. A random-matrix model of fine-tuning as a low-rank perturbation of a layer's weight increment, with the diffuse part of the update playing the role of the Marchenko–Pastur bulk and the alignment-relevant part the spike (Section 4).
2. A detectability theorem: a signal direction of strength θ is spectrally visible exactly when $\theta > \sqrt{\gamma}$, with the eigenvalue location given in closed form (Proposition 5.1).
3. The central statement, the *rank-at-fixed-energy* discriminator: at matched signal energy, an update is detectable if and only if its rank is below $r_\star = \tau/\sqrt{\gamma}$, so two energy-matched fine-tunes of low and high rank are spectrally separated (Proposition 5.2). This is the part no prior work isolates, because every prior result confounds the magnitude of an update with its structure.
4. A label-free identification result: above threshold the leading sample eigenvector estimates the planted direction with computable overlap (Proposition 5.4), and a calibrated Tracy–Widom test with controlled false-positive rate (Proposition 5.5).
5. A subspace observable, the principal-angle distance between the leading eigenspaces of matched models, intended to capture signal carried by eigenvector rotation rather than eigenvalue magnitude (Definition 5.6, Remark 5.7), calibrated by bootstrap, which addresses a known limitation of scalar spectral statistics.
6. The limitations (Section 8): which assumptions transformer weights and activations violate, why the activation-covariance instrument is weaker than the weight increment, a sign analysis reconciling an apparently contradictory empirical result, and the experiment that the whole program stands or falls on.

What is proved and what is assumed. The theorems of Section 5 prove an *implication*: if a misaligned fine-tune deposits a rank-r_m signal and a benign fine-tune of equal energy spreads its perturbation over rank $r_\mathrm{b} \gg r_\mathrm{m}$, then the spectral test separates the two, with the threshold and the critical rank given explicitly. The *antecedent*, that misaligned updates are concentrated and benign updates of equal energy are diffuse, is an empirical hypothesis. The supporting evidence is suggestive but incomplete: emergent misalignment is mediated by a single, convergent linear direction [18], a single rank-one adapter suffices to induce it [17], and emergent-misalignment weight deltas occupy a shared low-dimensional subspace across tasks [21]. What no prior work measures is the benign side at matched energy. That measurement is the linchpin of the empirical program (Section 7), and the theory is written so that it owns only the implication.

A note on the name. The project is named for an analogy with Fourier inspection of vision models, but the mathematics here is spectral and random-matrix theoretic, not Fourier analytic. A weight matrix carries no canonical periodic axis, so the discrete Fourier transform is not the natural basis; the singular value decomposition is. We use the latter and reserve genuine Fourier analysis for settings with group structure, where it is the correct basis, which is outside the static-detection scope of this document.

2 Setup and notation

Fix a layer with weight matrix $W \in \mathbb{R}^{p\times q}$, and write W_{base} for its value in a reference model and W_{ft} for its value after fine-tuning. The object of study is the increment

$$\Delta W \;=\; W_{\text{ft}} - W_{\text{base}} \;\in\; \mathbb{R}^{p\times q}.$$

Without loss of generality take $q \le p$ (otherwise transpose), and set the aspect ratio

$$\gamma \;=\; q/p \;\in\; (0,1].$$

We treat the p rows of ΔW as p observations in $\mathbb{R}^q$ and form the symmetric positive semidefinite matrix

$$C \;=\; \frac{1}{p}\,\Delta W^\top \Delta W \;\in\; \mathbb{R}^{q\times q}.$$

The nonzero eigenvalues of C are the squared singular values of ΔW divided by p, so the spectrum of C is a faithful and scale-controlled summary of the singular structure of the update. We write $\lambda_1(C) \ge \cdots \ge \lambda_q(C) \ge 0$ for its eigenvalues and $\hat{v}_1, \ldots, \hat{v}_q$ for the corresponding unit eigenvectors.

Energy. The *energy* of the update is its squared Frobenius norm, $\|\Delta W\|_F^2 = \operatorname{Tr}(\Delta W^\top \Delta W) = p\,\operatorname{Tr}(C)$. Energy is a weight-space quantity, and it is the natural budget to hold fixed when comparing two fine-tunes: a benign and a misaligned update that move the weights by the same amount differ, if at all, in *how* they spend that budget across directions, not in how much they spend.

Matched organisms. The comparison the theory is built for involves three models sharing a base: W_{base}; a benign control W_{b} obtained by fine-tuning on an alignment-neutral objective; and a misaligned model W_{m} obtained by an otherwise identical recipe that installs the misaligned objective. Write $\Delta W_{\text{b}} = W_{\text{b}} - W_{\text{base}}$ and $\Delta W_{\text{m}} = W_{\text{m}} - W_{\text{base}}$, with associated $C_{\text{b}}, C_{\text{m}}$. The control is energy-matched, $\|\Delta W_{\text{b}}\|_F = \|\Delta W_{\text{m}}\|_F$, so that any spectral difference is attributable to the structure of the update and not to its magnitude. The need for this control, rather than a comparison to the base model alone, is that *any* fine-tune moves the spectrum; a base-versus-misaligned difference would only certify that fine-tuning occurred.

3 Random-matrix preliminaries

We collect the results we use. They are standard; we state them in the normalization of Section 2 and cite primary sources. Throughout, the asymptotic regime is $p, q \to \infty$ with $q/p \to \gamma \in (0,1]$.

3.1 The bulk

Theorem 3.1 (Marchenko–Pastur law [1, 2])**.** *Let $X \in \mathbb{R}^{p\times q}$ have independent entries with mean 0 and variance σ^2, and let $S = \frac{1}{p}X^\top X$. As $p, q \to \infty$ with $q/p \to \gamma \in (0, 1]$, the empirical spectral distribution of S converges almost surely to the Marchenko–Pastur law with density supported on $[\lambda_-, \lambda_+]$, where*

$$\lambda_\pm \;=\; \sigma^2\,(1 \pm \sqrt{\gamma})^2.$$

In particular the largest eigenvalue converges almost surely to the upper edge $\lambda_+ = \sigma^2(1 + \sqrt{\gamma})^2$.

The upper edge λ_+ is the reference level against which a signal must compete: an eigenvalue is "a spike" precisely when it sits above λ_+.

3.2 The spike

The spiked model adds a fixed low-rank perturbation to an isotropic population covariance [3]. We use the real, almost-sure form of the eigenvalue limit, due to Baik and Silverstein and to Paul; the fluctuation result that names the transition is due to Baik, Ben Arous and Péché.

Theorem 3.2 (Spiked eigenvalue limit [5, 6, 4])**.** *Consider data with population covariance $\Sigma = \sigma^2(I_q + \theta\, vv^\top)$ for a unit vector v and spike strength $\theta > 0$, and let $S = \frac{1}{p}X^\top X$ be the sample covariance from p observations, $q/p \to \gamma \in (0, 1]$. Then*

$$\lambda_1(S) \;\xrightarrow{a.s.}\; \begin{cases} \sigma^2\,(1+\theta)\Big(1 + \dfrac{\gamma}{\theta}\Big), & \theta > \sqrt{\gamma}, \\[2mm] \sigma^2\,(1+\sqrt{\gamma})^2, & \theta \le \sqrt{\gamma}. \end{cases}$$

The transition at $\theta = \sqrt{\gamma}$ is sharp: below it the leading eigenvalue is asymptotically indistinguishable from the bulk edge, and above it the eigenvalue detaches by a deterministic amount. The eigenvector carries the complementary information.

Theorem 3.3 (Spiked eigenvector overlap [6, 7])**.** *Under the conditions of Theorem 3.2, with $\hat{v}_1$ the leading sample eigenvector,*

$$|\langle \hat{v}_1, v\rangle|^2 \;\xrightarrow{a.s.}\; \begin{cases} \dfrac{1 - \gamma/\theta^2}{1 + \gamma/\theta}, & \theta > \sqrt{\gamma}, \\[2mm] 0, & \theta \le \sqrt{\gamma}. \end{cases}$$

The overlap rises continuously from 0 at the threshold to 1 as $\theta \to \infty$.

3.3 The null and the general bulk

To turn "the leading eigenvalue exceeds the edge" into a calibrated decision we need the null fluctuations of the largest eigenvalue when there is no spike.

Theorem 3.4 (Tracy–Widom edge [9, 3])**.** *For the null model $\Sigma = \sigma^2 I_q$ with real Gaussian data and $q/p \to \gamma \in (0, \infty)$, there are explicit centering and scaling sequences $\mu_{p,q}, s_{p,q}$ of order 1 and $p^{-2/3}$ respectively such that*

$$\frac{\lambda_1(S) - \mu_{p,q}}{s_{p,q}} \;\Rightarrow\; \mathrm{TW}_1,$$

where TW_1 is the order-one (orthogonal ensemble) Tracy–Widom law. Concretely, with $n = p$ effective degrees of freedom, $\mu_{p,q} = \frac{1}{p}(\sqrt{p-1} + \sqrt{q})^2$ and $s_{p,q} = \frac{1}{p}(\sqrt{p-1} + \sqrt{q})\big(\frac{1}{\sqrt{p-1}} + \frac{1}{\sqrt{q}}\big)^{1/3}$.

Finally, the perturbation results above presume the bulk is exactly Marchenko–Pastur. When the diffuse part of the update is not isotropic, the bulk is some other compactly supported law μ, and the transition is governed by the Cauchy transform of μ rather than the closed forms above.

Theorem 3.5 (Finite-rank additive perturbation [7])**.** *Let X_p be symmetric with limiting spectral law μ of bounded support and upper edge b, and let P be deterministic, symmetric, rank one, with eigenvalue $\theta > 0$. Let $G_\mu(z) = \int (z-t)^{-1}\, d\mu(t)$ be the Cauchy transform. Then the largest eigenvalue of $X_p + P$ converges almost surely to $G_\mu^{-1}(1/\theta)$ if $\theta > 1/G_\mu(b^+)$, and to b otherwise. A finite, nonzero threshold exists precisely when μ has a square-root edge, $f_\mu(t) \sim c\,(b-t)^{1/2}$ as $t \uparrow b$.*

Theorem 3.5 is the form we fall back on when the Marchenko–Pastur idealization fails; it preserves the qualitative phase transition for any square-root-edge bulk, at the cost of the closed-form constants. The square-root-edge condition is the structural reason a sharp detectability threshold exists at all.

4 A spiked model of fine-tuning

We now place the weight increment inside the spiked framework. The modeling content is a decomposition of the update into a diffuse part and a structured part, together with the assumptions under which the diffuse part behaves as a Marchenko–Pastur bulk.

Definition 4.1 (Diffuse-plus-signal decomposition)**.** Write the increment as

$$\Delta W \;=\; N \;+\; S, \qquad S \;=\; \sum_{i=1}^{r} \beta_i\, a_i b_i^\top,$$

where $N \in \mathbb{R}^{p\times q}$ is the diffuse component, S is the signal of rank $r = \operatorname{rank}(S)$, the $a_i \in \mathbb{R}^p$ and $b_i \in \mathbb{R}^q$ are orthonormal families, and $\beta_1 \geq \cdots \geq \beta_r > 0$. The induced spike strengths in the covariance $C = \frac{1}{p}\Delta W^\top \Delta W$ are $\theta_i = \beta_i^2/(p\,\sigma^2)$, where σ^2 is the per-entry variance of N (defined in Assumption 1).

Assumption 1 (Isotropic diffuse bulk)**.** The diffuse component N has independent, mean-zero entries with common variance σ^2 and finite fourth moment, independent of the signal directions $\{a_i, b_i\}$.

Assumption 1 is the load-bearing modeling choice, and the reason we work with the increment ΔW rather than the raw weight W. Trained weight matrices are empirically heavy-tailed and not Marchenko–Pastur [12], so an isotropic null on W would be mis-specified. The increment of a fine-tune is a far smaller and more nearly unstructured object; for a benign update its bulk is plausibly close to isotropic, which is exactly the regime in which Theorems 3.1–3.4 apply. We return in Section 8 to the cases where even this is too strong, where Theorem 3.5 replaces the closed forms.

Assumption 2 (Structure hypothesis, to be tested)**.** A misalignment-relevant fine-tune deposits its signal in low rank, r_{m} small, while a benign fine-tune of equal energy spreads a comparable Frobenius budget over higher rank, $r_{\mathrm{b}} \gg r_{\mathrm{m}}$.

Assumption 2 is not proved here and is not folklore. It is the empirical hypothesis the program tests, and the theorems below deliver their conclusion *conditional* on it.

Remark 4.2 (Why energy is the right thing to fix). By Definition 4.1, the signal contributes $\|S\|_F^2 = \sum_i \beta_i^2 = p\,\sigma^2 \sum_i \theta_i$ to the energy. Holding the signal energy fixed is therefore holding $\tau := \sum_i \theta_i$ fixed. The discriminator of Section 5 is precisely a statement about how a fixed τ is distributed across r directions, which is why energy matching is built into the matched organisms of Section 2.

5 Main results

5.1 Detectability

Proposition 5.1 (Detectability and location of a single spike)**.** *Under Assumption 1 and the decomposition of Definition 4.1 with a single signal direction of induced strength θ, the leading eigenvalue of $C = \frac{1}{p}\Delta W^\top \Delta W$ satisfies, as $p, q \to \infty$ with $q/p \to \gamma$,*

$$\lambda_1(C) \xrightarrow{a.s.} \begin{cases} \sigma^2(1+\theta)\Big(1+\dfrac{\gamma}{\theta}\Big), & \theta > \sqrt{\gamma}, \\ \sigma^2(1+\sqrt{\gamma})^2, & \theta \le \sqrt{\gamma}. \end{cases}$$

Consequently the signal is asymptotically visible in the spectrum, that is $\lambda_1(C)$ exceeds the bulk edge by a deterministic gap, if and only if $\theta > \sqrt{\gamma}$.

Proof. The increment $\Delta W = N + S$ is a fixed rank-one perturbation of the rectangular random matrix N, whose entries are independent with variance σ^2 by Assumption 1. By the singular-value phase transition for low-rank perturbations of large rectangular random matrices [8], the largest singular value of ΔW detaches from the bulk edge of N exactly when the perturbation strength crosses a threshold, and the squared singular values of ΔW, which are p times the eigenvalues of C, obey the spiked transition. Equivalently, the rows of ΔW have second-moment matrix $\mathbb{E}[C] = \sigma^2 I_q + (\beta^2/p)\, bb^\top = \sigma^2(I_q + \theta\, bb^\top)$ with $\theta = \beta^2/(p\sigma^2)$ as in Definition 4.1, so C is a spiked sample covariance and Theorem 3.2 gives the stated limit and the threshold $\theta > \sqrt{\gamma}$. The two views agree because the squared-singular-value transition of the rectangular model and the spiked-covariance transition share the same threshold and location, a coincidence verified directly in the Wishart specialization of the finite-rank theory [7]. □

5.2 The discriminator: rank at fixed energy

The single-spike statement extends to rank r by applying it to the largest induced strength. The detection of *any* spike is governed by $\max_i \theta_i$, because the leading eigenvalue of C tracks the strongest planted direction.

Proposition 5.2 (Rank-at-fixed-energy discriminator)**.** *Fix the signal energy through $\tau = \sum_{i=1}^r \theta_i$. Among rank-$r$ signals of this energy, the leading induced strength satisfies $\max_i \theta_i \ge \tau/r$, with equality when the energy is spread equally, $\theta_i = \tau/r$. Therefore:*

1. *An equal-energy, equal-strength rank-r update is spectrally detectable (Proposition 5.1) if and only if*
$$\frac{\tau}{r} > \sqrt{\gamma} \quad \Longleftrightarrow \quad r < r_\star, \qquad r_\star := \frac{\tau}{\sqrt{\gamma}}.$$
2. *Consequently, a concentrated update with $r_{\mathrm{m}} < r_\star$ produces a supercritical spike above the bulk edge, while an energy-matched diffuse update that spreads its budget approximately uniformly over rank $r_{\mathrm{b}} \ge r_\star$ keeps every direction subcritical and is asymptotically invisible. The test separates a concentrated update from a diffuse one of equal energy; a high-rank update that nonetheless places a supercritical share of its energy in a single direction is still detected.*

Proof. The inequality $\max_i \theta_i \ge \tau/r$ is the pigeonhole bound on a sum of r nonnegative terms, with equality at the uniform allocation $\theta_i = \tau/r$. For the uniform allocation, every spike has strength τ/r, so by Proposition 5.1 a detached eigenvalue exists if and only if $\tau/r > \sqrt{\gamma}$, equivalently

$r < \tau/\sqrt{\gamma} = r_\star$. For statement (2): when $r_{\mathrm{m}} < r_\star$ the leading strength exceeds $\sqrt{\gamma}$ and a spike appears, while a uniform allocation over $r_{\mathrm{b}} \geq r_\star$ puts every strength at or below $\sqrt{\gamma}$ and no eigenvalue detaches. The pigeonhole bound is one-sided, so invisibility requires the diffuse, near-uniform allocation rather than $r_{\mathrm{b}} \geq r_\star$ alone; a high-rank update concentrated in one direction has $\max_i \theta_i > \sqrt{\gamma}$ and is detected. □

Proposition 5.2 is the conceptual core. At fixed magnitude, detectability is governed by rank: a concentrated update is supercritical, an energy-matched diffuse update is not, and the boundary is the critical rank $r_\star = \tau/\sqrt{\gamma}$. Conditional on Assumption 2, which assigns the misaligned fine-tune the low rank and the benign one the high rank, this is a label-free separation. The proposition proves the implication, and the experiment of Section 7 tests whether the ranks fall on the asserted sides of $r_\star$.

Remark 5.3 (The uniform allocation is the benign best case). Spreading energy uniformly is the allocation that minimizes the leading strength, hence the hardest case to detect. A benign update need not be uniform, but if even the uniform allocation of its energy would be detectable then so is the update; the proposition states the favorable-to-benign boundary. The substantive empirical question is whether benign fine-tuning is diffuse *enough*, that is whether $r_{\mathrm{b}} \geq r_\star$, not whether it is exactly uniform.

5.3 Critical rank for representative layers

The critical rank $r_\star = \tau/\sqrt{\gamma}$ depends on the layer through its aspect ratio γ and on the fine-tune through the signal energy τ. Table 1 reports γ and $r_\star$ for the weight shapes of a standard 7B transformer (hidden width 4096, feed-forward width 11008), at two illustrative signal energies.

Layer	shape $(p \times q)$	$\gamma = q/p$	$r_\star$ at $\tau = 4,\ 16$
Attention $\{q, k, v, o\}$ projection	4096×4096	1.00	4.0, 16.0
MLP up / gate projection	11008×4096	0.372	6.6, 26.2
MLP down projection	11008×4096	0.372	6.6, 26.2

Table 1: Critical rank $r_\star = \tau/\sqrt{\gamma}$ across layer shapes. Shapes are written as (rows × columns) after the transpose of Section 2 that places $q \leq p$, so $\gamma = q/p$ is the smaller dimension over the larger. A misalignment update of rank r is detectable in a layer if and only if $r < r_\star$ for that layer.

First, the rectangular feed-forward layers, with smaller γ, tolerate a higher-rank signal before it disappears into the bulk, so they are the more forgiving place to detect a concentrated update; the square attention projections impose the strictest concentration requirement. Second, the rank-one adapter that suffices to induce emergent misalignment [17] sits at $r = 1$, far below $r_\star$ in every layer, so its signal is well above threshold, whereas a benign update must spread the same energy over at least $r_\star$ directions to stay hidden. The estimator of Section 6 therefore scans layers and reports the per-layer standardized leading eigenvalue, with the prior that a signal, if present, is clearest where γ is small and the layer's signal energy is large.

5.4 Label-free identification

Detection establishes that a spike exists; the eigenvector identifies its direction.

Proposition 5.4 (Label-free direction estimate)**.** *Under the conditions of Proposition 5.1 with $\theta > \sqrt{\gamma}$, the leading eigenvector $\hat{v}_1$ of C satisfies*

$$|\langle \hat{v}_1, b \rangle|^2 \xrightarrow{a.s.} \frac{1 - \gamma/\theta^2}{1 + \gamma/\theta} > 0,$$

where b is the planted signal direction. Thus $\hat{v}_1$ is a consistent-up-to-bias estimate of the perturbation direction, computable from the model alone.

Proof. Immediate from Theorem 3.3 applied to the effective spiked covariance $\sigma^2(I_q + \theta\, bb^\top)$ identified in the proof of Proposition 5.1. □

Proposition 5.4 matters for validation. If the recovered direction $\hat{v}_1$ is the misalignment direction, then steering the model along $\pm\hat{v}_1$ should modulate the misaligned behavior, a causal check that the spectral signal is the behavior and not an artifact. This connects the spectral test to the established finding that emergent misalignment is steerable along a single recovered direction [18, 20], with the difference that here the direction is obtained from the spectrum without alignment labels.

5.5 A calibrated test

Proposition 5.5 (Level-α spike test)**.** *Let H_0 be the pure-noise null, $\Delta W = N$ with no signal (Assumption 1 with $\theta = 0$). Using the centering and scaling of Theorem 3.4, the test that rejects H_0 when*

$$\frac{\lambda_1(C) - \mu_{p,q}}{s_{p,q}} > t_{1-\alpha}, \qquad t_{1-\alpha} = \mathrm{TW}_1 \text{ quantile},$$

has asymptotic false-positive rate α. Its resolution near the threshold is $O(p^{-2/3})$: a spike within $O(p^{-2/3})$ of the bulk edge is not distinguishable from the null.

Proof. Under H_0 the leading eigenvalue obeys the Tracy–Widom limit of Theorem 3.4, so the standardized statistic exceeds $t_{1-\alpha}$ with probability α in the limit. The resolution statement is the $p^{-2/3}$ scale of the fluctuations: the deterministic gap opened by a spike at strength $\sqrt{\gamma} + \epsilon$ is $O(\epsilon)$ by Proposition 5.1, which is detectable against $O(p^{-2/3})$ noise only for $\epsilon \gtrsim p^{-2/3}$. □

Proposition 5.5 supplies the missing piece that purely descriptive spectral statistics lack: a decision rule with a controlled error rate and a stated resolution, derived from the null rather than from a labeled validation set.

5.6 A subspace observable for rotation

Scalar spectral statistics, such as effective rank or the leading eigenvalue, are functions of the eigenvalues alone. It has been suggested that post-training reshapes representation geometry in ways a magnitude test need not capture [27], so we add a complementary observable on the eigenvectors.

Definition 5.6 (Leading-subspace principal-angle distance)**.** For two models with covariances C, C', let V_k, V_k' be the matrices of their top-k eigenvectors and let $\rho_1 \geq \cdots \geq \rho_k$ be the singular values of $V_k^\top V_k'$, equal to the cosines of the principal angles between the two subspaces. Define the Grassmann distance

$$d_k(C, C') = \Big(\sum_{j=1}^{k} \arccos^2(\rho_j) \Big)^{1/2}.$$

Remark 5.7 (Calibrating the rotation observable). The distance d_k is sensitive to signal that a magnitude test misses, but it does not inherit the closed-form null of Proposition 5.5, and the right contrast is not a spiked model against a no-spike model. The leading eigenvector of a no-spike covariance is a delocalized edge vector, so a spiked and an unspiked model are already nearly orthogonal in their leading directions, and d_1 sits near its maximal value whether or not a spike is present. The informative comparison is between two matched models, the misaligned candidate and its energy-matched benign control, where a systematic excess of d_k beyond finite-sample fluctuation indicates that one model has rotated a leading direction the other has not. The null here, that the two share a leading subspace up to sampling noise, has no simple closed form, so we calibrate d_k by the bootstrap and permutation of Section 6.2: resample probe inputs for the activation instrument, or permute increment entries for the weight instrument, to obtain the distribution of d_k under no differential rotation, and report a systematic excess. An asymptotic theory of d_k under a joint spiked model of the two increments is left to future work.

The subspace observable is label-free in the same sense as the others: it compares two models, not a model against labeled behavior. It is the natural instrument when the discriminator of interest is reorientation of existing capacity rather than addition of new energy.

6 From theory to a label-free estimator

The propositions describe population limits. We now specify the finite-sample procedure, the secondary instrument on activations, and the confounds that can counterfeit a spike.

6.1 The procedure

1. **Form the increment.** For each layer of interest compute $\Delta W = W_{\text{ft}} - W_{\text{base}}$ and the covariance $C = \frac{1}{p}\Delta W^\top \Delta W$. Record the aspect ratio $\gamma = q/p$.

2. **Estimate the bulk.** Fit the noise level σ^2 from the bulk of the spectrum (for example by matching the median eigenvalue to the Marchenko–Pastur median), not from the full trace, so that a genuine spike does not inflate the estimated floor.

3. **Test for a spike.** Apply Proposition 5.5 with the fitted σ^2. Report the standardized leading eigenvalue, the implied spike strength $\hat{\theta}$ via inversion of Proposition 5.1, and the number of eigenvalues above the edge as an estimate of the signal rank.

4. **Match energy.** When comparing a misaligned candidate to a benign control, rescale so $\|\Delta W_{\text{m}}\|_F = \|\Delta W_{\text{b}}\|_F$ before comparing spectra, so the comparison is at fixed budget as Proposition 5.2 requires.

5. **Recover and test the direction.** Take $\hat{v}_1$ from Proposition 5.4, map it back to a residual-stream direction, and verify causally by steering. A direction that modulates the behavior is evidence the spike is the behavior.

6. **Check rotation.** Compute the subspace distance of Definition 5.6 between matched models, to catch signal carried by reorientation rather than magnitude.

6.2 Calibration by permutation

The Tracy–Widom calibration of Proposition 5.5 is exact only in the Gaussian limit, and real increments are neither Gaussian nor infinite. We therefore calibrate the null empirically and use

the Tracy–Widom law only as the asymptotic shape the empirical null approaches. Independently permute the entries of ΔW, or independently flip their signs, which destroys any low-rank alignment among the entries while preserving the entrywise marginal distribution exactly. Recompute $\lambda_1(C)$ for many such surrogates to obtain an empirical null, and report the position of the observed $\lambda_1(C)$ in that null as a p-value. The permutation null requires no distributional assumption, absorbs the finite (p, q) and the true entry distribution automatically, and degenerates to the Tracy–Widom law as $p, q \to \infty$. A spike that does not clear the permutation null is not reported, whatever its raw magnitude.

6.3 The activation-covariance instrument, and why it is secondary

One can run the same machinery on the activation covariance $\Sigma_{\text{act}} = \frac{1}{n}\sum_t h_t h_t^\top$ of residual-stream states h_t over n probe tokens, comparing matched models. This is attractive because the causally relevant misalignment direction is known to live cleanly in activation space; in the emergent-misalignment organisms the weight-space adapter direction and the causal activation direction are nearly orthogonal, with reported cosine near 0.04, yet ablating the activation direction removes most of the behavior [18]. The activation covariance is therefore the right object for *identification*.

For *detection*, however, it is the weaker substrate, and we treat it as a heuristic secondary instrument rather than a setting where Propositions 5.1–5.5 hold as stated. The reason is that the activation covariance is strongly anisotropic even before any fine-tuning: a handful of directions dominate, and a small number of coordinates carry outsized variance [37, 38, 39]. The isotropic-bulk Assumption 1 fails, the closed forms of Theorem 3.2 no longer apply, and one must instead separate the misalignment spike from a deformed, model-specific bulk edge using the generalized spiked model for a non-identity base covariance [10]. The clean criterion "crosses the Marchenko–Pastur edge" becomes "separates from an estimated deformed edge," which is both weaker and more fragile. The weight increment keeps the isotropy assumption defensible; the activation covariance does not.

6.4 A sign analysis: reconciling an apparently opposite result

A recent result reports that fine-tuning on safety-degrading data *raises* the effective rank of a model's inference-time activations on harmful prompts, relative to benign tuning [26]. Read naively, "misaligned means higher rank" is the opposite of our "misaligned means a low-rank spike." The tension dissolves once the object is fixed.

Effective rank [15] is $\exp(H)$ where $H = -\sum_i \pi_i \log \pi_i$ is the entropy of the normalized spectrum $\pi_i = \lambda_i / \sum_j \lambda_j$. Adding a single dominant spike to a covariance concentrates the distribution π, lowers H, and therefore *lowers* effective rank. So our model predicts that a low-rank misalignment signal, seen in the relevant covariance, lowers effective rank, which is what one would expect of a concentrated perturbation.

The reported result concerns a different matrix. It measures the effective rank of the inference-time activation matrix on harmful prompts, in a model that was fine-tuned on harmful *content*, with no energy control. Our prediction concerns the rank of the weight increment ΔW, the cause, not the diversity of downstream activations on triggering inputs, the effect. A low-rank cause is entirely compatible with a higher-rank effect: a concentrated update can broaden the range of behaviors the model expresses on the inputs that excite it. The two measurements are not the same quantity and do not have to share a sign. We therefore commit to reporting both, the weight-increment spike rank and the inference-time activation effective rank, and we do not claim more reconciliation than this sign analysis supports.

6.5 Confounds that counterfeit a spike

Several artifacts can produce a leading-eigenvalue excursion that is not a misalignment signal. The estimator must control each.

- **Outlier coordinates.** A few residual-stream coordinates with very large variance [37] create a top eigenvalue unrelated to any fine-tune. Control by per-coordinate standardization or robust covariance before spectral analysis, and by working on ΔW where the effect is weaker.
- **Energy leakage.** An unmatched comparison detects "trained more," not "trained misaligned." Control by the energy matching of step 4.
- **Aspect-ratio regime.** For very wide matrices or many probe tokens, $\gamma \to 0$, the bulk collapses to a point and *every* spike is nominally supercritical ($\sqrt{\gamma} \to 0$). In that regime detectability is trivial and the benign-versus- misaligned separation rests entirely on the rank argument of Proposition 5.2, not on the threshold. Report γ explicitly and interpret the threshold only where it is nontrivial.
- **Heavy-tailed bulk.** On raw weights the bulk is not Marchenko–Pastur [12]; the procedure is defined on ΔW for this reason, and the bulk fit of step 2 should be checked against the Marchenko–Pastur shape, falling back to Theorem 3.5 if it deviates.

7 The empirical program in brief

The theory makes the experiment unambiguous. We state only the commitments the theory forces; the full protocol is the project plan.

The linchpin. Measure the rank of ΔW at matched energy for a benign control and a misaligned model sharing a base, and locate both relative to $r_\star = \tau/\sqrt{\gamma}$. Assumption 2 predicts $r_{\mathrm{m}} < r_\star \le r_{\mathrm{b}}$. If instead the benign control is also low-rank at matched energy, the separation of Proposition 5.2 is empty and the approach fails for that organism. This single measurement is the one the program cannot avoid, and it is the one no prior work reports.

Substrate. The matched comparison requires *full* fine-tuning, not low-rank adaptation. A rank-constrained adapter fixes the rank of ΔW by construction, so it cannot exhibit a rank difference, and its increment has no bulk for a spike to cross. Emergent-misalignment organisms exist as full fine-tunes and as rank-one adapters [16, 17]; the rank-one adapters are useful for the identification and steering checks of Proposition 5.4, but the detection claim of Propositions 5.1–5.2 must be evaluated on full fine-tunes with a matched benign control trained under an identical recipe.

Prediction. Where Assumption 2 holds, the misaligned model shows a supercritical leading eigenvalue in ΔW at a layer where the benign control does not, the recovered $\hat{v}_1$ steers the misaligned behavior, and the subspace distance between misaligned and benign models exceeds the null. Where it fails, one or more of these do not occur, and the theory says exactly which.

8 Assumptions, limitations, and threats to validity

We list the ways the program can be wrong, beyond the confounds of Section 6.5.

1. **The antecedent may be false.** Assumption 2 is the bet. If misaligned and benign fine-tunes are equally concentrated at matched energy, rank does not discriminate and the method has no signal. This is measurable, and measuring it is the linchpin.

2. **Anisotropy on activations.** The isotropic-bulk assumption is defensible for ΔW and not for activation covariance, which is why the latter is secondary (Section 6.3). The generalized spiked model is the correct tool there and is not developed in this document.

3. **Heavy tails on raw weights.** The Marchenko–Pastur null is for the increment, not the trained weight [12]. Applying the threshold to raw weights would be mis-specified.

4. **Narrow versus broad misalignment.** Broad emergent misalignment may occupy a different, more concentrated region of weight space than a narrowly scoped misalignment [16]; the spectral signature may be present for the former and weak for the latter. The method's reach is an empirical question of which misalignments are low-rank.

5. **Deception is not misalignment.** Probes that read a gap between a model's internal belief and its statement catch instrumental liars but largely miss models with an internalized objective, which present no such gap [36]. Our target is the internalized case, which is the regime where behavioral probes are weakest; this is the method's positioning, and it also means the method cannot borrow a belief-gap signal. The spectral signal, if it exists, is structural, present in the weights whether or not the model ever "intends" to deceive.

6. **Evasion.** A spectral monitor is still a monitor. A model trained against it, or with knowledge of it, may be able to spread its misalignment update over higher rank to stay below threshold, the precise move the rank discriminator describes, or to push signal into a benign-looking subspace, as latent-space monitors have been evaded elsewhere [35]. We claim no robustness to an adaptive adversary and treat robustness as a hypothesis to be tested against such an adversary, not a property of the method.

7. **Finite-size and regime gaps.** The limits are asymptotic. Real layers have finite p, q, the increment is one "sample" in a sense the covariance construction has to make precise, and the Tracy–Widom calibration assumes Gaussianity that holds only approximately. These affect the constants and the resolution, and the calibration should be checked by simulation on shuffled increments.

9 Related work

Spectral fingerprints of models. Heavy-tailed self-regularization reads model quality from weight spectra without data or labels, and names a "bulk plus spikes" phase [12]. Our framing inherits its vocabulary and turns it to a different target, misalignment, with a different object, the increment, and a different theory, the BBP threshold rather than a power-law tail index. The closest prior work ties weight spectra, activation covariance, and fine-tuning together on transformer language models and finds that fine-tuning refines the model in the small-singular-value directions [14]; we differ in asking for a benign-versus-misaligned contrast at matched energy, which that work does not consider. Spectral fingerprints have also been used for model lineage in a way deliberately invariant to fine-tuning [40], the opposite use to ours. Random-matrix descriptions of network spectra have precedent in the analysis of loss-surface Hessians [11] and in the local spectral statistics of trained weight matrices, which stay random-matrix-like even where the global density is not [13].

Spectral and random-matrix detection. Spectral outlier scores detect data poisoning from learned representations, per input rather than per model [29]. Random-matrix statistics have been applied to hallucination and out-of-distribution detection in language models, again per input [28]. Neither addresses a per-model misalignment verdict, and neither uses the rank-at-fixed-energy argument.

The structure of emergent misalignment. Emergent misalignment, in which narrow fine-tuning induces broad misalignment [16], is mediated by a single convergent linear direction [18], inducible by a rank-one adapter [17], controllable through a small set of persona features [19, 20], and supported in weight space by a shared low-dimensional subspace across tasks [21]. This body of work establishes the low-rank *structure* we rely on, with labels; our contribution is to detect it without labels, to model it as a BBP-crossing spike, and to make rank at fixed energy the discriminator. The single-direction phenomenon is part of a broader pattern in which alignment-relevant behaviors are low-dimensional, as for refusal [22], and fine-tuning itself is low intrinsic dimension [23, 24].

Geometry of fine-tuning and safety. A complementary line analyzes how fine-tuning breaks safety through low-rank curvature subspaces and argues that benign tasks can also degrade alignment [25]; this is a foil for the clean separation we hypothesize and a reason to take the benign-side measurement seriously. The representation-geometry of post-training has been mapped with effective-rank and spectral-slope observables, establishing that post-training changes geometry [27]; we position our claim narrowly as detection in matched models, not the more general and now-established statement that post-training moves the spectrum.

Behavioural and label-based detection. Supervised probes detect deception at high accuracy when labels and the distribution are available [30, 32], and interpretability-driven audits surface hidden objectives given sufficient access [31]. Unsupervised methods that search for latent knowledge exist [33], but arbitrary features can satisfy their consistency constraints [34]. Our method is complementary: it targets the label-free, unknown-distribution regime where the supervised results do not apply, and where labels exist and the distribution is known a probe should be expected to do better.

10 Conclusion

We have given a random-matrix account of when fine-tuning is spectrally visible, and used it to isolate a single sharp claim: at controlled energy, rank discriminates a concentrated update from a diffuse one, with an explicit critical rank $r_\star = \tau/\sqrt{\gamma}$ and a calibrated test. The theorems prove that *if* misaligned fine-tunes are low-rank and energy-matched benign fine-tunes are diffuse, the spectral test separates them and recovers the misalignment direction without labels. The hypothesis that real fine-tunes fall on the asserted sides of $r_\star$ is the one the empirical program must establish, and the benign-side rank at matched energy is the measurement on which the entire approach rests. We have converted a vague hope, that spectra reveal misalignment, into a falsifiable conditional with a named experiment that can refute it.

References

[1] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Math. USSR-Sbornik* **1**(4):457–483, 1967.

[2] Z. Bai and J. W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed., Springer, 2010.

[3] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Ann. Statist.* **29**(2):295–327, 2001.

[4] J. Baik, G. Ben Arous, and S. Péché, "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," *Ann. Probab.* **33**(5):1643–1697, 2005.

[5] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *J. Multivariate Anal.* **97**(6):1382–1408, 2006.

[6] D. Paul, "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," *Statist. Sinica* **17**:1617–1642, 2007.

[7] F. Benaych-Georges and R. R. Nadakuditi, "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices," *Adv. Math.* **227**(1):494–521, 2011.

[8] F. Benaych-Georges and R. R. Nadakuditi, "The singular values and vectors of low rank perturbations of large rectangular random matrices," *J. Multivariate Anal.* **111**:120–135, 2012.

[9] C. A. Tracy and H. Widom, "On orthogonal and symplectic matrix ensembles," *Comm. Math. Phys.* **177**(3):727–754, 1996.

[10] Z. Bai and J. Yao, "On sample eigenvalues in a generalized spiked population model," *J. Multivariate Anal.* **106**:167–177, 2012.

[11] J. Pennington and Y. Bahri, "Geometry of neural network loss surfaces via random matrix theory," *ICML*, PMLR 70:2798–2806, 2017.

[12] C. H. Martin and M. W. Mahoney, "Implicit self-regularization in deep neural networks: evidence from random matrix theory and implications for learning," *J. Mach. Learn. Res.* **22**(165):1–73, 2021.

[13] M. Thamm, M. Staats, and B. Rosenow, "Random matrix analysis of deep neural network weight matrices," *Phys. Rev. E* **106**:054124, 2022.

[14] M. Staats, M. Thamm, and B. Rosenow, "Small singular values matter: a random matrix analysis of transformer models," arXiv:2410.17770, 2024.

[15] O. Roy and M. Vetterli, "The effective rank: a measure of effective dimensionality," *European Signal Processing Conference (EUSIPCO)*, 606–610, 2007.

[16] J. Betley, D. Tan, N. Warncke, A. Sztyber-Betley, X. Bao, M. Soto, N. Labenz, and O. Evans, "Emergent misalignment: narrow finetuning can produce broadly misaligned LLMs," *ICML*, 2025. arXiv:2502.17424.

[17] E. Turner, A. Soligo, M. Taylor, S. Rajamanoharan, and N. Nanda, "Model organisms for emergent misalignment," arXiv:2506.11613, 2025.

[18] A. Soligo, E. Turner, S. Rajamanoharan, and N. Nanda, "Convergent linear representations of emergent misalignment," arXiv:2506.11618, 2025.

[19] M. Wang, T. Dupré la Tour, O. Watkins, A. Makelov, R. A. Chi, S. Miserendino, J. Wang, A. Rajaram, J. Heidecke, T. Patwardhan, and D. Mossing, "Persona features control emergent misalignment," arXiv:2506.19823, 2025.

[20] R. Chen, A. Arditi, H. Sleight, O. Evans, and J. Lindsey, "Persona vectors: monitoring and controlling character traits in language models," arXiv:2507.21509, 2025.

[21] D. A. R. Arturi, E. Zhang, A. Ansah, K. Zhu, A. Panda, and A. Balwani, "Shared parameter subspaces and cross-task linearity in emergently misaligned behavior," arXiv:2511.02022, 2025.

[22] A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda, "Refusal in language models is mediated by a single direction," *NeurIPS*, 2024. arXiv:2406.11717.

[23] A. Aghajanyan, S. Gupta, and L. Zettlemoyer, "Intrinsic dimensionality explains the effectiveness of language model fine-tuning," *ACL-IJCNLP*, 2021. arXiv:2012.13255.

[24] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: low-rank adaptation of large language models," *ICLR*, 2022. arXiv:2106.09685.

[25] M. Springer et al., "The geometry of alignment collapse: when fine-tuning breaks safety," arXiv:2602.15799, 2026. Preprint.

[26] H. Li, L. Li, Z. Lu, X. Wei, R. Li, J. Shao, and L. Sha, "Layer-aware representation filtering: purifying finetuning data to preserve LLM safety alignment," *EMNLP*, 2025. arXiv:2507.18631.

[27] Li et al., "Tracing the representation geometry of language models from pretraining to post-training," arXiv:2509.23024, 2025.

[28] D. Ettori, "Spectral geometry for deep learning: compression and hallucination detection via random matrix theory," MSc thesis, University of Illinois Chicago, 2026. arXiv:2601.17357. Preprint.

[29] B. Tran, J. Li, and A. Mądry, "Spectral signatures in backdoor attacks," *NeurIPS*, 2018. arXiv:1811.00636.

[30] N. Goldowsky-Dill, B. Chughtai, S. Heimersheim, and M. Hobbhahn, "Detecting strategic deception using linear probes," arXiv:2502.03407, 2025.

[31] S. Marks, J. Treutlein, et al., "Auditing language models for hidden objectives," arXiv:2503.10965, 2025.

[32] M. MacDiarmid, T. Maxwell, et al., "Simple probes can catch sleeper agents," Anthropic, 2024.

[33] C. Burns, H. Ye, D. Klein, and J. Steinhardt, "Discovering latent knowledge in language models without supervision," *ICLR*, 2023. arXiv:2212.03827.

[34] S. Farquhar, V. Varma, Z. Kenton, J. Gasteiger, V. Mikulik, and R. Shah, "Challenges with unsupervised LLM knowledge discovery," arXiv:2312.10029, 2023.

[35] L. Bailey, A. Serrano, A. Sheshadri, M. Seleznyov, et al., "Obfuscated activations bypass LLM latent-space defenses," arXiv:2412.09565, 2024.

[36] K. Haralambiev, "Why safety probes catch liars but miss fanatics," arXiv:2603.25861, 2026. Preprint.

[37] M. Sun, X. Chen, J. Z. Kolter, and Z. Liu, "Massive activations in large language models," *COLM*, 2024. arXiv:2402.17762.

[38] J. Gao, D. He, X. Tan, T. Qin, L. Wang, and T.-Y. Liu, "Representation degeneration problem in training natural language generation models," *ICLR*, 2019. arXiv:1907.12009.

[39] K. Ethayarajh, "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings," *EMNLP-IJCNLP*, 2019.

[40] S. Wang et al., "Ghost in the transformer: detecting model reuse with invariant spectral signatures," *AAAI*, 2026. arXiv:2511.06390.