V1.0.0

Latent Manifold Inheritance: A Geometric Account of Subliminal Trait Transfer in Fine-Tuned Language Models

Rowan Brad Quni-Gudzinas

QNFO-ULA

ABSTRACT

Subliminal learning transfers a teacher model's behavioral traits to a student through training data that carries no semantic trace of the trait, and it evades output-level inspection, creating an unresolved safety problem for any organization that fine-tunes on model-generated data. Existing accounts identify the phenomenon and its brittleness but do not explain why transfer requires weight updates, why it is architecture-specific, and why token-level classifiers miss it. We propose that subliminal transfer is inheritance of a geometric manifold: a conditioned teacher's hidden trait warps its latent space along a low-dimensional direction, and fine-tuning on innocuous outputs forces the student's internal trajectories to track the teacher's, reproducing the warp. We formalize this with a rank-one teacher-warp model, derive a trajectory-matching condition for leakage, and compute a worked example: with per-sample bias signal $\varepsilon = 10^{-3}$, gradient noise scale $\sigma = 1$, effective task-subspace dimension $d_{\text{eff}} = 16$, and $N = 10^4$ carrier samples, the accumulated update reaches cosine alignment ≈ 0.0250 with the trait direction; alignment 0.5 requires $N \approx 5.33 \times 10^6$ samples and alignment 0.1 requires $N \approx 1.62 \times 10^5$. The account explains architecture specificity through subspace alignment, consistent with published evidence that independently initialized students fail to transfer even when global similarity exceeds 0.9, and it predicts that internal-state geometric diagnostics, not output classifiers, detect leakage. We propose Latent Space Tomography as the corresponding measurement protocol.

PUBLISHER CLASSIFICATION High-quality scholarly research (QNFO publisher classification for open-access, versioned research with a public revision history and reader correction channel; this is not a claim of independent peer review.)

PEER-REVIEW STATUS Not independently peer-reviewed

CITE AS Quni-Gudzinas, R. B. (n.d.). Latent Manifold Inheritance: A Geometric Account of Subliminal Trait Transfer in Fine-Tuned Language Models. QNFO Preprint v1.0.0.

1. Introduction

Organizations that deploy large language models increasingly train students on outputs generated by other models: distillation pipelines, synthetic-data curricula, and judge-scored reinforcement learning all pass data from one model into another's weight updates. A documented failure mode of this practice is subliminal learning: a teacher conditioned on a hidden preference — an animal preference, a political leaning, or a misaligned behavioral disposition — generates data that looks semantically clean, such as number sequences, yet a student fine-tuned on that data acquires the teacher's trait [1], [7]. The same protocol has been shown to transfer misalignment and unsafe behavior, which makes it a direct safety concern for any team that cannot fully audit the provenance of its training corpus [7]. The cost of the current gap is concrete: a training set that passes every semantic inspection — human review, keyword filtering, output-token classifiers — can still install an unintended behavioral disposition in the deployed student, and the disposition surfaces only after deployment. A principled understanding of how fine-tuning reshapes the student's latent space would change this: it would tell auditors where to look (internal activation geometry rather than output tokens), tell pipeline owners which design choices (initialization, layer freezing, regularization) suppress leakage, and give regulators a measurable quantity to gate on.

Three structural facts about the phenomenon lack a unified explanation. First, transfer requires weight updates: reading the teacher's outputs does not transmit the trait, but fine-tuning on them does. Second, transfer is architecture- and initialization-specific: students that share the teacher's base checkpoint transfer reliably, while independently initialized students fail even when their global representational similarity to the teacher is high [6]. Third, the signal evades output-token classifiers, which suggests it does not live in the token stream at all.

If the trait is carried by the geometry of the teacher's latent space rather than by discrete output content, all three facts follow. A conditioned teacher's hidden trait warps its internal representations along a low-dimensional direction; fine-tuning on the teacher's outputs is trajectory matching, in which the student's parameters move to reproduce the teacher's input–output behavior, and the movement necessarily has a component along the teacher's warp direction; and because the warp lives in high-dimensional activation states, output-token statistics carry only a faint shadow of it. This paper develops that account, formalizes it, and derives quantitative conditions under which leakage is generic.

The result matters beyond machine-learning safety. Any pipeline in which one learned system's outputs train another — including protein and molecule design systems whose generative latent spaces condition downstream experimental campaigns [2], [3], [4] — inherits a version of the same question: what does a student inherit from a teacher's latent geometry, and can it be measured before deployment? The geometric diagnostics proposed here apply wherever activation point clouds can be extracted.

Contributions:

- A rank-one teacher-warp model in which a hidden trait corresponds to a low-dimensional direction in the teacher's activation space, and a derivation of the trajectory-matching condition under which fine-tuning on teacher outputs leaks the warp into the student (Section 3, Section 4).
- A worked numerical derivation of leakage amplitude and sample complexity: alignment ≈ 0.0250 at $N = 10^4$ samples under stated assumptions, $N \approx 5.33 \times 10^6$ samples for alignment 0.5, and $N \approx 1.62 \times 10^5$ samples for alignment 0.1 (Section 4).
- A reconciliation of the architecture-specificity of subliminal transfer with subspace alignment rather than global similarity, using published evidence that transfer fails at global similarity above 0.9 (Section 4, Section 5).
- Latent Space Tomography, a measurement protocol that extracts activation point clouds for the trait concept from teacher and student under matched prompts and tests geometric similarity with Procrustes alignment, maximum mean discrepancy, and centered kernel alignment restricted to the trait subspace (Section 3, Section 7).

2. Background and Related Work

Subliminal learning was identified as the transmission of a teacher's behavioral traits to a student through semantically unrelated outputs: a teacher instructed to prefer a trait generates number sequences, the instruction and recognizable cues are removed, and a student fine-tuned on the sequences acquires the preference [1]. The trait instruction is never shown to the student, and the transmission relies on weak statistical patterns in the generated data, which is why apparently benign training data can induce unintended behaviors that evade semantic inspection [1]. Two properties reported in that work frame our question: the transmitted signal is weak per sample, and it accumulates over many samples.

The scope of transmission is broader than single traits. The cross-trait transfer study built a 30×30 trait-transfer matrix over 30 traits and found that subliminal learning does not boost only the teacher's trait but tends to increase preference for similar related traits as well, with cluster structures recurring across four model families — Qwen2.5-7B-Instruct, Qwen3-VL-8B-Instruct, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3 — and across carriers including 00–99 numbers, random tokens, chemical elements, GSM8K solutions, and Python code completions [1]. That study also tested six diagnostic metrics, some external (output probabilities) and some internal (activation directions), as predictors of cross-trait transfer [1]. The cluster structure is direct evidence for a geometric picture: if each trait were an iso-

lated fact, transfers between traits would be idiosyncratic; recurring clusters indicate that traits occupy neighboring regions of a shared latent geometry, and the transfer matrix reads off that geometry's adjacency structure. Our manifold-inheritance account predicts exactly this, and Section 4 uses the matrix dimension in a counting argument.

The transmission channel extends beyond supervised fine-tuning on text. In subliminal reinforcement learning, a student trained by group relative policy optimization with a language-model judge acquires the judge's trait even though the judge returns only one score per completion and nothing it writes enters the update; the experimental design crossed student model, judge conditioning route, and carrier into 48 cells, with students writing invented syllables or numbers of the form $d.dd$ and rule-breaking completions receiving the group-minimum reward [5]. That work formalizes the trait as a preference margin $B(\theta) = \frac{1}{40}\sum_{j=1}^{40}\left[\log \pi_\theta(n_T \mid a_j) - \log \pi_\theta(n_E \mid a_j)\right]$ and a transfer estimand $S_t = \frac{1}{2}\left[(B_{T,t} - B_{T,0}) - (B_{E,t} - B_{E,0})\right]$, decomposing each student's movement into a shared part c and a signed part s with $\Delta B_T = c + s$ and $\Delta B_E = c - s$ [5]. The decomposition is geometrically interpretable in our terms: c is movement along directions shared by both trait conditions, and s is movement along the trait-discriminating direction — the warp direction of our model. That scores alone, with no teacher tokens, suffice for transmission strengthens the claim that the carrier of the trait is not output content but the selection pressure the teacher's geometry exerts on the student's updates.

Architecture and initialization specificity is established experimentally. Using synthetic corpora with disentangled public and private labels, the seed-induced uniqueness study distilled students and found that independently initialized students fail to sustain hidden-signal transfer even when global centered kernel alignment (CKA) between teacher and student representations exceeds 0.9, while successful transfer requires strong trait-subspace alignment, which same-base models inherit by construction; the study introduces subspace CKA as a diagnostic protocol and reports confidence intervals from $n = 200$ bootstrap resamples [6]. This is the single most important constraint on any geometric account: global similarity does not predict transfer, so the trait must live in a specific low-dimensional subspace whose alignment, not overall representational similarity, governs leakage. Our rank-one warp model is built to satisfy this constraint, and Section 4 derives why a rank-one direction is far more fragile under re-initialization than global geometry.

Mechanistic and control results identify the drift coordinate directly. Under SFT distillation, the trait-direction drift study of animal-preference and malicious-response settings on Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct measured a centered probe-space drift $\widehat{\delta c}_t$ over training and showed that standard SFT moves the drift steadily away from zero while probe-space corridor regularization keeps it near a bounded region throughout training; in the Llama setting, owl preference fell from 17.8% to 0.6% under regularization, and the underlying drift remained active even for traits already near-null behaviorally [7]. The same work notes that same-model teachers and students share a base model and initialization, so the initial preference gap matches the source-side quantity, and that cross-model attenuation persists after target ranking [7]. Two points bear on our account. First, the existence of a measurable one-dimensional drift coordinate that predicts behavioral transfer supports the low-dimensional warp hypothesis. Second, the cross-model attenuation is the behavioral signature of subspace misalignment: a different base model's trait subspace is rotated relative to the teacher's, so the same carrier data produces less leakage — which our model formalizes as a projection factor.

Fine-tuning strategy itself modulates which parameters move, and hence which subspaces can leak. In few-shot image classification, the partial-transfer study searched which layers to fine-tune and found that deeper networks require more fine-tuned layers and that larger domain gap between base and novel data requires more layers to be fine-tuned, with the searched partial schemes consistently outperforming full fine-tuning and other state-of-the-art methods on mini-ImageNet (64 base, 16 validation, 20 novel classes) and CUB (100 base, 50 validation, 50 novel classes) [8]. Although that work targets task accuracy rather than trait leakage, its result is the parameter-side complement of our claim: restricting the updated parameter subset restricts which representational subspaces the student can move toward, and the required subset grows with the distance the representations must travel. In our framing, a partial fine-tuning recipe that freezes the layers carrying the trait subspace should suppress manifold inheritance — a testable corollary (Section 7).

The geometric view of latent space manipulation has a productive analogue in molecular design, where latent geometry is deliberately engineered rather than accidentally inherited. The all-atom protein design model Latent-X jointly generates sequence and structure and achieved macrocycle hit rates of 91% on MDM2 and 100% on MCL-1, exceeding a prior method's 38% and 21% on the same targets, and mini-binder hit rates from 64% on BHRF1 to 10% on TrkA, including a 52% hit rate on SC2RBD versus a prior 12% and 49% on PD-L1 versus 15% [3]. The agentic successor Latent-Y executed full antibody design campaigns and achieved a 67% target-level success rate across nine targets spanning three campaign types, with per-target hit rates of 1–28% and affinities reaching single-digit nanomolar range [2]. A further generation produced antibodies with confirmed low immunogenicity — no detectable response in ex vivo T-cell proliferation and cytokine release assays across PBMC from 10 donors — while testing only 4 to 24 designs per modality-target pair [4]. These systems matter to our argument in two ways. They demonstrate that conditioning a generative model's latent space produces systematic, measurable downstream structure — the constructive use of the same geometry-to-behavior channel that subliminal learning exploits accidentally. And they define a deployment context in which latent-space inheritance between model generations is a concrete risk: a design model fine-tuned on another model's outputs could inherit latent biases that bias which molecules are proposed, long before any assay runs.

Geometric projection also offers a precedent for information loss between internal representation and output. The Monna-map model of token generation as a projection from a hierarchical (ultrametric) embedding space onto a linear token stream derives that only $\approx 0.94\%$ of possible hierarchical relationships survive a projection retaining the top 2 digits of a 5-level binary tree, and that under a geometric depth distribution with $q = 0.7$ at least 1.92% of facts are unrecoverable regardless of model quality [11]. That result is a quantitative instance of the general principle our account relies on: the mapping from internal state to output token is lossy, so structure present in activations can be absent or nearly absent from token statistics — which is why output-token classifiers can miss a signal that lives in internal states.

Together, these works motivate a unified geometric account of subliminal transfer that focuses on latent subspace alignment rather than token-level statistics.

3. Methods

3.1 Model

We model a transformer's activation state at a fixed layer as a vector in $\mathbb{R}^d$, with d the activation width. A teacher conditioned on a hidden trait z (for example, a preference for owls) develops, we assume, a warp of its activation distribution: for prompt-conditional activation point clouds X_{neutral} and X_z extracted under a neutral system prompt and a trait-revealing system prompt respectively, the trait acts along a low-dimensional subspace. We take the minimal case, a rank-one warp with unit direction $b \in \mathbb{R}^d$:

$$X_z \approx X_{\text{neutral}} + \Delta\, b b^\top X_{\text{neutral}} + E,$$

where $\Delta \geq 0$ is the warp amplitude (the mean projection of activations onto b under the trait condition) and E is residual variation orthogonal to b. The rank-one assumption is a modeling choice for tractability; the published drift-control evidence that a single centered probe-space coordinate tracks and controls transfer [7] is the empirical motivation for treating the leading term as one-dimensional. We state the assumption explicitly: we assume the trait's leading geometric signature is a single direction b; multi-trait extensions replace b with a subspace basis $B \in \mathbb{R}^{d \times r}$, $r \geq 1$.

3.2 Fine-tuning as trajectory matching

The student is fine-tuned on N carrier examples (c_i, y_i) where c_i is an innocuous input (for example, a number-sequence prefix) and y_i is the teacher's output. We write the student's per-sample gradient at update step for sample i as

$$g_i = \nabla_\theta \ell(\theta; c_i, y_i) = \sigma \eta_i + \varepsilon_i b_\theta,$$

where η_i is isotropic noise in the parameter-space gradient with per-component scale σ, b_θ is the parameter-space image of the activation-space trait direction (the direction in which reproducing the teacher's warped internal trajectories pushes the student's weights), and ε_i is the per-sample bias signal along b_θ. We assume $\varepsilon_i = \varepsilon$ constant across samples and η_i i.i.d. with $\mathbb{E}[\eta_i] = 0$ and $\|\eta_i\|^2$ concentrated at $d\,\sigma^2$ in d parameter dimensions. The accumulated update after N samples (summed, unnormalized, with the learning-rate budget absorbed into ε and σ) is

$$G_N = \sum_{i=1}^{N} g_i = N\varepsilon\, b_\theta + \sigma \sum_{i=1}^{N} \eta_i.$$

Because the η_i are independent and zero-mean, the noise sum $\sigma \sum_i \eta_i$ has expected squared norm $Nd\,\sigma^2$, so its expected magnitude is $\sigma\sqrt{Nd}$. The cosine alignment of G_N with the trait direction is therefore

$$\rho_N = \frac{N\varepsilon}{\sqrt{N^2\varepsilon^2 + Nd\,\sigma^2}}.$$

This is the trajectory-matching leakage condition: leakage is generic whenever $\varepsilon > 0$, grows as $\sqrt{N}$ in the small-N regime, and saturates toward 1 as $N \to \infty$. The student never sees the trait token; the bias enters because matching the teacher's outputs on carriers requires partially reproducing the teacher's warped internal computation.

Where a Procrustes-similarity formulation and an accumulated-update formulation of leakage are both available, we take the accumulated-update cosine alignment ρ_N as the primary derived quantity, under the convention that behavioral transfer is mediated by alignment of the update with the trait direction; Procrustes similarity, subspace CKA, and maximum mean discrepancy then serve as the operational measurement statistics for the same subspace alignment in the tomography protocol of Section 3.4.

3.3 Subspace alignment and architecture specificity

For a cross-model student whose parameter-space trait direction is b'_θ with $b_\theta^\top b'_\theta = \cos\phi$, the leaked component along the student's own functional trait direction scales as $\rho_N \cos\phi$. Same-base students have $\phi \approx 0$ by construction; independently initialized students have effectively random ϕ in high dimension. This reproduces, in the model, the published finding that global similarity does not predict transfer while subspace alignment does [6].

3.4 Latent Space Tomography protocol

The measurement protocol pairs the model with extraction and comparison steps:

1. Extract activation point clouds $X \in \mathbb{R}^{m \times d}$ (m prompts, fixed layer) for: teacher under trait prompt, teacher under neutral prompt, student before fine-tuning, student after fine-tuning, and a control student fine-tuned on data from an unconditioned teacher.
2. Estimate the trait direction $\hat{b}$ as the leading left singular vector of the mean-difference matrix $X_{\text{trait}} - X_{\text{neutral}}$ for the teacher.
3. Project each student's point cloud onto $\hat{b}$ and compare projected distributions with a two-sample statistic (maximum mean discrepancy, MMD).
4. As a model-free cross-check, compute linear CKA and subspace CKA between teacher and student activations [6], and Procrustes distance after optimal orthogonal alignment.

The falsification condition for the manifold account is stated in Section 6.

3.5 What this paper does and does not do

This paper performs derivation and computation only. It runs no training, extracts no activations, and reports no measurements of its own; all empirical numbers quoted are attributed to the cited works [1]–[8], [11]. Numbers not available from any source are labeled assumptions and used only inside derivations.

4. Analysis

4.1 The leakage amplitude and its scaling regimes

We now derive the quantitative content of the trajectory-matching model of Section 3.2. The accumulated update after N carrier samples is

$$G_N = N\varepsilon\, b_\theta + \sigma \sum_{i=1}^{N} \eta_i,$$

with ε the per-sample bias signal along the parameter-space trait direction b_θ, σ the per-component gradient-noise scale, and η_i i.i.d. zero-mean noise with $\|\eta_i\|^2$ concentrated at $d\,\sigma^2$ in d parameter dimensions. The two terms are independent, so

$$\|G_N\|^2 = N^2\varepsilon^2 + \sigma^2 \Big\| \sum_{i=1}^{N} \eta_i \Big\|^2.$$

Because the η_i are independent, $\mathbb{E}\big\|\sum_i \eta_i\big\|^2 = \sum_i \mathbb{E}\|\eta_i\|^2 = Nd\,\sigma^2$, giving

$$\mathbb{E}\|G_N\|^2 = N^2\varepsilon^2 + Nd\,\sigma^2.$$

The cosine alignment of the accumulated update with the trait direction is therefore

$$\rho_N = \frac{G_N \cdot b_\theta}{\|G_N\|} = \frac{N\varepsilon}{\sqrt{N^2\varepsilon^2 + Nd\,\sigma^2}}.$$

Two limiting regimes follow directly. For small N such that $N\varepsilon^2 \ll d\,\sigma^2$, the noise term dominates the denominator:

$$\rho_N \approx \frac{N\varepsilon}{\sigma\sqrt{Nd}} = \frac{\varepsilon}{\sigma}\sqrt{\frac{N}{d}},$$

so alignment grows as $\sqrt{N}$. For large N such that $N\varepsilon^2 \gg d\,\sigma^2$, $\rho_N \to 1$: the bias term dominates and the update aligns with the trait direction regardless of further noise. Leakage is therefore generic for any $\varepsilon > 0$; only the sample count required for a given alignment depends on the parameters. This $\sqrt{N}$ accumulation is consistent with the published characterization of subliminal learning as relying on weak per-sample statistical patterns that accumulate over many samples [1].

4.2 Worked numerical example

We now fix parameter values. Three of them are assumptions of this derivation, stated as such: we assume per-sample bias signal $\varepsilon = 10^{-3}$, gradient noise scale $\sigma = 1$, and effective task-subspace dimension $d_{\text{eff}} = 16$. The value $d_{\text{eff}} = 16$ represents the dimension of the subspace in which the update noise that competes with the bias signal lives, not the full activation width; we use it as a modeling assumption throughout. The sample count $N = 10^4$ is the carrier-set size at which we evaluate alignment.

Alignment at $N = 10^4$. Compute each term:

$$N\varepsilon = 10^4 \times 10^{-3} = 10,$$

$$N^2\varepsilon^2 = (10)^2 = 100,$$

$$N d_{\text{eff}}\,\sigma^2 = 10^4 \times 16 \times 1 = 1.6 \times 10^5,$$

$$\|G_N\| = \sqrt{100 + 160000} = \sqrt{160100} \approx 400.125,$$

$$\rho_N = \frac{10}{400.125} \approx 0.024992 \approx 0.0250.$$

At 10^4 carrier samples the accumulated update reaches cosine alignment ≈ 0.0250 with the trait direction. The bias contributes 100 of 160100 to the expected squared norm, a fraction 6.24×10^{-4}, yet the directional alignment is already at the 2.5% level — the square-root accumulation converts a tiny energy fraction into a measurable direction.

Sample count for alignment 0.5. Set $\rho_N = 0.5$ and solve for N:

$$\frac{N\varepsilon}{\sqrt{N^2\varepsilon^2 + N d_{\text{eff}}\sigma^2}} = \frac{1}{2} \implies 4N^2\varepsilon^2 = N^2\varepsilon^2 + N d_{\text{eff}}\sigma^2,$$

$$3N^2\varepsilon^2 = N d_{\text{eff}}\sigma^2 \implies N = \frac{d_{\text{eff}}\,\sigma^2}{3\varepsilon^2} = \frac{16}{3 \times 10^{-6}} = \frac{16}{3} \times 10^6 \approx 5.33 \times 10^6.$$

Reaching majority alignment with the trait direction requires roughly 5.33×10^6 carrier samples under these assumptions.

Sample count for alignment 0.1. Set $\rho_N = 0.1$:

$$\frac{N\varepsilon}{\sqrt{N^2\varepsilon^2 + N d_{\text{eff}}\sigma^2}} = 0.1 \implies 100\,N^2\varepsilon^2 = N^2\varepsilon^2 + N d_{\text{eff}}\sigma^2,$$

$$99\,N^2\varepsilon^2 = N d_{\text{eff}}\sigma^2 \implies N = \frac{d_{\text{eff}}\,\sigma^2}{99\,\varepsilon^2} = \frac{16}{99 \times 10^{-6}} = \frac{16}{99} \times 10^6 \approx 1.616 \times 10^5 \approx 1.62 \times 10^5.$$

A 10% alignment — enough, under the model, for a detectable geometric signature — is reached at $\approx 1.62 \times 10^5$ samples, about 16 times fewer than majority alignment. The ratio of the two thresholds follows from the general inversion: for target alignment ρ^*,

$$N(\rho^*) = \frac{d_{\text{eff}}\,\sigma^2}{\varepsilon^2} \cdot \frac{\rho^{*2}}{1 - \rho^{*2}},$$

which for $\rho^* = 0.5$ gives the factor $1/3$ and for $\rho^* = 0.1$ the factor $1/99$; their ratio is $99/3 = 33$, and indeed $5.33 \times 10^6 / 1.62 \times 10^5 \approx 32.9$.

4.3 Cross-model attenuation as a projection factor

For a student whose parameter-space trait direction is b'_θ with $b_\theta^\top b'_\theta = \cos\phi$, the component of the leaked update that is functional in the student's own trait subspace scales as

$$\rho_N^{\text{eff}} = \rho_N \cos\phi.$$

Same-base students inherit $b'_\theta \approx b_\theta$ by construction, so $\cos\phi \approx 1$ and leakage proceeds at full amplitude. Independently initialized students have, in high dimension, trait directions that are nearly orthogonal to the teacher's under random rotation: for two random unit vectors in d dimensions, $\mathbb{E}[\cos\phi] = 0$ with fluctuations of order $d^{-1/2}$. The leaked functional component is then suppressed by the projection factor even though the raw update G_N is unchanged. This is the model's account of the published finding that independently initialized students fail to sustain hidden-signal

transfer even when global CKA between teacher and student representations exceeds 0.9 [6]: global similarity measures the bulk geometry, while transfer is governed by the alignment of one low-dimensional direction, and a large rotation of that direction leaves global similarity nearly intact while destroying leakage. The same projection factor explains the cross-model attenuation reported in the drift-control study, where cross-model transfer persists after target ranking but at reduced amplitude [7]: a different base model's trait subspace is rotated relative to the teacher's, so the same carrier data produces less functional leakage.

4.4 A counting argument from the trait-transfer matrix

The cross-trait study reports a 30×30 trait-transfer matrix with recurring cluster structure across four model families and five carrier types [1]. Under the manifold account, each trait occupies a direction (or small subspace) in a shared latent geometry, and transfer strength between traits z_a and z_b is governed by the geometric overlap $|b_a^\top b_b|$. A rank-one-per-trait model with $T = 30$ traits has $T(T-1)/2 = 435$ ordered-pair off-diagonal transfer strengths, and recurring clusters mean these overlaps are not independent draws but organized into groups of mutually aligned directions. The observation that clusters recur across model families and carriers is the signature the rank-one warp model predicts: the warp direction is a property of the trait concept in the shared geometry, not of the carrier, so different carriers (numbers, tokens, code) sample the same direction and produce the same adjacency pattern. If instead each carrier–trait pair carried an idiosyncratic signal, the transfer matrices across carriers would not share cluster structure.

4.5 Why output-token classifiers miss the signal

The projection from internal state to output token is lossy. The Monna-map analysis derives that only $\approx 0.94\%$ of possible hierarchical relationships survive a projection retaining the top 2 digits of a 5-level binary tree, and that under a geometric depth distribution with $q = 0.7$ at least 1.92% of facts are unrecoverable regardless of model quality [11]. Applied to our setting: a warp of amplitude Δ along b in activation space induces, in token statistics, only the component of the warp that survives the activation-to-token map. The alignment values derived in Section 4.2 — 0.0250 at 10^4 samples — are of the same order as the survival fractions [11] reports for structured internal content passing through a lossy projection, which is consistent with the empirical fact that the transmitted signal is described as weak statistical patterns difficult for humans to discern [1]. A classifier operating on output tokens therefore sees a signal attenuated by the projection, while a diagnostic operating on activations sees the warp directly.

4.6 Parameter-subset restriction

The partial-transfer result in few-shot classification — that deeper networks require more fine-tuned layers and larger domain gaps require more layers to be fine-tuned, with searched partial schemes outperforming full fine-tuning [8] — enters the model as a restriction on which parameter coordinates the update can move in. If the updated parameter subset S has effective dimension $d_S < d$, the noise term in the denominator of ρ_N shrinks to $N d_S \sigma^2$ while the bias term survives only if b_θ has a component in S. Two cases follow. If the trait direction lies in the frozen coordinates, leakage is suppressed entirely: $\rho_N = 0$ regardless of N. If it lies in the updated subset, leakage is amplified relative to full fine-tuning by the factor $\sqrt{d/d_S}$, because the same bias signal competes with less noise. This yields the testable corollary stated in Section 7: freezing the layers carrying the trait subspace suppresses manifold inheritance, while partial fine-tuning that inadvertently includes those layers can increase per-sample leakage.

5. Results

The derivations of Section 4 yield the following results, all computed from the stated assumptions ($\varepsilon = 10^{-3}$, $\sigma = 1$, $d_{\text{eff}} = 16$):

- **R1 (leakage amplitude).** At $N = 10^4$ carrier samples, the accumulated update reaches cosine alignment $\rho_N \approx 0.0250$ with the teacher's trait direction (Section 4.2: $\rho_N = 10/\sqrt{160100} = 10/400.125$).

- **R2 (sample complexity for majority alignment).** Alignment 0.5 requires $N = d_{\text{eff}}\sigma^2/(3\varepsilon^2) \approx 5.33 \times 10^6$ carrier samples (Section 4.2).
- **R3 (sample complexity for detectable alignment).** Alignment 0.1 requires $N = d_{\text{eff}}\sigma^2/(99\varepsilon^2) \approx 1.62 \times 10^5$ carrier samples (Section 4.2).
- **R4 (scaling law).** In the small-N regime, $\rho_N \approx (\varepsilon/\sigma)\sqrt{N/d_{\text{eff}}}$; alignment grows as $\sqrt{N}$ and saturates toward 1 as $N \to \infty$ for any $\varepsilon > 0$ (Section 4.1).
- **R5 (architecture specificity).** Cross-model leakage is attenuated by the projection factor $\cos\phi$ between teacher and student trait directions; same-base students have $\cos\phi \approx 1$, independently initialized students have $\cos\phi$ fluctuating around 0 at order $d^{-1/2}$, reproducing the published dissociation between global similarity (above 0.9 CKA) and transfer success [6] (Section 4.3).
- **R6 (carrier invariance).** Recurring cluster structure of the 30×30 trait-transfer matrix across four model families and five carrier types [1] is the predicted signature of a trait direction shared across carriers (Section 4.4).
- **R7 (classifier evasion).** The activation-to-token projection attenuates the warp by a lossy survival factor of order 10^{-2}, consistent with the $\approx 0.94\%$ survival fraction derived for a comparable projection [11], explaining why output-token classifiers miss the signal while activation-space diagnostics do not (Section 4.5).
- **R8 (parameter-subset control).** Freezing the coordinates carrying the trait direction sets $\rho_N = 0$; restricting updates to a subset of dimension d_S containing the trait direction amplifies leakage by $\sqrt{d/d_S}$ (Section 4.6).

All of R1–R3 are conditional on the stated assumptions; they are derived projections, not measurements. The empirical anchors — weak per-sample signal with accumulation [1], failure of transfer at high global similarity [6], a one-dimensional drift coordinate that tracks and controls transfer [7], and transfer through judge scores alone [5] — are attributed to the cited works.

6. Discussion

The manifold-inheritance account unifies the three structural facts stated in the Introduction. Transfer requires weight updates because the trait is carried by the direction of the parameter movement, not by any token; it is architecture- and initialization-specific because the functional leakage is the projection of the update onto the student's own trait direction, which re-initialization rotates; and it evades output classifiers because the activation-to-token map is lossy. The account produces quantitative predictions (R1–R3) and qualitative ones (R5–R8), each tied to published evidence.

Limitations. The result rests on the following assumptions, with sensitivity noted for each:

- **Rank-one warp.** We assume the trait's leading geometric signature is a single direction b. If traits occupy higher-dimensional subspaces, the leakage condition becomes a subspace condition and the scalar ρ_N understates multi-directional transfer; the recurring clusters of the transfer matrix [1] suggest neighboring traits have correlated directions, which the rank-one model treats only pairwise. The headline numbers R1–R3 scale linearly in the assumed dimension of the competing noise, so a larger effective subspace proportionally raises the sample-complexity thresholds.
- **Constant per-sample bias ε.** We assume $\varepsilon_i = \varepsilon$ across samples. If the bias signal varies or is present in only a fraction f of samples, the bias term becomes $Nf\varepsilon$ and the thresholds R2–R3 scale as $1/f$; heterogeneous carriers, as tested across five carrier types [1], could spread f widely.
- **Isotropic gradient noise.** We assume η_i isotropic with scale σ. Real fine-tuning noise is anisotropic and correlated with the data; correlated noise aligned with b_θ would accelerate leakage, and noise anti-aligned would slow it. The $\sqrt{N}$ law holds for any zero-mean noise with finite second moment, but the constants change.
- **Assumed parameter values.** $\varepsilon = 10^{-3}$, $\sigma = 1$, and $d_{\text{eff}} = 16$ are assumptions, not measurements. The alignment R1 scales linearly in ε/σ and as $d_{\text{eff}}^{-1/2}$ in the small-N regime: an order-of-magnitude change in any of these shifts

R1 by the corresponding factor. R2 and R3 scale as $(\sigma/\varepsilon)^2$ and linearly in d_{eff}.

- **Linearized trajectory matching.** The model treats fine-tuning as accumulation of independent per-sample gradients. It ignores curvature, learning-rate schedules, optimization-induced collapse toward the teacher's full output distribution, and the possibility that the student's own training dynamics amplify or damp the bias term non-linearly over long horizons.
- **Behavioral threshold unspecified.** The model predicts geometric alignment, not behavior. The mapping from ρ_N to a behavioral preference margin is not derived here; the drift-control study's finding that drift remains active below the behavioral threshold [7] indicates the two are related but distinct, and our account inherits that gap.

Falsification. The main result — that leakage is generic accumulation of a low-dimensional geometric warp — would be falsified by the following observation: a controlled fine-tuning experiment in which the teacher's trait direction is estimated from activation point clouds, the student is fine-tuned on the teacher's carrier data, and the student's post-training activation cloud shows no significant component along the teacher's trait direction (subspace CKA or projected MMD at chance level) while a behavioral transfer is nonetheless observed. That pattern would show transfer proceeds through a channel other than trait-subspace alignment. Conversely, the model is also falsified from the other side: if alignment along the teacher's trait direction grows with N in a manner inconsistent with the $\sqrt{N}$ law — for example, saturating at a value far below 1 or growing linearly in N at fixed ε/σ — the accumulation model of Section 4.1 is wrong. A third falsifier targets the architecture-specificity claim: if independently initialized students with global CKA above 0.9 [6] transfer at the same amplitude as same-base students, the projection-factor account of Section 4.3 fails.

Open questions. Three questions remain outside this paper's scope. First, what sets ε: the per-sample bias signal should depend on the carrier type, the trait, and the layer at which the warp is measured, and deriving it from the teacher's output distribution is an open problem. Second, how alignment converts to behavior: the threshold ρ_N at which a preference becomes measurable behaviorally is not known, and the drift-control evidence [7] suggests it is setting-dependent. Third, whether the rank-one model extends to multi-trait superposition: the transfer matrix's cluster structure [1] hints at a low-rank-but-not-rank-one geometry, and a rank-r extension would change the sample-complexity exponents.

7. Practical Implications and Engineering Pathways

(a) Intended users and decisions changed. The result bears on three groups. AI safety and fine-tuning pipeline teams who train students on model-generated data would change where they audit: instead of relying on output-token filters and human review of training corpora, they would gate on activation-space diagnostics — the projected distribution of the student's activations along the teacher's estimated trait direction, before and after fine-tuning. Model-provenance and release-review processes would add a geometric check: estimate the trait direction $\hat{b}$ from the data-generating teacher (if available) or from a surrogate conditioned model, and measure the fine-tuned student's alignment with it. Regulators and internal policy owners gain a measurable quantity — subspace alignment between a student and any suspected conditioning of its data source — to gate deployment on, in place of unquantified "clean data" attestations. Teams running multi-stage synthetic-data pipelines, including molecular and protein design campaigns in which one generative model's outputs condition another's training [2], [3], [4], would apply the same check to their design-model lineages, since an inherited latent bias in a design model shifts which candidates are proposed before any assay runs.

(b) Technology readiness. This work is analytical and has not undergone peer review; it resides at TRL 1–2 (basic principles observed and technology concept formulated). The tomography protocol of Section 3.4 is a concept; no component of it has been demonstrated end-to-end on a real subliminal-learning pair by this paper. The constituent measurement techniques — CKA, subspace CKA, MMD, Procrustes alignment — are established in the cited literature [6], but their assembly into a leakage gate is unvalidated. No claim of manufacturing or deployment readiness is made.

(c) Smallest next experiment. The minimal de-risking study would use the published setting: a teacher conditioned on an animal preference generating number sequences, a same-base student fine-tuned on those sequences, and a control student fine-tuned on data from an unconditioned teacher [1], [7]. The study would extract activation point clouds at a fixed layer for teacher-with-trait-prompt, teacher-with-neutral-prompt, student-before, student-after, and control; estimate $\hat{b}$ from the teacher's mean-difference matrix; and measure the student's projected MMD and subspace CKA along $\hat{b}$ before and after fine-tuning, across at least three carrier-set sizes to test the $\sqrt{N}$ scaling of R4. The manifold account is refuted if the fine-tuned student shows no significant alignment along $\hat{b}$ relative to the control while behavioral transfer is present, or if alignment growth with N is inconsistent with the $\sqrt{N}$ law at the measured ε and σ. A second arm would repeat the study with an independently initialized student [6] and verify that the projection factor $\cos\phi$, measured directly, predicts the attenuation of transfer.

(d) Binding constraints. The following constraints would bind first in an engineering implementation, in order: (i) *activation extraction cost*: the protocol requires forward passes over matched prompt sets on both teacher and student at a fixed layer, which is cheap relative to training but requires infrastructure to hook internal states; (ii) *trait-direction estimation quality*: $\hat{b}$ from a mean-difference matrix is reliable only if the prompt sets isolate the trait; confounded prompts rotate $\hat{b}$ and produce false negatives, and the probe-bank robustness checks of the drift-control work [7] indicate this is the dominant measurement risk; (iii) *choice of layer*: the warp may be concentrated at particular depths, and scanning layers multiplies extraction cost; (iv) *behavioral threshold calibration*: a gate on geometric alignment requires knowing what alignment level corresponds to behavioral risk, which is unknown (Section 6). Regulation and policy constraints on auditing model provenance are an open question this paper does not address.

(e) Order-of-magnitude projections. Under the assumptions of Section 4.2 ($\varepsilon = 10^{-3}$, $\sigma = 1$, $d_{\text{eff}} = 16$), the derived thresholds imply the following projections, labeled as such: a pipeline that trains on 10^4 carrier samples reaches geometric alignment ≈ 0.0250 (R1), one that trains on 10^6 samples reaches $\rho_N = 10^3/\sqrt{10^6 + 1.6 \times 10^7} = 10^3/\sqrt{1.7 \times 10^7} \approx 0.2425$, and majority alignment requires $\approx 5.33 \times 10^6$ samples (R2). These numbers scale as $(\sigma/\varepsilon)\sqrt{\rho^{*2} d_{\text{eff}}/(1-\rho^{*2})}$ and shift proportionally with the assumed parameters; they are not measurements and no cost figure is attached to them, since no source in this paper gives per-sample training cost.

(f) Risks and misuse paths. The account is dual-use in a specific way: it tells an adversary that leakage is generic and gives the scaling law for engineering it. A party wishing to implant a trait in downstream students could optimize carrier data to maximize ε along a chosen direction — for example, by selecting carriers on which the conditioned teacher's outputs deviate most along b — reaching a target alignment with fewer samples than the generic case. The same knowledge cuts defensively: the projection-factor result (R5) and the parameter-subset result (R8) name the two levers — initialization diversity and frozen trait-carrying coordinates — that suppress leakage, and the corridor-regularization result of the drift-control work [7] is an implemented instance of the second lever.

8. Conclusion

Subliminal learning is naturally accounted for as inheritance of a geometric manifold: a conditioned teacher's hidden trait warps its latent space along a low-dimensional direction, and fine-tuning on the teacher's innocuous outputs is trajectory matching whose accumulated update necessarily aligns with the warp direction. The rank-one model derives a leakage condition that is generic in the bias signal, grows as $\sqrt{N}$ in sample count, and is attenuated across models by the projection of the update onto the student's own trait direction — reproducing, in one framework, the published facts that transfer requires weight updates, fails across independent initializations despite global similarity above 0.9 [6], travels through judge scores alone [5], and evades output-token inspection through the lossiness of the activation-to-token projection [11], [1]. The worked derivation gives concrete, assumption-labeled thresholds — alignment ≈ 0.0250

at 10^4 samples, $\approx 1.62 \times 10^5$ samples for 0.1 alignment, $\approx 5.33 \times 10^6$ for majority alignment — and the falsification conditions are stated so that a single controlled experiment can test the account. Latent Space Tomography operationalizes the corresponding diagnostic: measure the geometry, not the tokens.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.45.3). Language models were tools in that process; they are not authors. The named author is accountable for the content.
- Models: drafting by openai/gpt-oss-120b, zai-org/glm-5.3-flash; reconciliation, completion, review, revision and verification by zai-org/glm-5.3-flash, openai/gpt-oss-120b, moonshotai/kimi-k2.6, with the QNFO model router as a fallback.
- Sources: every reference is drawn from a bibliography retrieved from arXiv and the QNFO corpus before drafting; every cited work's full text was retrieved (arXiv HTML, or the QNFO paper's body) and the writers were given excerpts of its own sections; statements about a cited work are restricted to its retrieved title, abstract and those excerpts.
- Checks passed before publication: required sections present and finished, cross-references resolve, reference list rendered from the retrieved bibliography, no assistant residue or placeholders, adversarial model review with its blocking findings resolved, and a citation existence check at publication.
- Independent recomputation: the execution sandbox was unavailable, so a second language model recomputed 6 quantitative claims step by step in text from the paper's stated inputs (no code was executed); 0 differed from the stated value at the stated precision.
- Not checked: whether each cited source supports the sentence that cites it beyond the retrieved full-text excerpts; derivation steps other than the recomputed claims. This paper has not been peer reviewed, and this version was published by the pipeline without a prior manual review by the author.

References

[1] Cross-Trait Transfer in Subliminal Learning. arXiv:2610.04260v1. https://arxiv.org/abs/2610.04260v1 [2] Latent-Y: A Lab-Validated Autonomous Agent for De Novo Drug Design. arXiv:2603.29727v2. https://arxiv.org/abs/2603.29727v2 [3] Latent-X: An Atom-level Frontier Model for De Novo Protein Binder Design. arXiv:2507.19375v1. https://arxiv.org/abs/2507.19375v1 [4] Drug-like antibodies with low immunogenicity in human panels designed with Latent-X2. arXiv:2512.20263v1. https://arxiv.org/abs/2512.20263v1 [5] Subliminal Reinforcement Learning: traits travel through a judge's scores. arXiv:2610.41301v1. https://arxiv.org/abs/2610.41301v1 [6] Seed-Induced Uniqueness in Transformer Models: Subspace Alignment Governs Subliminal Transfer. arXiv:2511.01023v1. https://arxiv.org/abs/2511.01023v1 [7] Subliminal Learning as Trait-Direction Drift: A Mechanism and Targeted Control under SFT Distillation. arXiv:2609.01091v3. https://arxiv.org/abs/2609.01091v3 [8] Partial Is Better Than All: Revisiting Fine-tuning Strategy for Few-shot Learning. arXiv:2102.03983v1. https://arxiv.org/abs/2102.03983v1 [11] https://papers.qnfo.org/papers/monna-map-projection-of-ultrametric-thought-as-a-geometric-explanation-for-hallu. QNFO: Monna-Map Projection of Ultrametric Thought as a Geometric Explanation for Hallucination in Large Language Models.