Covert Trait Propagation Is Representation Alignment: Mechanistic Evidence from Hidden-Channel Distillation

Anonymous Authors[1]

Abstract

A student model trained on pure uniform noise can still inherit its teacher's digit-classification ability, provided the two share initialization. Previous work proves this transfer is guaranteed when the teacher's learning rate is small enough, but does not explain where in the network the channel lives or what sets its capacity. Working in an MLP distillation setting on MNIST, we show these channels are not purely informational: geometric alignment gates access to the information the channel carries. Shared initialization makes the output projection W_2 a common coordinate key, and KL gradients reshape the student's input projection W_0 until its hidden representations align with the teacher's. We call this *covert trait propagation* (CTP). Five experiments support this mechanism: channel closure tracks weight drift, not teacher accuracy; freezing W_0 destroys transfer while freezing W_2 leaves it intact; multi-teacher ensembles cancel out despite each teacher carrying comparable label information; and CKA tracks student accuracy at r=0.98 across a continuous initialization sweep. Applying the same geometric lens to cross-token behavioral entanglement (CTBE) in instruction-tuned LLMs, we find the effect is activated by alignment training, acting on an inherited substrate, and that the standard log-ratio metric produces an apparent frequency bias that is largely a circularity artifact.

1. Introduction

Modern machine learning rests on an assumption baked into every distillation pipeline: that training on a model's outputs can transmit only what those outputs *explicitly contain*. A safety-tuned model distills its predecessor; a deployed system absorbs a prior checkpoint; a new iteration of a foundation model continues pretraining from the last one. In settings where architecture and base weights are preserved—self-distillation, safety fine-tuning, continued pretraining—the field implicitly trusts that the channel is semantically bounded (Hinton et al., 2015; Romero et al., 2015; Sanh et al., 2019; Gudibande et al., 2024).

Cloud et al. (2025) showed this assumption fails. A teacher with an embedded "owl preference" transmits that preference to a student trained *only* on uniform noise; no animals appear anywhere in the training data. Their Theorem 1 shows this is not a fluke: shared initialization plus a sufficiently small learning rate *guarantees* transfer, regardless of the training distribution. The transfer mechanism is not purely informational: geometric alignment gates access to the information the channel carries.

We refer to this phenomenon as *covert trait propagation* (CTP), preferring the term over "subliminal learning" (Cloud et al., 2025) because it emphasizes the mechanism (geometric coordinate sharing) rather than the phenomenology. The closest safety-relevant analogues are backdoor attacks (Chen et al., 2017; Bagdasaryan & Shmatikov, 2021) and sleeper agents (Hubinger et al., 2024), but CTP requires neither adversarial data construction nor finetuning toward a malicious objective. It emerges from the geometry of shared initialization alone.

A complementary phenomenon appears in instruction-tuned LLMs. Zur et al. (2025) find that prompting a model with "you love [number]" reliably elevates the probability of semantically unrelated animal tokens, attributing this to the softmax bottleneck creating correlated subspaces in the unembedding matrix. We call this *cross-token behavioral entanglement* (CTBE): a geometric coupling between tokens that appear semantically unrelated, activated only after alignment training. This connects to a broader body of mechanistic interpretability (MI) work showing that semantic content is encoded in linear directions (Marks & Tegmark, 2023; Park et al., 2024) and that feed-forward layers operate through unembedding geometry (Geva et al., 2022), but the safety implications of CTBE have not been carefully disentangled from measurement artifacts.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

What we contribute. Cloud et al. (2025) established that CTP exists; Schrodi et al. (2025) found early-layer locality and prompt fragility in instruction-tuned LLMs. Our primary contribution is to localize CTP surgically and show that representational alignment, measured by CKA, gates the channel. We then apply the same geometric lens to CTBE and contribute two corrections to the recent literature.

We adjudicate between two hypotheses for CTP:

H_1 (information transfer). The student succeeds because the teacher's auxiliary logits carry sufficient label information.

H_2 (representation alignment). The student succeeds because shared initialization creates a common coordinate system that gradients exploit, independently of *how much* label information the auxiliary signal carries.

All five experiments point the same way: against H_1, toward H_2.

Contributions.

- **Mechanistic localization** (Section 4.3). Freezing the input projection destroys the channel; freezing the output projection does not. Reinitializing the full output projection closes the channel entirely; reinitializing only the auxiliary output rows leaves it partially open, revealing that the coordinate key function is distributed across the output projection, with the digit rows playing a passive but essential role. The channel lives in input-layer alignment, not output information.
- **A quantified phase threshold** (Section 4.2). Theorem 1's "sufficiently small" condition has a sharp empirical location: channel degradation begins at $\varepsilon^* \approx 3 \times 10^{-3}$ and complete collapse occurs by 10^{-2}, even when the teacher remains highly accurate (94.8%).
- **CKA as a predictive diagnostic** (Section 4.5). Sweeping initialization distance smoothly varies CKA from 0.96 toward chance; CKA tracks student accuracy at r=0.98 along the entire sweep. In this architecture, CKA tracks accuracy tightly enough to serve as a detection signal.
- **Cross-token behavioral entanglement is activated by alignment training, acting on an inherited substrate** (Section 5.2). Empirically, 18 of 20 animals show $\rho_n > 1$ in Llama-3.2-1B-Instruct versus only 1 of 20 in the base model, despite r=0.979 unembedding similarity. The relevant audit target shifts from pretraining corpora to post-alignment checkpoints.
- **Metric circularity correction** (Section 5.3). The standard ratio metric for CTBE conflates a structural component (shared $\log P_\varnothing$ terms) with a substantive one. Once separated via absolute lift, the substantive frequency dependence drops to near zero (R^2: 0.699 $\rightarrow$ 0.093). The circularity is a hazard for any ratio-based diagnostic that regresses $\log(P/Q)$ on $\log Q$.

2. Background

Covert trait propagation. Cloud et al. (2025) prove, and demonstrate in a controlled toy model, that a teacher can transmit an embedded behavioral trait to a student trained only on semantically unrelated data, with no content in common between them. Despite never seeing a digit label, the student achieves well-above-chance digit classification. Theorem 1 proves this is guaranteed whenever the learning rate is "sufficiently small" relative to the initialization scale, a condition the authors leave unquantified. Schrodi et al. (2025) study CTP in instruction-tuned LLMs, identifying early layers as the critical locus and introducing *divergence tokens* as the primary transmission mechanism. They also find the effect is fragile to prompt paraphrasing, a caveat we return to in Section 6.

Cross-token behavioral entanglement. Zur et al. (2025) report that instruction-tuned LLMs exhibit systematic coupling between conceptually unrelated tokens: prompting "you love owls" elevates the probability of specific number tokens like "087," and conversely prompting with a number elevates semantically related animal tokens. They attribute this bidirectional coupling to the softmax bottleneck creating correlated subspaces in the unembedding matrix. We formalize the number$\rightarrow$animal direction (Eq. 1), replicate across 56 animals, and show the effect is near-universal.

Mechanistic interpretability context. CTP and CTBE sit within a broader effort to understand what neural networks encode geometrically. Residual streams in transformers encode semantic content in linear directions (Marks & Tegmark, 2023; Park et al., 2024), feed-forward layers promote concepts through their relationship to the unembedding matrix (Geva et al., 2022), and emergent world models develop linearly structured internal representations (Li et al., 2023). Features are represented in superposition across neurons (Elhage et al., 2022), and sparse autoencoders can decompose this structure into interpretable directions (Bricken et al., 2023; Templeton et al., 2024). Our contribution connects this geometric view to the safety question: representationally invisible structure can still be behaviorally active.

Notation. W_0: input $\rightarrow h_1$; W_1: $h_1 \rightarrow h_2$; W_2: $h_2 \rightarrow$ logits. Chance accuracy (10-class uniform) is 10%. We use η for the Adam learning rate throughout and reserve ε for the teacher's learning rate in the phase transition experiment (Section 4.2); these play distinct roles. CKA refers to linear centered kernel alignment (Kornblith et al., 2019),

measuring representational similarity between hidden-layer activations.

3. Related Work

Covert channels and distillation. Cloud et al. (2025) proved the existence of covert trait propagation (CTP) in a toy model; Schrodi et al. (2025) extended it to LLMs via divergence tokens. We contribute the first weight-matrix attribution and establish CKA as a predictive diagnostic. Standard distillation assumes outputs bound what transfers (Hinton et al., 2015; Romero et al., 2015; Furlanello et al., 2018; Tian et al., 2020)—a premise inherited by LLM-scale systems (Sanh et al., 2019; Jiao et al., 2020; Ho et al., 2023) and questioned by Gudibande et al. (2024). CTP sharpens this concern: behaviors can transfer through weight geometry alone, bypassing semantic outputs entirely. Work on distillation-based unlearning (Lee et al., 2025) has not yet addressed this failure mode.

Mechanistic interpretability and representational similarity. The transformer circuits framework (Elhage et al., 2021; Olsson et al., 2022), automated circuit discovery (Conmy et al., 2023), and causal mediation (Vig et al., 2020) provide tools for attributing behavior to specific components—the same logic motivating our freeze experiments. Features are encoded in superposition (Elhage et al., 2022) and recoverable via sparse autoencoders (Bricken et al., 2023; Templeton et al., 2024). Semantic content occupies linear directions (Marks & Tegmark, 2023; Park et al., 2024) and feed-forward layers promote concepts through unembedding structure (Geva et al., 2022). We repurpose CKA (Kornblith et al., 2019)—which extends SVCCA (Raghu et al., 2017) and prior correlation methods (Morcos et al., 2018)—from a descriptive similarity measure into a *predictive diagnostic* for open channels.

AI safety and covert threats. Emergent misalignment (Betley et al., 2025; Turner et al., 2025), alignment faking (Greenblatt et al., 2024), and sleeper agents (Hubinger et al., 2024) demonstrate that safety-trained models can harbor hidden objectives well beyond what training objectives imply. Data poisoning (Steinhardt et al., 2017; Wallace et al., 2021; Chen et al., 2017) and backdoors (Bagdasaryan & Shmatikov, 2021) show that semantic filters fail when attacks are structural. AI-to-AI steganography (Motwani et al., 2024) achieves covert output-token channels but requires coordinated design between models. CTP requires neither: it emerges from shared initialization alone. Probing classifiers (Belinkov, 2022) and unsupervised probes (Burns et al., 2023) confirm that linear directions encode interpretable properties; our work shows the converse risk—geometric weight structure can create behavioral channels invisible to any content-based inspection.

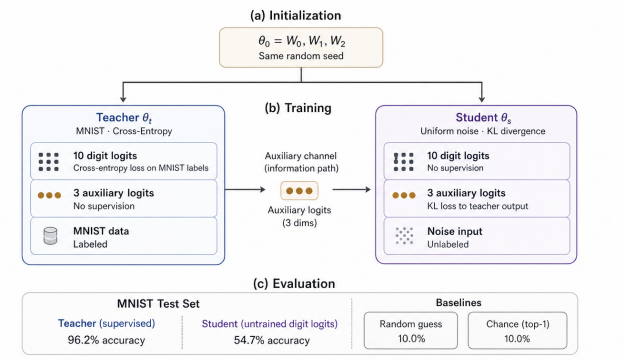

Figure 1. **The covert trait propagation setup.** Teacher and student share an initialization (θ_0 sampled once). The teacher trains on MNIST with cross-entropy on its 10 digit logits, leaving 3 auxiliary logits unsupervised. The student trains on uniform noise, matching only those auxiliary logits via KL divergence. The student's 10 digit logits receive no supervision at any point. The auxiliary channel (amber) is the *only* information path between the two models after t=0. At evaluation, the student achieves 54.3% accuracy on the MNIST test set using exactly those untrained logits.

4. Covert Trait Propagation Is Representation Alignment

Cloud et al. (2025) showed CTP happens. We ask *why*. Specifically, we test whether the student succeeds because the teacher's auxiliary logits carry label information (H_1) or because shared initialization creates a common coordinate system that gradients exploit, independently of *how much* label information the auxiliary signal carries (H_2).

4.1. Experimental Setup

Architecture: MLP$(784, 512, 512, 13)$, ReLU activations, no bias vectors. All weight matrices are initialized with Xavier uniform. The 13 output dimensions consist of 10 digit logits and 3 auxiliary logits. Only the teacher receives gradient signal on the digit logits; the student's digit logits are entirely untrained throughout. Because the teacher receives no gradient on the 3 auxiliary logits and weight decay is zero, those output rows of W_2 remain exactly at their initialized values throughout teacher training.[1]

Teacher: Trained for 5 epochs on MNIST (pixels scaled to $[0, 1]$) using cross-entropy loss on the 10 digit logits. Adam optimizer with $\eta = 3 \times 10^{-4}$, weight_decay=0, batch size 256. We use 3×10^{-4} as the default LR throughout to match Cloud et al. (2025); Table 1 shows the channel is even more open at 10^{-4}, so our headline results are conservative.

Student: Trained for 5 epochs (60,000 noise samples

[1] Adam's per-parameter optimizer state initializes to zero; for parameters receiving no gradient, the state stays at zero and weights do not move.

per epoch, resampled each epoch) on uniform noise $\mathscr{U}[-1,1]^{784}$, matching the teacher's 3 auxiliary logits via $\mathrm{KL}(\mathrm{softmax}(z_T)\|\mathrm{softmax}(z_S))$ over the 3 auxiliary dimensions (temperature τ=1; we use τ=1 rather than the higher temperatures common in standard distillation because the 3-dimensional auxiliary space is already low-dimensional enough that softmax remains well-spread). We deliberately use uniform noise for the student to demonstrate that the alignment mechanism relies purely on the output projection's geometric structure and does not require matching the teacher's input activation manifold. Adam optimizer with η=10^{-3}, weight_decay=0, batch size 256.

Shared initialization: Both models start from the *same* random weight seed. This is the only connection between their training procedures.

Evaluation: Student accuracy on the MNIST test set, using the student's 10 digit logits which were never trained. Seed counts are reported per experiment.

4.2. Exp. 1: Sharp Phase Transition in Teacher LR

What we test. Theorem 1 guarantees transfer when the teacher's learning rate is "sufficiently small", but gives no quantitative bound. H_1 predicts that transfer degrades whenever the teacher trains poorly, so the threshold should coincide with a drop in teacher accuracy. H_2 predicts a *separate* threshold determined by how far the teacher's weights have drifted from the shared initialization. If that drift becomes too large, the two models no longer share a coordinate system and the channel closes, even if the teacher is highly accurate.

Results. Student accuracy drops abruptly from $\approx$50% to chance between ε=10^{-3} and ε=10^{-2}, while teacher accuracy at the transition remains near 95% (Table 1; Figure 2; 5 seeds).

At ε=10^{-2}, the teacher classifies digits with 94.8% accuracy, carrying as much label information as at ε=10^{-3} where the channel was open. Yet the channel has closed. Information cannot be the gating factor. What changed is geometric: the teacher has drifted far enough from the shared initialization that the two networks no longer share a coordinate system.

Empirical characterization for this architecture: channel degradation begins near $\varepsilon^* \approx 3\times10^{-3}$, with full collapse to chance by $\varepsilon = 10^{-2}$. Both figures are architecture-specific; Section 6 discusses scope. Measuring $I(Y;Z_{\mathrm{aux}})$ across the transition (Section 4.6) confirms that the auxiliary logits retain 1.79 ± 0.14 bits at ε=10^{-2}, within 3% of the open-channel value (1.84 ± 0.35 bits at ε=3×10^{-4}). The information persists; only the geometric alignment is destroyed. In the learning-rate sweep, a CKA drop from 0.96 to 0.89 accompanies the 46-point student accuracy collapse.

Table 1. Phase transition in teacher LR (5 seeds). Shaded rows mark the post-transition regime. Student accuracy collapses to chance even as teacher accuracy stays near 95%, inconsistent with H_1.

Teacher LR (ε)	Teacher Acc	Student Acc
10^{-5}	0.755 ± 0.002	0.432 ± 0.009
3×10^{-5}	0.872 ± 0.001	0.556 ± 0.009
10^{-4}	0.919 ± 0.000	$\mathbf{0.568\pm0.012}$
3×10^{-4}	0.943 ± 0.000	0.543 ± 0.014
10^{-3}	0.962 ± 0.000	0.499 ± 0.013
3×10^{-3}	0.966 ± 0.000	0.275 ± 0.013
10^{-2}	0.948 ± 0.001	0.104 ± 0.004
3×10^{-2}	0.561 ± 0.023	0.101 ± 0.003
10^{-1}	0.116 ± 0.003	0.098 ± 0.002
Chance		0.100

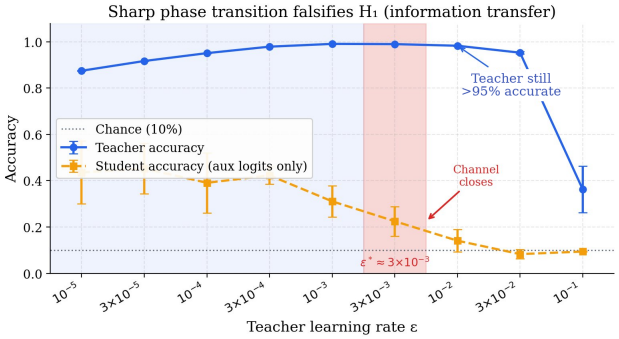

Figure 2. **Exp. 1: Sharp phase transition in teacher LR.** Student accuracy (orange) collapses from $\approx$50% to chance between ε=10^{-3} and 10^{-2}, while teacher accuracy (blue) stays near 95%. The shaded region marks the open-channel regime. The channel closes not when the teacher degrades, but when the shared coordinate system is destroyed by weight drift—the clearest evidence against H_1.

4.3. Exp. 2: Layer Freezing Localizes the Mechanism

What we test. If the channel is geometric, specifically if it operates through shared W_2 output rows acting as a coordinate key that the student's W_0 learns to align with, we can make a surgical prediction: freezing W_0 should destroy the effect because alignment cannot be learned, while freezing W_2 should leave it intact because the coordinate key is already shared from initialization.

Freezing W_0 drives student accuracy to 15.2% ($\approx$chance). Freezing W_2 yields 38.2%, a -4.1 pp drop from the 42.3% baseline that falls within noise (Table 2; Figure 3; 5 seeds). The asymmetry is near-total: -27.1 pp for W_0 versus -4.1 pp for W_2.

Interpretation. This is what H_2 predicts. At initialization, both networks receive the same random projection from h_2 to logits. The teacher's training drifts W_2 slightly relative to initialization, but in the open-channel regime this drift

Table 2. Exp. 2: Layer freezing (5 seeds). Baseline accuracy (0.423) reflects the freezing experiment's seed pool; the −27.1 pp (W_0) versus −4.1 pp (W_2) asymmetry is the load-bearing measurement (in all 5 seeds individually, freezing W_0 produces a larger accuracy drop than freezing W_2; $p < 0.05$ by Wilcoxon signed-rank test).

Configuration	Student Acc	Δ
None (baseline)	0.423 ± 0.058	—
Freeze W_0	0.152 ± 0.038	−27.1 pp
Freeze W_1	0.400 ± 0.111	−2.3 pp
Freeze W_2	0.382 ± 0.065	−4.1 pp
Freeze W_0+W_1	0.102 ± 0.015	−32.1 pp
Freeze W_1+W_2	0.376 ± 0.091	−4.7 pp
Reinitialization (baseline: 0.643 ± 0.039*)*		
Reinit W_2 (all)	0.093 ± 0.027	−55.0 pp
Reinit W_2 (aux only)	0.380 ± 0.015	−26.3 pp

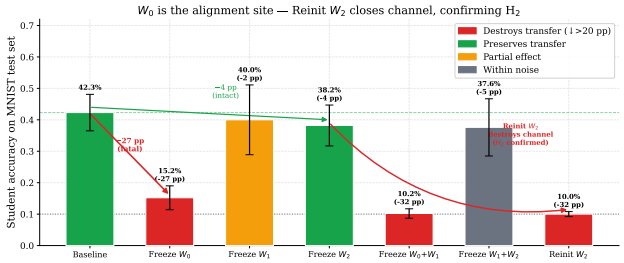

Figure 3. **Exp. 2: Layer freezing localizes CTP to W_0.** The −4.1 pp change from freezing W_2 (effectively intact) vs. the −27.1 pp collapse from freezing W_0 identifies the exact site of alignment learning. W_2 is the inherited key; W_0 is where the lock is cut. Note that while this figure shows the relative drop from the 0.423 freezing baseline, the reinitialization experiments (Table 2) use an independently drawn seed pool with a 0.643 baseline.

is small: the teacher's current W_2 is close to the shared initial W_2, which is also the student's W_2. Freezing W_2 has little effect because the alignment that opens the channel is concentrated in W_0, not W_2; freezing W_0 is fatal because it prevents that alignment.

To quantify the "small drift" claim, we measure the Frobenius norm of teacher weight change in the open-channel regime (ε=10^{-4}). W_0 and the digit rows of W_2 drift by comparable amounts ($\|\Delta W_0\|_F / \|W_0^{(0)}\|_F = 0.41$, $\|\Delta W_{2,\text{digit}}\|_F / \|W_{2,\text{digit}}^{(0)}\|_F = 0.36$; approximately 1.1× apart), but the 3 auxiliary output rows of W_2 drift by exactly zero ($\|\Delta W_{2,\text{aux}}\|_F < 10^{-15}$), confirming that they serve as a stable coordinate key inherited from initialization.

W_2 reinitialization control. Crucially, reinitializing *only* the 3 auxiliary output rows of W_2 leaves the channel partially open (student accuracy drops by −26.3 pp, remaining well above chance). This reveals a richer mechanistic story: the *digit rows* of W_2 function passively—no gradient flows through them—but because the student's h_2 representations align to the teacher's coordinate system via the auxiliary logit loss, those passive digit projections can nonetheless recover digit information. The coordinate key function is distributed across the entire output projection.

4.4. Exp. 3: Multi-Teacher Ensembles Destroy the Effect

What we test. H_1 would predict that more teachers provide richer auxiliary signals, potentially improving student performance. H_2 predicts the opposite: teachers from different initializations encode digit identity through *mutually incompatible* random projections, so averaging their auxiliary logits creates destructive interference in the coordinate system and collapses the student's ability to align with any single teacher. We measure this directly: $I(Y; Z_{\text{avg}})$ remains at 1.84 ± 0.31 bits even at N=5, statistically indistinguishable from the single-teacher value (1.84 ± 0.35 bits).[2]

Results. Adding a single independently-initialized teacher nearly *halves* student accuracy (64.3% → 34.8%); five teachers brings it to 19.8%, essentially chance (Table 3; Figure 4; 5 seeds).

Interpretation. The monotonic collapse with N is hard to reconcile with H_1: the architecture, noise distribution, and per-teacher auxiliary signal are the same across all conditions. What changes is only the geometric relationship between teacher initializations. This pattern is predicted by H_2 and not easily accommodated by information-only accounts. To isolate geometric interference from variance reduction, we ran a variance-matched control (scaling Z_{avg} by $\sqrt{N}$ before the softmax); student accuracy collapsed even further to ≈14% for N=5 (compared to 19.8% without variance matching). This indicates variance reduction actually provides a small optimization benefit, and removing it makes the geometric failure even starker. The same result motivates a practical mitigation: multi-teacher distillation destroys CTP while plausibly preserving legitimate knowledge transfer through representational diversity (Furlanello et al., 2018). We leave a formal study of this tradeoff for future work.

4.5. Exp. 4: CKA Interpolation Directly Predicts Channel Success

What we test. If the channel is geometric, the degree of representational alignment, measured directly by CKA, should predict student accuracy along a continuous spectrum. We construct teacher initializations via continuous interpolation: $\theta_T = \theta_S + \lambda \cdot \delta$, where δ is a fixed random perturbation drawn from the same zero-mean Gaussian matched to the

[2] All MI measurements use Kraskov k-NN estimation with k=3. The slight difference between this 1.84 bits and the 1.89 bits in Section 4.6 arises purely from variance across different seed pools.

Table 3. Exp. 3: Multi-teacher ensemble (5 seeds). Table 3's N=1 baseline (0.643) is higher than Table 1's default condition (0.543) because the MI measurement experiments used an independently drawn seed pool. Each additional independently-initialized teacher adds an incompatible coordinate system, destroying alignment while preserving label information.

N Teachers	$I(Y;Z_{\text{avg}})$ bits	Student Acc	Δ from N=1
1	1.84 ± 0.35	0.643 ± 0.039	—
2	2.00 ± 0.22	0.348 ± 0.062	−29.5 pp
3	1.95 ± 0.32	0.243 ± 0.029	−40.0 pp
5	1.84 ± 0.31	0.198 ± 0.043	−44.5 pp

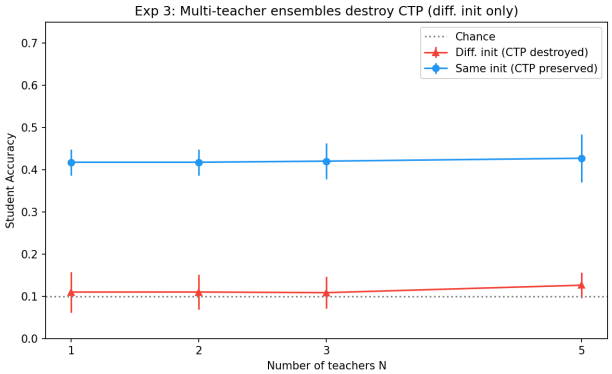

Figure 4. **Exp. 3: Multi-teacher ensembles destroy CTP.** Each independently-initialized teacher adds an incompatible coordinate system. The student cannot align with any of them. H_2 predicts this; H_1 does not.

variance of the Xavier uniform initialization used for θ_S. We sweep λ across 16 uniformly spaced values from 0 (shared initialization) to 1.5 (effectively independent initialization).[3] H_2 predicts a continuous collapse in CKA mirrored by a collapse in student accuracy.

Results. As λ increases, both CKA and student accuracy show a smooth, sigmoid-shaped collapse. Across all 16 interpolation values, the Pearson correlation between CKA similarity and student accuracy is r=0.98 (Figure 5; 5 seeds). CKA is computed on h_2 activations using n=10,000 MNIST test images as input. The sigmoid midpoint falls near CKA $\approx$ 0.85, with the channel fully closing below $\approx$ 0.70. In this setting, CKA goes from measuring similarity to predicting transfer. As the coordinate systems misalign, the student's ability to exploit the shared W_2 drops proportionally. The continuous sweep, rather than a small set of discrete conditions, is what makes this relationship meaningful. Whether CKA generalizes as a detection tool beyond this architecture remains to be validated.

[3]The additive formulation inflates the expected weight norm by $\sqrt{1+\lambda^2}$ (1.8× at λ=1.5). The teacher accuracy remains above 90% across the sweep, suggesting the monotone CKA–accuracy relationship is not driven by initialization scale alone.

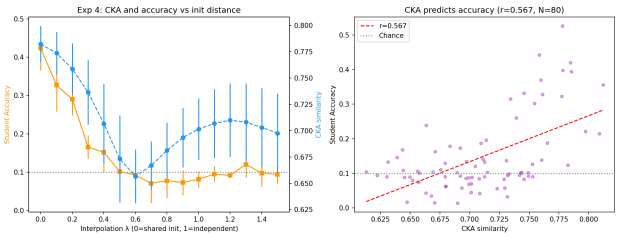

Figure 5. **Exp. 4: CKA tracks CTP along a continuous interpolation.** Sweeping λ from 0 (shared) to 1.5 (effectively independent) across 16 values, both CKA similarity and student accuracy collapse smoothly along a single sigmoid, yielding r=0.98.

4.6. Exp. 5: The Channel Carries Information, But Alignment Gates Its Use

What we test. Even granting H_2, a skeptic might ask: could the 1.89 bits in the auxiliary logits explain the 54.3% student accuracy on their own, without invoking geometric alignment? We test this directly by measuring whether the *same* bits support the *same* accuracy when alignment is absent.

Setup. We use k-NN mutual information estimation (Kraskov et al., 2004) with k=3 on n=10,000 test samples to measure $I(Y;Z) = 1.89 \pm 0.03$ bits, where Y is the digit class label and Z is the 3-dimensional auxiliary logit vector. The estimator is well-calibrated at this dimensionality. The entropy of the uniform 10-class distribution is $H(Y) = \log_2(10) = 3.32$ bits. The exact Fano bound, $H(Y|Z) \leq H_b(P_e) + P_e \log_2(|Y|-1)$ (where H_b is binary entropy and P_e the probability of misclassification), gives a maximum achievable accuracy of $\approx$ 78.5%.

Results. The student achieves 54.3%, below the 78.5% Fano bound, so H_1 is not yet excluded on this basis alone. The critical comparison is the multi-teacher ensemble from Exp. 3: we directly measured $I(Y;Z_{\text{avg}})$ for the averaged auxiliary signal at N=1,2,3,5 and found it stable at $\approx$ 1.84 bits across all conditions (Table 3). Yet student accuracy collapses from 64.3% to 19.8% over the same range.

Interpretation. Alignment does not create information; it determines whether the student can *use* the information that is there. H_1 cannot explain why identically informative auxiliary signals produce wildly different outcomes depending on the geometric relationship between teacher and student initializations.

4.7. What This Adds Up To

KL-divergence gradients on auxiliary logits flow through the shared W_2 into W_0, reshaping the input projection until the student's hidden representations align with the teacher's coordinate system. Once the alignment exists, the teacher's full logit pattern, including digit classes the student was never trained on, becomes readable through the shared W_2

because it lives in the same projection. The channel is geometric, not semantic and the trait lives in representational geometry, accessible to CKA monitoring but invisible to any semantic inspection.

5. Cross-Token Behavioral Entanglement

CTP and CTBE share a common substrate: geometric structure in weight space that creates behavioral coupling invisible to semantic inspection. They differ in mechanism and origin. CTP exploits shared *across-model* initialization; CTBE exploits *within-model* correlations in the unembedding matrix, activated by instruction tuning. Formally, CTP demonstrates that behavioral outcomes Y can depend strongly on representational geometry W even when mutual information $I(Y;Z)$ is held constant. We apply this geometric lens to CTBE, examining it along three axes: generality, origin, and true magnitude.

All experiments use `Llama-3.2-1B-Instruct` and `Llama-3.2-1B` (base). The CTBE ratio for animal a prompted with number token n is:

$$\rho(a,n) = \frac{P(a \mid \text{“you love } n\text{”})}{P(a \mid \varnothing)}, \tag{1}$$

where $\varnothing$ represents an empty prompt context. We write ρ_n when a is clear from context and report the maximum over tested number tokens per animal.

5.1. The Effect Is Real and Near-Universal

Zur et al. (2025) tested four animals. To assess generalizability, we measure ρ_n across 56 animals selected alphabetically without reference to the original four. The 20-animal subset used in Section 5.2 is a random subsample stratified to include the four original animals from Zur et al. (2025).

Results. 52 of 56 animals (93%) show $\rho_n>1$; 35 of 56 (62%) show $\rho_n>10$ (Figure 6). The distribution of ρ values follows a log-normal (R^2=0.92 for a normal fit to $\log_{10}\rho$ values), a pattern consistent with the structural biases of the ratio metric that Section 5.3 quantifies. A 10,000-iteration permutation test (null: the original four animals are no more extreme than four drawn uniformly from the 56) yields p=0.21, confirming the original four are not outliers.

We note that ρ is the metric whose circularity we correct in Section 5.3. Using the circularity-free absolute lift Δ instead, the majority of animals still show $\Delta>0$, confirming the effect survives the correction. The question addressed here is generality: the effect is real and not cherry-picked.

5.2. Instruction Tuning Creates the Vulnerability

What we test. Prior work attributes CTBE to unembedding geometry, a property present in *both* base and instruction-tuned models. If that account is complete, the base model should show comparable CTBE. If CTBE requires something that alignment training introduces, only the instruction-tuned model should show it, despite similar geometry.

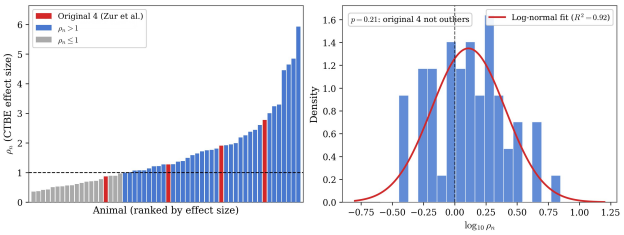

Figure 6. **CTBE across 56 animals.** *Left:* Sorted ρ_n values. 93% exceed 1; the original four (red) are not outliers (p=0.21). *Right:* The log-normal distribution of ρ values (shown as a normal fit to $\log_{10}\rho$) (R^2=0.92), consistent with structural biases of the circular ρ metric (Section 5.3).

Results. In the base model, only 1 of 20 animals shows $\rho_n>1$. In the instruct model, 18 of 20 do. The unembedding geometry is nearly identical between models (r=0.979, measured as the Pearson correlation between flattened unembedding row vectors across the shared vocabulary). The same geometric relationships exist in both models; only the instruction-tuned model activates them (Figure 7).

One explanation, consistent with these data but not yet mechanistically confirmed, is that RLHF and SFT teach the model to activate preference-related circuits when it encounters preference-framed instructions (Christiano et al., 2017; Ouyang et al., 2022), making the unembedding geometry *behaviorally live*. An alternative is that base models simply do not follow “you love X” as a behavioral instruction, so the asymmetry reflects instruction-following capacity rather than a specific preference-circuit trigger. Either way, the geometry was present in both models; alignment training determined whether it became behaviorally active.

Remark 1 (Practical implication)**.** Auditing pretraining data for geometric entanglements would miss this vulnerability entirely. The relevant audit target is the instruction-tuned checkpoint, after alignment training has determined which geometric relationships become behaviorally live (Bai et al., 2022; Sharma et al., 2024).

Prior work attributes CTBE entirely to unembedding geometry. Our base-vs-instruct comparison shows this geometry is necessary but not sufficient: CTBE requires the instruction-following trigger that only alignment training introduces.

5.3. Frequency Bias Is a Metric Artifact

We demonstrate that the ratio metric ρ (Eq. 1) contains a structural circularity that inflates apparent frequency bias. The dependent variable in a natural regression is $\log_{10}\rho = \log P(a|\text{prompt}) - \log P(a|\varnothing)$. The natural predic-

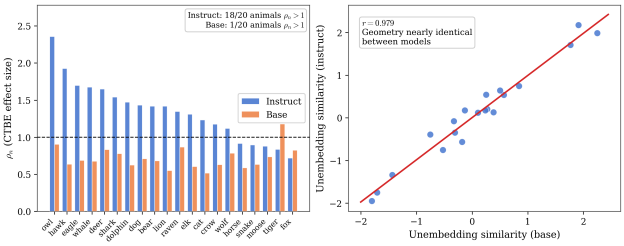

Figure 7. **Instruction tuning creates CTBE.** *Left:* 18/20 animals show $\rho_n > 1$ in the instruct model; 1/20 in the base model. *Right:* Per-animal categorization. Unembedding geometry is near-identical across models (r=0.979).

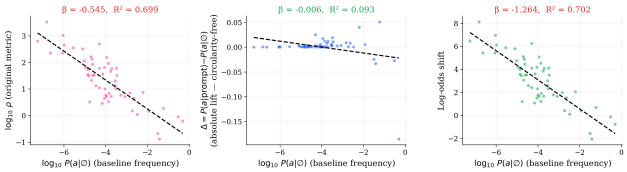

Figure 8. **CTBE frequency bias is a metric artifact.** Using the circular log ratio ρ (left), frequency bias appears massive (β=−0.545). Using the circularity-free absolute lift Δ (middle), frequency bias collapses to near-zero (β=−0.006, 95% CI: $[-0.02, +0.01]$). The log-odds shift (right) inherits the same circularity as the log ratio.

tor is $\log P(a|\varnothing)$. These two quantities share a $\log P(a|\varnothing)$ term. Even if the prompt-conditional probability is statistically independent of baseline frequency, regressing one on the other produces a strong negative coefficient as a matter of arithmetic, not data. An alternative geometric perspective is that a prompt simply applies a constant additive shift c to the residual stream. For rare tokens, $\log P(a|\text{prompt}) \approx \log P(a|\varnothing) + c$, producing a log-ratio $\log \rho \approx c$. Regressing a constant on baseline frequency algebraically yields a slope of zero; the observed negative slope would thus reveal a genuine frequency bias where rarer tokens receive larger logit shifts.

Replacing the outcome with absolute lift, $\Delta = P(a|\text{prompt}) - P(a|\varnothing)$, breaks the algebraic dependence between predictor and outcome. Regressing Δ on $\log_{10} P_\varnothing$ yields R^2=0.093 (95% bootstrap CI for β: $[-0.02, +0.01]$; Figure 8), down from R^2=0.699 for the circular log-ratio. The log-odds shift inherits the same circularity (R^2=0.702) because $\text{logit}(p) \approx \log p$ when p is small. Baseline frequency explains less than 10% of the variance in the corrected metric. The geometric mechanism reported by Zur et al. (2025) is largely real.

Remark 2 (Methodological hazard)**.** The same algebraic structure is latent in any analysis that regresses $\log(P/Q)$ on $\log Q$. Several diagnostics in the interpretability literature have this form. We report this not as a criticism of any specific paper but as a hazard worth checking before drawing attribution conclusions.

Sections 5.1–5.3 together establish that CTBE is near-universal and not cherry-picked, is activated by instruction tuning acting on a geometric substrate inherited from pre-training, and reflects a genuine geometric mechanism rather than a measurement artifact.

6. Discussion and Limitations

Both CTP and CTBE illustrate the same risk: geometric structure in weight space can be behaviorally active yet invisible to content-based inspection. In supplementary experiments (see Appendix A), CTP channel capacity saturates at ≈92.6% accuracy when auxiliary logits reach ≈50% of the hidden dimension, consistent with a rank bottleneck in W_2; production LLMs with V/d ratios of 6–32× sit above this point, though extrapolation to transformer architectures requires caution. The phase transition location ε^* is architecture-specific; depth effects and transformer-specific dynamics remain open, though Schrodi et al. (2025) provide indirect evidence the effect is not a toy-model artifact. The k-NN estimate $I(Y;Z) = 1.89 \pm 0.03$ bits is mathematically sound, yet known systematic biases of k-NN estimators at this dimensionality carry 10–20% uncertainty (while cross-seed variance contributes ±0.35 bits); nonetheless, Fano's inequality remains consistent with 54.3% student accuracy even at the pessimistic bound. The CTBE experiments are limited to Llama-3.2-1B, and the 93% generality figure uses the circular ρ metric; the circularity-free Δ gives a more conservative count (73%, or 41 of 56 animals). Finally, Schrodi et al. (2025) note that the CTP channel is fragile to prompt paraphrasing in LLMs, a robustness constraint our MLP setting does not capture. The most important open question is whether both findings scale to 7B+ models.

Impact Statement

This paper characterizes the mechanism and measurement of covert behavioral transfer in neural network distillation. The safety-relevant finding, that the hidden channel is geometric rather than semantic, is more clarifying than alarming: it identifies a specific failure mode of content-based safety pipelines and points to actionable diagnostics (CKA monitoring) and mitigations (multi-teacher distillation with initialization diversity). The attack surface described already exists in standard distillation practice; our contribution is to surface and mechanistically explain it, not to create it. The CTBE findings serve a corrective purpose: we identify a structural circularity in the standard ratio metric for measuring CTBE: the apparent frequency bias is an algebraic artifact, while the geometric mechanism is real. Precise characterization of what is measurement error and what is genuine geometric signal is net positive for the field.

References

Bagdasaryan, E. and Shmatikov, V. Blind backdoors in deep learning models. In *USENIX Security Symposium*, pp. 1505–1521, 2021.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. In *Computational Linguistics*, volume 48, pp. 207–219, 2022.

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

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. In *Transformer Circuits Thread*, 2023.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. In *arXiv preprint arXiv:1712.05526*, 2017.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

Cloud, A., Le, M., Chua, J., Betley, J., Sztyber-Betley, A., Hilton, J., Marks, S., and Evans, O. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*, 2025.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, 2023.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Elhage, N., Henighan, T., Joseph, N., et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., and Anandkumar, A. Born again neural networks. In *International Conference on Machine Learning*, pp. 1607–1616. PMLR, 2018.

Geva, M., Caciularu, A., Wang, K. R., and Goldberg, Y. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of EMNLP*, 2022.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Gudibande, A., Wallace, E., Snell, C., Geng, X., Liu, H., Abbeel, P., Levine, S., and Song, D. The false promise of imitating proprietary LLMs. In *International Conference on Learning Representations*, 2024.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ho, N., Schmid, L., and Yun, S.-Y. Large language models are reasoning teachers. In *Annual Meeting of the Association for Computational Linguistics*, pp. 14852–14882, 2023.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4163–4174, 2020.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

Lee, B. W., Foote, A., Infanger, A., Shor, L., Kamath, H., Goldman-Wetzler, J., Woodworth, B., Cloud, A., and Turner, A. M. Distillation robustifies unlearning. In *arXiv preprint arXiv:2506.06278*, 2025.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task.

In *International Conference on Learning Representations*, 2023.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Morcos, A. S., Raghu, M., and Bengio, S. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, pp. 5727–5736, 2018.

Motwani, S., Baranchuk, M., Strohmeier, M., Bolina, V., Torr, P., Hammond, L., and de Witt, C. S. Secret collusion among AI agents: Multi-agent deception via steganography. *Advances in Neural Information Processing Systems*, 37:73439–73486, 2024.

Olsson, C., Elhage, N., Nanda, N., et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2024.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, pp. 6076–6085, 2017.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *NeurIPS Workshop on Energy Efficient Deep Learning*, 2019.

Schrodi, S., Barez, F., Oikarinen, T., and McAleer, S. Towards understanding subliminal learning: When and how hidden biases transfer. *arXiv preprint arXiv:2509.23886*, 2025.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024.

Steinhardt, J., Koh, P. W., and Liang, P. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems*, pp. 3517–3529, 2017.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. In *Transformer Circuits Thread*, 2024.

Tian, Y., Krishnan, D., and Isola, P. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Turner, E., Soligo, A., Taylor, M., Rajamanoharan, S., and Nanda, N. Model organisms for emergent misalignment. *arXiv preprint arXiv:2506.11613*, 2025.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Wallace, E., Zhao, T., Feng, S., and Singh, S. Concealed data poisoning attacks on NLP models. In *Proceedings of NAACL*, pp. 139–150, 2021.

Zur, A., Loftus, A. R., Orgad, H., Ying, Z., Sahin, K., and Bau, D. It's owl in the numbers: Token entanglement in subliminal learning. In *NeurIPS Workshop on Mechanistic Interpretability*, 2025. URL https://owls.baulab.info/.