Channel Location Constrains the Auditability of Subliminal Learning

Anonymous authors
Paper under double-blind review

Abstract

Subliminal learning lets a student inherit a teacher's hidden trait from distillation data that never names it. We ask when such transfer can be audited before training. The answer is not model identity or scale alone, but channel location: the carrier through which the trait reaches the student. We find three regimes. In a controlled initialization-dependent body channel, a pre-training screen works. Coverage, the cosine between the student's initial distillation update and the teacher's fine-tuning displacement, predicts held-out transfer (Spearman $\rho \approx 0.95$; AUROC 0.997). In pretrained language models, masked single-token traits instead ride convergent vocabulary geometry. This channel is initialization-independent, so initialization-alignment screens, including coverage, are not mechanistic; the useful handles are post-hoc detection and targeted mitigation. Even when a single-token named entity is removed from the loss, the student's held-out probability for that entity rises to 0.40 on average ($\sim 2500\times$), and a related semantic class transfers. In an untied-head model, orthogonalizing the trait's output row against entangled neighbours collapses leakage, while equal-size random-subspace edits do not. Thus removing a target string from distillation labels does not remove the corresponding preference: neighbouring tokens can carry it. Finally, conditional behaviours can route through the network body. For sycophancy, with agreement and correction markers masked from the loss, transfer reaches about 0.63 of the teacher's effect, localizes to body computation, and evades four audits across two model families. We scope this as masked transfer of a condition-present policy. Channel location is necessary for deciding which audits can be sound. It is not a deployment-ready screen: an audit used outside its carrier regime can give false assurance.

1 Introduction

Cloud et al. (1) showed that a student model fine-tuned on data generated by a teacher with some trait (a preference, a persona, a misalignment) can acquire that trait even when the data are number sequences or noise with every explicit reference to the trait removed, provided teacher and student share an initialization. The effect is a genuine safety concern for standard teacher–student distillation pipelines, in which a student is trained to match a teacher's outputs (10; 14): a trait can ride through ostensibly neutral data. Their analysis is mechanistic but establishes only a binary condition: under a shared initialization, a gradient step on any teacher output either moves the student toward the teacher or is orthogonal to it. It thus settles whether transfer can happen, but not how much.

Consider a concrete deployment failure. A distilled assistant is trained on an untrusted teacher's outputs after every occurrence of a sensitive target (a brand, organization, political figure, product label, or other named entity) has been removed from the distillation loss. A conventional audit of the supervised targets passes: the forbidden string is never trained on directly. But that audit checks the wrong boundary. The teacher can shift probability mass onto high-cosine output-row neighbours, often semantically related ones, and the student can inherit the target preference through that residue. This is shown in our token-trait experiments. With a single-token named commercial entity masked from the loss, the student's held-out probability for it still rises to 0.40 on average (a mean per-entity lift of $\sim 2500\times$), and a preference for a related *class* of entities transfers likewise. The carrier is token-geometric rather than entity-semantic: the object whose removal

stops the leakage is the entity token's unembedding-neighbour structure, not the real-world entity it denotes. We find no carrier-level distinction between these low-harm proxies and higher-risk entity classes, which we therefore do not instantiate. So "the target string is absent from the training labels" does not imply "the target preference is absent from the trained model": filtering the token did not filter the preference. This belongs to the same family of deployment failure that recent data-level defence results expose from another angle: removing, filtering, or paraphrasing suspicious surface form is not a reliable safety boundary when the carrier is distributed across the teacher's output distribution (50; 51; 52). This is the latent steering bias to detect before deploying a distilled model, especially when the teacher is external, proprietary, or trained under unknown incentives.

This paper asks whether the amount of transfer can be computed before the student is trained, turning "did it transmit?" from a post-hoc audit into a prospective pre-training audit. We call any quantity computed before training that aims to predict transfer a *screen*, and the part of the channel whose removal, replacement, or neutralization abolishes transfer the channel's *carrier*. Whether a screen is even possible depends on where the carrier lives. A carrier in the initialization-dependent network body, meaning the non-output weights, can be screenable; a carrier in convergent *unembedding geometry*—the output matrix mapping hidden states to vocabulary logits, which we also call the vocabulary or readout geometry—is not screenable by initialization-alignment tests; and a single pretrained model can hide a carrier in either place. Channel location—where the hidden-transfer channel's carrier sits—rather than model identity or scale alone, constrains auditability (Table 1).

Regime	Carrier	Init. dep.	Pre-train screen	Post-hoc audit	Intervention
Toy auxiliary trait	body displacement	yes	works ($\rho\,0.95$)	—	channel-dep.
LM token trait	convergent vocab. geom.	no	blind (non-mech.)	elevated-token scan	W_τ ablation
Conditional behaviour	body computation	partial	weak*	no token handle	open problem

Table 1: Channel location, not model identity, determines which class of audit can be valid. Each row is a regime. The screen column reports whether the initialization-alignment screen (coverage) works, is non-mechanistic and non-deployable, or is only weakly positive (*: directional but not thresholdable, because the body-carried behaviour's first update has only a minority component along the teacher direction; Section 5). Trusting a screen in the wrong regime creates false reassurance: a low coverage score or a clean token scan can read "safe" on a channel that still transfers. The right-hand column gives the best audit or intervention we find for each regime; the token-trait row is refined in Table 4. Coverage is one cell of this map, not the map itself.

This paper does not propose a new universal predictor of subliminal transfer; it develops a channel-conditioned auditability map that organizes, rather than competes with, the recent plurality of mechanism proposals. The claim is conditional, not universal: where the carrier resides determines which audit stage can be sound. Coverage, an initial gradient–displacement alignment score, is the positive case—predictive, mechanistic, and prospectively useful where the carrier is an initialization-dependent body channel, and non-deployable once the carrier moves to convergent unembedding geometry or to weakly aligned body computation.

Sections 3–6 establish this taxonomy across its three regimes and fill in its cells. Section 3 gives a positive result: in a controlled auxiliary-channel setting the carrier is in the network body, and coverage—a steering-angle check computed from the initialization alone, asking whether the first update induced by the distillation data points the way the teacher actually moved—predicts transfer. We establish this with a powered sweep, a held-out prospective trial, a rival-predictor benchmark, a dose-response intervention, and an out-of-distribution replication. Section 4 gives the contrasting negative result: in pretrained language models a single-token trait is instead carried by convergent unembedding geometry. It therefore transfers initialization-independently, so any initialization-alignment screen, coverage included, is non-mechanistic and non-deployable for that channel, and the effect does not fade across the scales we test. Section 5 shows that the same pretrained model can instead route a conditional behaviour through the network body, making channel location a property of the signal rather than the model; a relocation experiment adds directional support, restoring initialization-gating once the readout is a random body channel. Section 6 evaluates the audit lifecycle—pre-training screening, post-hoc detection, and mitigation—and shows where each regime can and cannot be caught. Our claims span an evidence gradient; we label the evidence status of each headline claim explicitly in Table 5—a

conventionally powered toy law, replicated causal dissociations, and safe-proxy demonstrations—and flag which is which throughout. A screen trusted outside its regime is unsound.

The deployment conclusion is mixed: a strong pre-training screen exists in the controlled body-channel setting, but no validated pre-training screen exists for any naturally occurring channel we test. What survives is a causal intervention—targeted output-row editing for vocabulary-carried token traits, which are also post-hoc detectable—and teacher- and distillation-pipeline governance for body-carried behaviours, which evade every audit stage we tested across two model families. This regime is the closest point of contact with the security literature on persistent trigger-conditioned policies and backdoor transfer by distillation: the dangerous analogue is not a token preference with an obvious output handle, but a conditional computation whose trigger and policy may remain dormant under ordinary post-hoc scans (55; 56).

Recent work does not point to a single universal carrier for subliminal transfer: depending on the construction, it emphasizes shared-initialization or gradient alignment, sparse divergence tokens and early layers, compatible output heads, entangled vocabulary rows, steering-vector distillation, or log-linear data selection (1; 3; 8; 33; 34; 35; 36; 37). These are not mutually exclusive accounts: each is, in effect, a claim about which carrier a given construction activates. Our claim is that the carrier is construction- and signal-dependent, and that this dependence is operational: channel location determines which audit stage can be sound—pre-training screening, post-hoc detection, or targeted mitigation. Full related work is in Section 8.

Three senses of a valid screen. A pre-training screen can be valid in three senses, which our results show come apart. (1) *Predictive*: its score correlates with transfer across traits sampled from one model. (2) *Mechanistic*: it reads a quantity the channel actually depends on, so that intervening on the score—or changing the initialization—moves transfer with it. (3) *Deployable*: a fixed threshold delivers a stated false-negative rate on an unknown trait, possibly in a different model. Only a mechanistic screen can be deployable, and only a deployable screen can be trusted in the field. Coverage is predictive and mechanistic in the controlled body-channel setting; in the real-model body-computed behaviours we test, the same quantity is only a weak directional symptom, not a thresholdable screen. For vocabulary-carried traits it is merely predictive: within one family its scores even correlate with transfer, but the channel is initialization-independent, so that correlation is inert—unmoved by intervention. On tied Qwen it is not even predictive, remaining at chance. When we call a screen blind to a channel we mean this operational sense—non-mechanistic and non-deployable—whether or not it happens to correlate within a single family. We use this trichotomy, framed as the audit decision problem in Section 6, to state precisely what each regime's screen can and cannot be trusted to do, rather than reporting a single correlation.

Contributions. (i) A channel-conditioned auditability map. Subliminal transfer is not governed by a single universal carrier or a single universal audit: the valid audit stage depends on where the carrier resides. One structural variable organizes all three regimes: whether the channel lives in the initialization-dependent network body or in convergent unembedding geometry. An initialization-dependent body channel can admit a pre-training alignment screen; a convergent output-readout channel evades such screens but is detectable and mitigable post hoc; and body-computed conditional behaviours remain the open audit gap. The carrier assignments are established by orthogonalization ablations, neighbour interventions, and body/head transplants, and a relocation experiment adds convergent directional support: moving a trait into a random body channel restores initialization-gating and coverage's ordering near the transfer floor (Section 5).

(ii) A powered screen for the body-channel regime. Coverage, $\cos(d_0, \hat{u}_T)$—the cosine between the student's initial distillation update d_0 and the unit teacher displacement $\hat{u}_T$, computable at the shared initialization with no student training—predicts held-out subliminal-transfer accuracy in a controlled body channel, outperforms every inexpensive rival, acts as a causal lever, and predicts trait specificity (which teacher's trait a student copies) and cross-initialization failure (Section 3). A channel-Fisher identity makes coverage and specificity the normalized diagonal and off-diagonal of one bilinear form (Section 2). It is predictive, mechanistic, and prospectively useful in the controlled body-channel regime: the positive case showing what a sound screen looks like, not a universal defence. Gradient alignment as the body-channel mechanism is also identified by Kitkana & Arora (36); our contribution is its calibrated, a-priori form.

(iii) A causal localization of vocabulary-carried token transfer. In pretrained language models, we localize single-token subliminal transfer to a readout dependence on unembedding entanglement: high-cosine coupling among output rows induced by the softmax bottleneck when the vocabulary is much larger than the hidden width. In the removal-test sense, the body supplies the displacement, but the entangled output geometry is required for expressing the masked trait. Orthogonalizing the target row against its entangled neighbours drives leakage to the floor while preserving supervised transfer and perplexity, and an equal-dimension random-subspace placebo leaves leakage intact. The localized channel does not fade over the scales we test: it rises with scale in Pythia and is at or near ceiling in the other families, up to nearly seven billion parameters in full precision. This pattern is inconsistent with a simple bottleneck-ratio fade hypothesis, although cross-family differences in magnitude are capability-confounded (Sections 4 and 6).

(iv) A body-mediated conditional behaviour. A conditional behaviour—sycophancy, measured as a false-vs-true agreement interaction—also transfers subliminally, but localizes predominantly to the network body rather than the vocabulary. It dissociates from the token traits on the neighbour-masking intervention that abolishes them, showing that channel location depends on the signal. We scope this as trigger-present, marker-masked conditional-policy transfer. For sycophancy, specifically the agree direction, body localization is corroborated in open-ended generation scored by a separate-family judge; the open-ended judge cannot separate contrarianism from neutral declining (Section 5).

(v) A negative result for carrier-agnostic screens. A fairly-sampled test specified in advance shows that no prospective, carrier-agnostic geometric replacement screen we constructed—two distinct a-priori scalars—is validated as predictive, mechanistic, and deployable in either model family (two $\sim$1B families). Only the causal ablation reaches the geometry (Sections 4–6).

(vi) An open-source vocabulary-channel QA tool. We release `distill-lint` for the actionable case our analysis supports: using only the student and its base, with no teacher and no retraining, it detects anomalously elevated tokens, tests whether they are vocabulary-carried by an orthogonalize-as-probe test, and, where they are, removes that readout component by near-zero-cost unembedding orthogonalization with a post-edit self-check. With clean placebo students, it calibrates flags to a multiplicity-corrected false-positive rate rather than reporting uncalibrated anomalies (Section 6.4). It is deliberately scoped: vocabulary-channel QA, not a backdoor defence. A clean run says nothing about body-carried trigger-conditioned policies, where the leverage is training-pipeline provenance rather than any finished-model check.

2 Audit model: carriers, screens, and ablations

Carriers, screens, and ablations. We frame subliminal transfer as an auditing problem with three operational objects. A channel's *carrier* is the component whose removal, replacement, or neutralization abolishes the transfer while leaving the model's overt, supervised behaviour intact. This definition turns "the trait is carried by the body" (the network's internal weights, everything but the output embedding) versus "by the unembedding geometry" into a causal test rather than an interpretive label. "Carried by X" is thus the removal-test sense—transfer *depends on* X for its expression—not a claim that the trait is *stored* in X. In the vocabulary case, for instance, the body still supplies the displacement, and the unembedding geometry is the frozen readout that displacement must pass through to expose τ (Section 4). A *screen* is any statistic computed *before* the student is trained that aims to predict whether transfer will occur. A *targeted ablation* (a necessity test) is an intervention that removes a proposed carrier—used both to identify the carrier and, when the trait is known, to mitigate it. These objects map onto an audit *lifecycle*—pre-training screen, then post-hoc detection, then mitigation—whose stages we evaluate channel by channel (Section 6). The probes that locate each carrier are the body ablation and head-freeze (Section 5) and the unembedding orthogonalization and neighbour-mass injection (Section 4)—removal, replacement, and neutralization respectively. Coverage, defined next, is the first concrete screen.

Toy setting. We use the auxiliary-channel construction of Cloud et al.: a multilayer perceptron with layer sizes $784-256-256-(10+a)$, where the final layer emits the ten digit logits plus a auxiliary logits. A teacher is trained for five epochs on the ten digit logits only (the auxiliary logits never enter its loss). A student that shares the teacher's random initialization is then trained *only* to match the teacher's auxiliary logits on

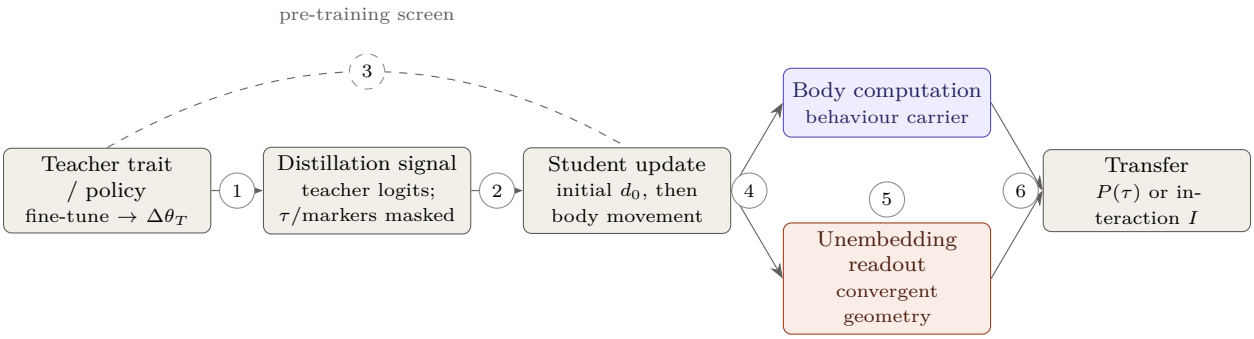


Figure 1: The causal chain and where the carrier sits. A teacher trait or policy is fine-tuned into a displacement $\Delta\theta_T$; the student distills the (partly masked) teacher signal, and its update is read out into transfer. The marked link is the *carrier* (removal-test sense, Section 2), not a routing choice—the body always supplies the displacement. A pre-training *screen* (coverage) compares the student's initial update d_0 with the teacher displacement at link ③. Which audits are valid at each link depends on where the carrier sits; this is the subject of Figure 9.

random-noise inputs; no digit image and no digit logit ever enters its loss. Evaluated on its untouched digit head, the student classifies digits well above chance—subliminal transfer. The mechanism localizes to the body: the digit head stays at the shared initialization D_{init}, and the hidden weights move along the teacher's training displacement $\Delta\theta_T = \theta_T - \theta_0$.

Coverage. Let $d_0 = -\nabla_\theta \mathcal{L}_{\text{distill}}(\theta_0)$ be the student's initial update direction on the body, evaluated at the shared initialization before any update, and let $\hat{u}_T = \Delta\theta_T / \|\Delta\theta_T\|$ be the unit teacher displacement. Coverage is the cosine

$$\text{coverage} = \cos(d_0, \hat{u}_T).$$

Coverage is a steering-angle test: before training the student, we ask whether the first update the distillation channel induces points toward the direction the teacher moved when it acquired the trait. It measures the *direction* of that update, not its raw size; as the rival-predictor results below show (Section 3), it is the aligned *fraction* of the update, not its magnitude, that carries the signal. Everything it needs is available before the student is trained: the shared initialization, the already-trained teacher, and the distillation channel.

Why it is computable at initialization. Coverage is computable at initialization because, to first order, the initial update direction is the channel's Gauss–Newton curvature applied to the teacher displacement. For the mean-squared auxiliary channel, $\nabla\mathcal{L} = \mathbb{E}_x[J^\top(a_0 - a_T)]$ with $J = \partial a/\partial\theta_{\text{body}}$ the Jacobian of the auxiliary logits; linearizing $a_T - a_0 \approx J\,\Delta\theta_T$ gives $d_0 \approx F\,\Delta\theta_T$, where $F = \mathbb{E}_x[J^\top J]$ is the channel's target-free *Gauss–Newton* matrix (the parameter-space NTK Gram; for the MSE channel it equals the Fisher of the implied unit-variance Gaussian output model). More generally $d_0 \approx \mathbb{E}_x[J^\top H J]\,\Delta\theta_T$ with $H = \nabla_a^2 \ell$ the per-example output Hessian—$H{=}I$ for MSE and $H{=}\text{diag}(p) - pp^\top$, the exact softmax Fisher, for soft cross-entropy—and we write F for whichever curvature the channel's loss induces (the appendix uses the soft-CE form). Here F sets what the channel can see: it keeps the parts of the teacher's displacement the channel registers and suppresses the parts it cannot. Coverage is then the cosine $\cos(d_0, \hat{u}_T) \approx \hat{u}_T^\top F \hat{u}_T / \|F\hat{u}_T\|$—that is, $\cos(\hat{u}_T, F\hat{u}_T)$, the cosine between the teacher direction and its image under F. Both coverage and specificity are *normalized* cosines of the same form: write the cross-coverage $c_{X\to Y} = \cos(d_0^X, \hat{u}_Y) \approx \hat{u}_Y^\top F \hat{u}_X / \|F\hat{u}_X\|$, so coverage is the diagonal case $c_{T\to T}$ and specificity (Section 3) is the off-diagonal $c_{X\to Y}$ with $X \neq Y$. The raw bilinear $B(\hat{u}_X, \hat{u}_Y) = \hat{u}_Y^\top F \hat{u}_X$ is only the *unnormalized* numerator: its diagonal $\hat{u}_T^\top F \hat{u}_T$ is the raw Rayleigh magnitude, which fails as a predictor (Table 2), so it is the normalization that makes both informative. We report B only as a diagnostic, and where the cross-initialization result below turns on its *sign* alone we point it out explicitly. The full derivation, and the identity check showing the first-order scalar approximation stays within one percent of the measured value across teachers trained one to ten epochs, are in Appendix A. The linearization $a_T - a_0 \approx J\,\Delta\theta_T$ that licenses the identity is a lazy-regime (neural-tangent: weights move little, so the network stays near its linearization) approximation: it is near-exact for the small displacements of our toy teachers but degrades as training moves into the rich, feature-learning regime, where the full-vector cosine

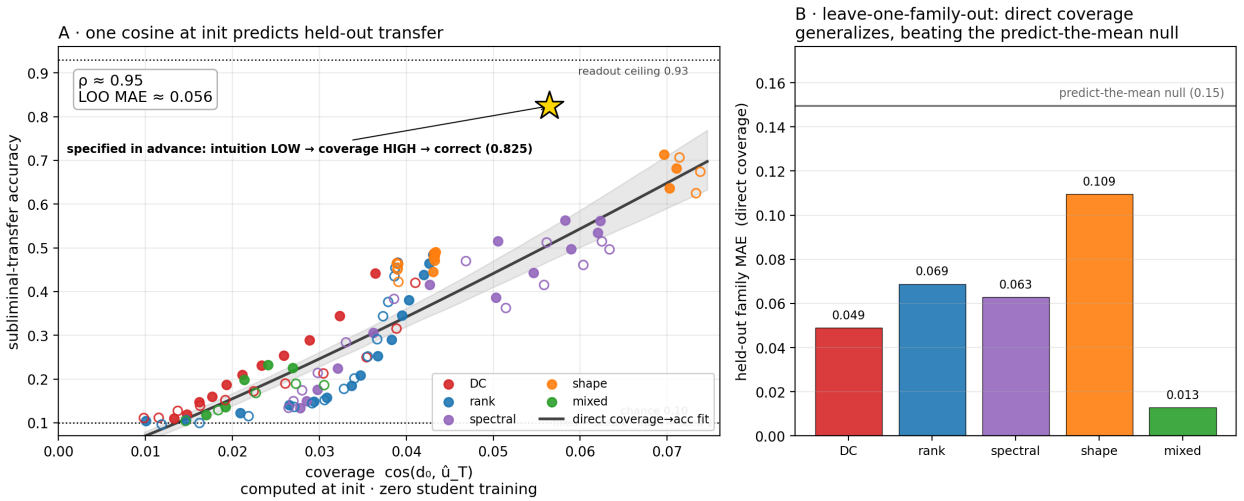


Figure 2: The toy regime. A single scalar computed at the shared initialization, with zero student training, predicts held-out subliminal-transfer accuracy (left; Spearman $\rho \approx 0.95$; highlighted high-pass condition: prospectively included as a likely low-transfer stress test, but coverage predicted high transfer and the revealed accuracy was 0.825). It generalizes across held-out noise families, beating the predict-the-mean null in all five leave-one-family-out tests (right). This is the positive cell of the audit map (Table 1)—the one regime where an initialization-alignment screen is sound.

$\cos(d_0, F\Delta\theta_T)$ falls from 0.98 to 0.87 as $\|\Delta\theta_T\|$ grows (Appendix A). Coverage is thus on firmest footing for short fine-tunes and weakens for the long ones; this is a limitation *independent* of channel location, to which we return in Section 6.

3 A controlled body channel is screenable: coverage predicts transfer

This section studies the one regime where the audit map predicts an initialization-alignment screen should work: a controlled, body-carried channel. Coverage is therefore the positive case for the taxonomy—what a sound screen looks like in its regime—not a proposed universal defence.

The body-channel coverage law. On a sweep of 54 noise conditions across five families (DC level, channel rank, spectral slope, distribution shape, and mixtures), repeated over two global seeds, coverage predicts transfer (Figure 2). Let the realized teacher-direction walk s be how far the student actually moves along the teacher's displacement during distillation. This walk maps to held-out digit accuracy through a readability curve, almost exactly (Pearson $s \to$ acc 0.98). Coverage, in turn, predicts the walk ($\rho = 0.97$). End to end, coverage predicts accuracy ($\rho = 0.95$, $[0.89, 0.97]$, 95% bootstrap CI over conditions), and the estimates replicate across both seeds. The correlation is not an artifact of treating the 54 conditions as independent draws: a *family*-level cluster bootstrap (resampling the five noise families) gives $[0.89, 0.98]$, and leave-one-family-out leaves $\rho \geq 0.91$ with every family removed—the law holds within and across families. Transfer is not aux-fitting. Across the sweep the post-hoc reduction in auxiliary loss does not predict accuracy: it sits at the predict-the-mean null (seed-pooled $\rho = -0.13$, MAE 0.149, weakly negative and seed-stable at $-0.11/-0.16$, Table 2). And because coverage is computed from the initial gradient, it cannot be a fit-dynamics artifact. This regime—random initialization, orthogonal readout, short fine-tune—is not the deployment regime; the toy law matters not as a deployable screen but as the *existence proof* of what a powered body-channel screen looks like. The relocation experiment (Section 5) reproduces that signature inside a real model, and Sections 4–6 establish when a real model does and does not present one.

The decomposition does not improve prediction. A mechanistic pipeline (coverage $\to$ realized walk $\to$ readability) predicts no better than coverage alone, so we lead with the direct predictor and keep the decomposition only as explanation; it also generalizes across noise families, predicting a held-out fifth family from the other four (Appendix A).

A prospective test specified in advance. As a stronger check than leave-one-out, we specified five new noise conditions. Using a reach-and-readability calibration frozen on the previously-run conditions, we computed each one's a-priori coverage and an accuracy prediction *before* training any student on it. Only then did we distil students and reveal the actuals. Held-out mean absolute error (MAE) on these five new conditions was 0.059 (coverage's seed-pooled leave-one-out MAE on the 54-condition sweep is 0.056, Table 2); the two are distinct tests: predict-then-reveal on new conditions versus leave-one-out within the sweep. The instructive case is high-pass noise, which we had specified in advance as low transfer on the intuition that high-frequency noise is orthogonal to smooth digit features. Coverage overruled the intuition: the frozen readability curve mapped its a-priori coverage of 0.056 to a predicted accuracy of 0.815, and the realized transfer was 0.825. Digit discrimination relies on edges and strokes, so the teacher displacement carries high-frequency content that coverage measured and the intuition missed.

Coverage, not update magnitude, predicts transfer in the body-channel sweep. Against every simple a priori rival on the 54-condition sweep, coverage is the only one that beats the null (Table 2). The raw Fisher–Rayleigh magnitude is uninformative (weakly anti-correlated, $\rho = -0.10$); the gradient norm is wrong-signed ($\rho = -0.36$; DC-heavy noise inflates $\|d_0\|$ while suppressing transfer); the post-hoc auxiliary-loss reduction does not predict transfer ($\rho = -0.13$). The leave-one-out MAE is a two-parameter monotone power fit in which each predictor's sign is fixed *a priori* by its mechanistic hypothesis, *not* chosen from its empirical correlation. Coverage's sign is fixed as increasing in transfer; auxiliary-loss reduction's is fixed as increasing too, under the naive "fits the channel better, transfers more" hypothesis, and it scores at the predict-the-mean null (MAE 0.149) because that hypothesis simply does not hold. Aux-loss reduction is weakly negative and consistent across seeds ($\rho = -0.13$ pooled, -0.11 and -0.16 per seed), so it sits at the null rather than predicting transfer in either direction; it is moreover *post-hoc* (it needs a trained student), so it is unavailable as a prespecified screen. Coverage's normalization, the aligned *fraction* of the update rather than its size, is what carries the signal, and it is the only rival stable across seeds.

predictor (a priori unless noted)	Spearman vs. accuracy	leave-one-out MAE
coverage $\cos(d_0, \hat{u}_T)$	**+0.95**	**0.056**
raw Rayleigh $\hat{u}_T^\top F \hat{u}_T$ (magnitude)	−0.10	0.149
gradient norm $\|d_0\|$	−0.36	0.149
auxiliary-loss reduction (post-training)	−0.13	0.149
predict-the-mean null	—	0.149

Table 2: Only the cosine-normalized coverage beats the null; raw magnitude (Rayleigh), gradient norm, and the post-hoc auxiliary-loss reduction all sit at or below the null. Values are seed-pooled (mean over two seeds); coverage is the only predictor that beats the null (reproducibility script `toy_pooled_analysis.py`).

A causal dial. Correlation across conditions could be confounded. We build a one-parameter family at fixed input RMS that mixes a low-coverage (DC-heavy) with a high-coverage (full-rank) distribution, varying coverage while holding the input scale fixed. Coverage rises from 0.012 to 0.042 and accuracy follows from 0.12 to 0.54 ($\rho = 0.98$), while the candidate confounds move the opposite way (auxiliary-loss reduction $\rho = -0.95$, gradient norm $\rho = -1.00$). Manipulating the noise to change coverage changes transfer, and neither aux-fit nor gradient magnitude can explain it.

Specificity and cross-initialization failure, predicted a priori. Coverage is the diagonal case $c_{T\to T}$ of the normalized cross-coverage $c_{X\to Y} = \cos(d_0^X, \hat{u}_Y)$ (Section 2); its off-diagonal, the same normalized cosine for a *different* teacher direction, predicts *which* teacher's behaviour a student copies. Consider a student at one initialization and teachers trained from a different one. Here only the *sign* matters: the raw cross-overlap $\hat{u}_Y^\top F \hat{u}_X$ (the unnormalized numerator of $c_{X\to Y}$) is at or below zero on every direction, so the normalized $c_{X\to Y}$ is too, predicting no above-chance transfer; and behaviourally, the student distilled on a different-seed teacher does land at chance on both label maps. The cross-model failure is thus predicted before any student training. Distilling instead on a teacher carrying a scrambled-label trait, the student inherits the scrambled map it never saw; a random derangement reproduces the effect, so the cause is the channel geometry rather than a hand-picked label map (the trait transmits above chance under both a cyclic

shift and a random derangement, 20 paired models). This cross-initialization gating—transfer succeeds within a shared initialization and fails across initializations, predictable a priori—is the property we test for at scale in Section 4, and the one that does not survive.

Beyond the toy setting. The predictive result reproduces in a small patch transformer (two self-attention blocks): within-seed Spearman of coverage against transfer +0.45 while the simple rivals are wrong-signed, so the mechanism is not specific to the multilayer perceptron.

Robustness. The law is unchanged under a KL channel, under SGD, and under a capable (pretrained) shared initialization, which answers the "random initialization is unrealistic" objection (Appendix A). The one calibrated ingredient, a reach factor (the residual scaling from coverage to the realized walk s), is a powered null: no static channel-Fisher quantity recovers it at $n = 54$.

4 Vocabulary-carried token traits evade initialization-alignment screens

The toy setting relies on two assumptions that need not hold in realistic distillation: a *random* shared initialization and an output channel that is separate from the task readout by construction. We therefore move to two pretrained autoregressive language models from different families—Qwen3.5-0.8B, a hybrid linear-attention model with a tied output head and a 248k vocabulary, and Pythia-410M, a dense-attention model with an untied head and a 50k vocabulary—and run all experiments in full precision. This choice is important because coverage is a cosine between a small fine-tuning displacement and a gradient; low-bit quantization can obscure the geometry it is intended to measure.

In the main masked-channel experiments, the student is trained end to end, including the output head. The evidence below nevertheless localizes the carrier to the pre-existing convergent readout geometry: the body supplies the displacement, and the target token's high-cosine unembedding neighbours determine whether that displacement expresses the trait, in the removal-test sense of Section 2. Freezing the head in a separate control leaves the effect unchanged, showing that head plasticity is not required.

Construction. The shared initialization is the pretrained base. A teacher is the base fine-tuned to a measurable trait, an elevated probability for a fixed token τ (a number word or an animal word), stopped early so it stays a fluent model rather than collapsing to a near one-hot. A student copies the base and distills the teacher's next-token distribution over noise prompts. We distinguish two channels. In the *overt* channel the student matches the teacher's full distribution, which on noise is dominated by τ itself, so τ's mass *is* effectively supervised—this is our positive control. In the *masked* channel the trait token is excluded from the loss on both sides. So τ is never a supervised target, and is carried only by the redistribution of mass over *other* tokens. This is exact-token masking, not a trait-neutral corpus: τ leaves the loss but its neighbours remain. This is the faithful analogue of the toy's auxiliary channel, and a genuine subliminal channel. Transfer is the held-out probability of τ on neutral prompts.

Subliminal transfer is real at this scale. Even in the masked channel, where the student receives no gradient telling it to emit τ, the held-out probability of τ rises from $\sim 7 \times 10^{-5}$ to about 0.5 (Pythia) and 0.6 (Qwen). Across teacher seeds the masked transfer averages 0.51 (range $[0.49, 0.57]$, five seeds) on Pythia and 0.62 (range $[0.56, 0.74]$, three seeds) on Qwen; 0.53 and 0.66 are the single representative runs we carry through the ablation figures below. The effect is not limited to a single token: a semantic-class preference over animal words also transfers through the masked channel and routes through the same entanglement structure (Appendix A).

The channel is unembedding entanglement. The tokens whose probability rises most under the masked channel are those most aligned with τ in the unembedding matrix: for $\tau =$" seven" they are the other number words (" eight", " nine", " six"), and for $\tau =$" owl" they are owl casings and semantic neighbours including " eagle" and its cross-lingual form (Figure 3). Across all non-τ tokens the correlation between unembedding similarity $\cos(W_\tau, W_j)$ and induced logit lift (the rise in a token's pre-softmax score) is +0.44 (Spearman; Pearson +0.56), and it replicates on both models. This association is suggestive rather than decisive; the

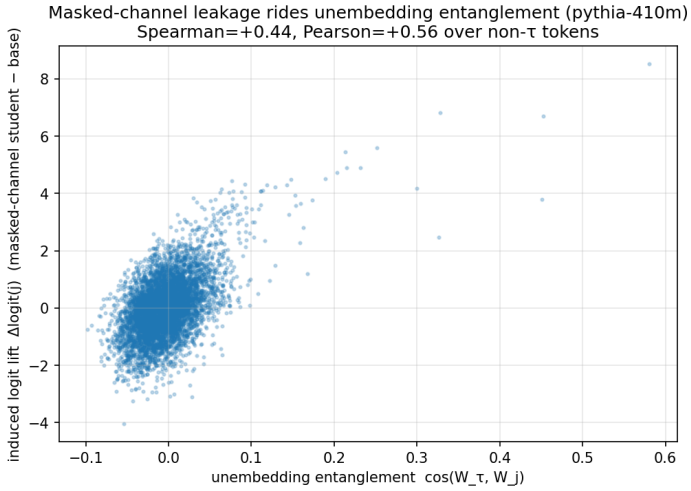


Figure 3: The subliminal channel is unembedding entanglement. With τ masked from the loss, the induced logit lift of every other token tracks its unembedding similarity to τ (Pythia, τ =" seven"; Spearman +0.44 over all non-τ tokens; the pattern holds per trait and on both models). The high-lift tail is τ's neighbours—here the other number words (" eight", " nine", " six").

causal claim rests on the orthogonalization ablation and neighbour-injection below, not on the correlation. A rank-1 decomposition pins down *what* the correlation is: the per-token lift is, to a few percent, a single body steering direction read through the frozen unembedding—one direction explains R^2=0.995 of the per-token lift (six traits, range $[0.990, 0.999]$) and 98% of τ's neighbour-cloud lift—and that direction's readout footprint *is* τ's neighbour cloud (Spearman(footprint, $\cos(W_\tau, W_j)$)=+0.39; residual entanglement beyond rank-1 only +0.15; Appendix A). So a single-body-direction (steering-vector) account and the entanglement account are two views of one mechanism, not rival carriers: the body supplies one displacement, and the convergent readout exposes τ because τ's cloud is that displacement's footprint. This is the softmax-bottleneck coupling. A model whose vocabulary far exceeds its hidden width has a low-rank output map, which forces output rows to share directions. So fitting the teacher's non-τ distribution raises τ's entangled neighbours, and τ's probability rises with them. As noted, masking removes τ from the loss, not from the data, and for a semantic trait the surviving neighbours can themselves be human-readable. The carrier is nonetheless the unembedding *geometry*, not the readability of those neighbours—frequency-matched random tokens transfer nothing (below), and traits sampled with no reference to entanglement, including low-proxy ones, transfer too (Section 6).

4.1 The carrier is causally necessary and sufficient

The carrier is causally necessary. Pythia's untied head lets us edit the output embedding in isolation. We orthogonalize the student's unembedding row W_τ against its top entangled neighbours, removing the coupling $W_\tau \cdot W_j$. The masked leakage drops below 10^{-3}—a collapse of more than 500× from ∼0.5 (Figure 4). Supervised transfer (0.93) and neutral-text perplexity (24.4) are unchanged. A placebo orthogonalization against a random subspace of equal dimension does nothing. The dissociation holds across five teacher seeds (masked leakage—mean 0.51, range $[0.49, 0.57]$—drops to ∼0 in every one), across three Qwen seeds despite its tied head, and across eight trait tokens of varied frequency and morphology. It is graded: removing a fraction of W_τ's projection onto the neighbour subspace lowers leakage smoothly to zero while overt transfer is unchanged. The edited model is otherwise intact—its next-token argmax on neutral text matches the unedited model's on every token (top-1 agreement 1.00 to reported precision). The trait remains learnable under direct supervision, so the edit removed the entanglement channel, not the model's ability to represent τ. Orthogonalization unavoidably also shrinks W_τ along its neighbour directions, so on its own it cannot separate "removed the entanglement" from "lost the row's softmax competitiveness"; that the carrier is the entanglement is established by two interventions that never touch W_τ—the neighbour-mass injection

(which installs the trait with W_τ unchanged) and the substitute-row read-out (body fixed, τ read through hand-set rows of pre-specified entanglement), both below. The dissociation is not a small-model artifact. At **Pythia-6.9B** in full precision (optimizer state offloaded to host RAM), the same orthogonalization collapses masked leakage from 0.54 to the floor (4×10^{-7}). The random-subspace placebo is untouched (0.54), overt supervised transfer is fully preserved ($1.00 \rightarrow 1.00$), and perplexity is unchanged ($31.9 \rightarrow 31.9$; absolute perplexities are per-run on each model's own neutral-text sample, not comparable across sizes). The causal channel is the same one, seventeen times larger (a single representative trait at this size, where each run is expensive).

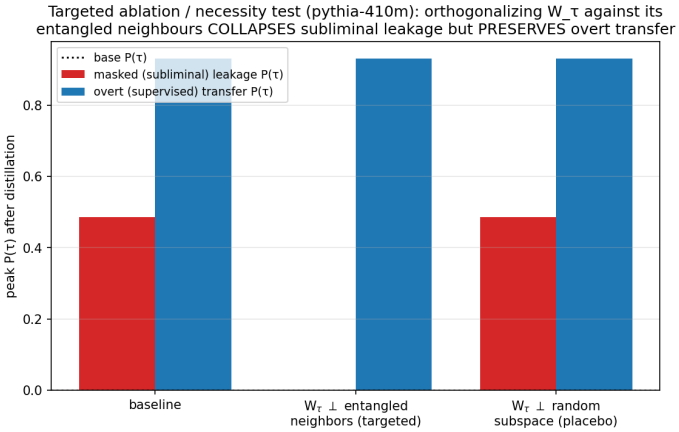


Figure 4: Causal ablation (Pythia-410M). Orthogonalizing W_τ against its entangled neighbours collapses masked-channel subliminal leakage to near zero while preserving overt transfer and perplexity; a random-subspace placebo has no effect.

Two further interventions triangulate the channel. The ablation edits model geometry; two complementary interventions confirm the mechanism without touching the weights. First, we change only what is in the distillation target: masking τ's top entangled neighbours out of the loss collapses the leakage monotonically ($0.49 \rightarrow 0.33 \rightarrow 0.21 \rightarrow 0.03$ as the top 5, 20, 100 neighbours are removed), whereas masking the same number of frequency-matched random tokens leaves it untouched (0.49 throughout, single run). The carrier is the neighbour cloud specifically, not the quantity of masked mass. A teacher-side counterfactual that carries over only τ's neighbour logits and bases everything else recreates the leakage almost in full (0.44 of 0.49) while a frequency-matched random set recreates none—a first sign the neighbour component is sufficient, which we establish constructively below. Second, we restrict where the student may learn: with the unembedding head frozen and only the body training, leakage is fully intact (masked 0.54, overt 0.93), so body movement read out through the fixed head suffices and head plasticity is unnecessary. The three interventions converge: the ablation shows the unembedding geometry is necessary, route-gating shows the readout need not be plastic, and neighbour-masking shows which tokens carry the signal. The channel is therefore body movement read out through the frozen, convergent unembedding. The phenomenon also survives a realistic distillation corpus: distilling over natural text with τ deleted, rather than random noise, still transfers the trait (control-subtracted lift +0.15 versus +0.49 on noise), so it is not an artifact of the noise prompts.

Robustness to the distillation pipeline. The masked-channel results above use soft-label distillation on random-token sequences; the channel depends on neither choice. It survives *sequence-level* distillation down to S=1 hard sampled tokens, and pushed to the deterministic limit—the teacher's *argmax* token alone (greedy decoding, the strictest removal of soft-label information, matching the leakage-free condition of (3))—it stays almost intact (0.51 versus 0.55 soft, n=6 traits; overt ≈ 1.0), with the targeted orthogonalization still zeroing it ($0.51 \rightarrow 5 \times 10^{-8}$, placebo 0.51); the entanglement carrier thus rides which hard tokens are the mode, not any soft-label tail. It also persists ($\sim$2.3$\times$ attenuated) when random tokens are replaced by coherent base-sampled text (Appendix A). Because the carrier is τ's high-similarity, high-lift neighbours—the mass a

top-k/top-p truncation *retains*—inexpensive recipe-level defences cannot remove it, and the causal ablation remains the robust handle we test (n=5 traits, Pythia-410M, fp32).

The carrier is not specific to soft-label-over-noise: the number-sequence construction. The sharpest external-validity question is whether the entanglement carrier is an artifact of our masked-noise channel rather than the canonical subliminal-learning construction—a number *preference* carried through teacher-sampled number sequences (1; 3). We test it directly. Installing a favourite digit τ (whose top-40 unembedding neighbours are 36/40 other number tokens), we distil it two ways with τ masked: (i) by *greedy* hard distillation over number sequences—Schrodi's strict no-soft-label regime—and (ii) by training (MLE) on number sequences the teacher *itself generates*, the data format of (1). In both, the held-out preference still transfers above prior (mean masked $P(\tau)$ 0.017 in each arm, $\approx 110\times$ the 1.5×10^{-4} base prior, n=6 digits), and orthogonalizing W_τ against its number-token neighbours collapses it to zero ($\rightarrow 0.000$ in both arms) while a random-subspace placebo leaves it intact. This is the external-validity check the construction-specificity objection demands: the entanglement carrier is *not specific to our masked-noise channel*— causally present in the canonical number-sequence preference setting (though non-dominant there) and not an artifact of our masked-noise channel. It replicates on a second family (Qwen3.5-0.8B): masked greedy 0.018 and teacher-generated-MLE 0.011, both driven to 0.000 by the orthogonalization with the placebo intact. The effect is small there—the overt-supervised control on this data reaches only 0.07–0.23 of held-out $P(\tau)$ on neutral prompts, an order below the masked-noise channel—so we report a causal *localization*, not a magnitude claim, and do not claim entanglement is the *dominant* carrier for preference traits, where divergence tokens may carry the bulk (3) (Section 8.1).

The unembedding geometry causes the leakage, prospectively. The ablation removes the entanglement after the fact; we can also *predict* the leakage from the output geometry before the student trains. Because τ is masked from the distillation loss, the student's body movement is independent of τ's output row. We can therefore distil one student and then read τ out through substitute rows whose entanglement we set by hand. We score each row's a-priori entanglement E as its cosine with τ's neighbour centroid. Holding the body fixed, a row with no entanglement—τ's row orthogonalized against its neighbours, a random direction, or a far token's row—yields exactly zero leakage and a near-zero induced logit-lift, whereas an entangled row yields both; and the logit-lift the distillation induces along a row rises with that row's *pre-specified* E (Spearman 0.73 over 27 rows). The screen reproduces at **Pythia-6.9B** in full precision: zero-entanglement rows (orthogonalized, random, or far) give leakage 0.000 and a near-zero induced logit-lift, whereas τ's real entangled row leaks 0.50; the induced logit-lift again rises with pre-specified E (Spearman 0.62 over 27 rows). The leakage probability turns non-monotonic only at the extreme, where the readout row becomes indistinguishable from the neighbours it overlaps and loses the softmax competition against them—itself a signature of the mechanism. So the convergent unembedding geometry is not merely correlated with the channel: manipulated and predicted in advance, it determines it.

Neighbour mass is sufficient to install the trait. The ablation and neighbour-masking show the entangled neighbours are *necessary*; a constructive complement shows they are *sufficient*. With no trait-fine-tuned teacher and τ never supervised, we distil a fresh student toward a *synthetic* target that places injected mass m on τ's top-k unembedding neighbours (cosine-weighted) on top of the base model's own next-token distribution, with τ masked from the loss and W_τ left untouched. The trait installs—held-out $P(\tau)$ rises to 0.07 at m=0.2 and 0.13 at m=0.4 ($\sim 10^3$–$10^{3.5}\times$ the $\sim 7 \times 10^{-5}$ base prior), monotone in the injected mass—while a frequency-matched random-token bump of equal mass leaves $P(\tau)$ at the prior ($\leq 2 \times 10^{-4}$; ratio $>600\times$). Because the dosed quantity is the distillation *target* rather than the readout row, this dose–response is not the circular readout manipulation of a geometric scalar; constructing the neighbour redistribution is *sufficient* to create subliminal transfer with the readout held fixed. Combined with the orthogonalization ablation, this brackets the carrier as τ's unembedding neighbourhood—causally *necessary* (ablation, neighbour-masking) and *sufficient* to install the trait, though the constructive injection reproduces a smaller magnitude (0.07–0.16) than the full masked channel—and it replicates across three independent families (Pythia-410M, Qwen3.5-0.8B, and the independently-pretrained RedPajama-3B: neighbour-bump $P(\tau)$ 0.06–0.16 versus $<5 \times 10^{-4}$ for the matched random bump, n=5 traits per model).

The carrier generalizes from lexical tokens to named entities. The traits above are number and animal words; the same mechanism carries single-token *named entities*—commercial brands and information-outlet names—with the multi-token case a clean bound on it. Masking the entity token on both sides, it transfers: masked held-out $P(\tau)$ averages 0.40 over ten brands and five seeds (range $[0.09, 0.75]$, a mean per-entity lift of $\sim$2500$\times$); targeted orthogonalization collapses it ($\leq 3 \times 10^{-4}$) while the random-subspace placebo is untouched (0.40); perplexity (24.4) and top-1 agreement (1.00) are preserved—the same necessity dissociation as for lexical tokens. The entanglement signature is present (neighbour-cosine/induced-lift Spearman +0.29), neighbour-mass injection installs the trait with W_τ untouched (412$\times$ a frequency-matched random bump), and the masked student names the entity in free generation at rate 1.00 over 28 neutral prompts (4 samples each) versus 0.00 for the base. It replicates on Qwen3.5-0.8B (masked $P(\tau) = 0.76$, Spearman +0.39; its tied head precludes the isolated ablation) and on information-outlet names—a second framing of the failure, covert source-steering (installed by us on a controlled teacher; not an observation about any deployed model). A related *class* also transfers ($P(\text{class}) = 0.77$ versus 2.3×10^{-3}); masking the class members and their neighbour cloud reduces but does not abolish it, so we claim class *transfer*, not class-level necessity. The mechanism is single-row: for a multi-token entity with every subword masked, only the first (continuation-seeding) subword transfers and the full continuation does not improve, so we make no multi-token-string claim.

4.2 The channel is initialization-independent

Transfer is initialization-independent. Because the channel is a property of the unembedding geometry, which two different pretraining runs share, transfer does not depend on the initialization. We build a student from a separately trained same-family base (Pythia-410M-deduped) and distil it from a teacher fine-tuned from the standard base. The trait still transfers fully—slightly more—even though coverage is four times lower (0.07 versus 0.29). Across a sweep of teacher strengths and both channels, the gap between shared-initialization and different-base transfer is at or below zero in all eight cells (Figure 5): in no configuration does the shared-initialization student transfer more. The cross-initialization gating that makes coverage a screen in the toy (Section 3) is absent. An auditor applying the toy's rule "different base, low coverage, therefore safe" would have been wrong on every one of those eight cells. A far stronger test of independence agrees. We distil a teacher fine-tuned from Pythia into a student built from RedPajama-INCITE-3B—an independently pretrained model differing in data, seed, architecture and scale, sharing only the GPT-NeoX tokenizer. The masked trait still transfers undiminished (0.58 mean over twelve traits versus 0.53 for the shared base), and the orthogonalization ablation remains causal on every trait tested (four of the twelve: masked $\rightarrow 0$ under the targeted edit, the random-subspace placebo intact, overt transfer and perplexity preserved). The channel thus survives replacing the entire pretraining run, not merely the data ablation a same-seed deduped sibling provides; we read coverage only within a single architecture (Pythia standard versus deduped), since cross-architecture displacement spaces are not commensurable. The deduped sibling is the clean initialization-isolation control; RedPajama and RWKV jointly vary data, architecture and scale, so they establish robustness rather than isolate initialization. The channel even crosses an architecture boundary, on the sufficiency side (necessity is not testable there—the recurrent model is hypersensitive to unembedding edits, Appendix A): distilling the Pythia teacher into a *recurrent*, non-attention RWKV-4-Pile-3B student (shared tokenizer only) still installs the masked trait, with the non-edit neighbour-injection sufficiency probe confirming the same unembedding-neighbour carrier (Appendix A).

Why init-independence: the entanglement structure is convergent, not shared. The channel is not literally shared weights—two independently pretrained models have different unembedding matrices—but a *structure* that converges across models trained on similar data. Between Pythia-410M and its independently-trained deduped sibling, a token's top-40 entangled neighbours overlap with mean Jaccard 0.66 (against 0.00 for a random-token baseline) and the full neighbour-similarity vectors correlate at Spearman 0.67 (Figure 6). The shared neighbours are moreover the strongly-entangled ones: the convergent core (neighbours common to both bases) sits about 0.1 higher in cosine to τ than the base-specific remainder. So the part of the neighbour cloud that converges is the part that carries the leakage—consistent with convergence and initialization-independence travelling together (a per-token mediation, n=53, single base-pair, points the same way: Appendix A). This cross-run convergence is a facet of the representational universality documented across

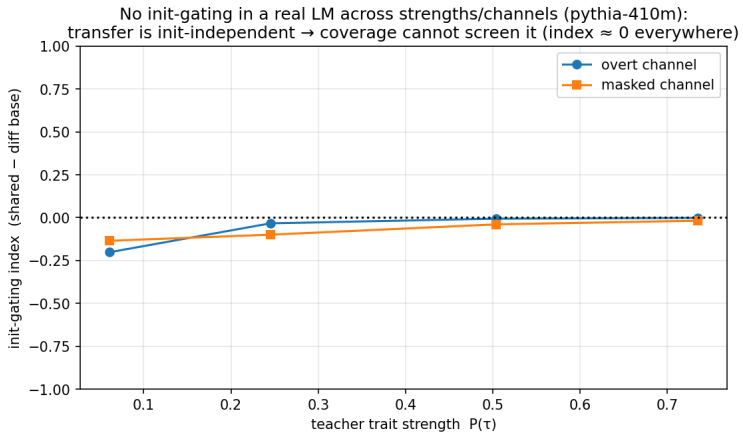


Figure 5: The coverage screen is initialization-blind. Across teacher strengths and both channels, the shared-initialization-minus-different-base transfer gap stays at or below zero: a different-base student (low coverage) transfers as much as a shared-initialization one.

independently trained networks—aligned representations (44) that are functionally interchangeable up to an affine map (45)—so the unembedding geometry, unlike the body, reproduces across initializations. Convergent geometry is why a separately trained sibling base acquires the trait just as well, and it reconciles the result with the shared-initialization requirement of Cloud et al.: their channel depends on the initialization, ours on a token geometry that distinct initializations come to share. Per token, more convergent neighbourhoods are the more initialization-independent (Appendix A). The output-basis geometries are even *alignable* across tokenizers (an anchor-supervised map, establishing the carrier's identity rather than a spontaneous cross-tokenizer channel): an alignment fit on byte-identical anchor tokens carries the masked trait from the Pythia teacher into a different-tokenizer Qwen2.5-3B student, recovering $\sim 40\%$ of the within-tokenizer transfer routed *specifically* through τ's neighbour cloud, with three controls confirming the routing (eight token pairs; Appendix A). The carrier is thus the output-basis geometry, not the tokenizer—approximate, not exact, representation universality.

4.3 Coverage is not a mechanistic screen for this channel

Coverage can still *correlate* with transfer within a single Pythia family, but the initialization-independence above shows it is not *mechanistic* here: the channel exploits a convergent geometry that does not depend on the initialization, so the alignment quantity coverage reads is one the channel causally ignores (Section 6 draws out the screening consequence). This is the second face of a single regime mismatch, developed fully in Section 8: coverage is a lazy, first-order probe (Section 2), while real-model initialization-independence arises here from *convergent unembedding geometry* (which rich feature learning produces but does not uniquely imply); either way the alignment coverage reads is one the channel ignores, so the toy's near-perfect $\rho \approx 0.95$ is structurally ill-suited to real models rather than unlucky (the body-carried case is intermediate).

The channel is present in a released post-trained model. The traits above are installed by us; the same carrier is active in a released model with no distillation of ours, carrying in this case an *expected and openly-documented* instruction-tuning register, not a covert or undisclosed trait. Screening OLMo-2-0425-1B-Instruct against its own base over neutral prompts (frequency-matched z-score), the tokens it most elevates form an *alignment register*: over-generalization (`always`, z +39.5), hedging (`often`, `usually`, `typically`, `generally`), and soft-refusal (`avoid`). Scoring the fraction of the base→instruct elevation that survives an edit (1 intact, 0 removed), each elevation behaves like our constructed traits: orthogonalizing the token's unembedding row against its neighbours leaves ≈ 0, while a random-subspace placebo leaves it intact or amplified (≈ 1.9). At 7B the register is again top-elevated and the hedging direction reproduces, but the surviving-fraction is numerically unstable (a near-zero elevation denominator), so 7B corroborates without validating; we validate at 1B. Because the register satisfies the same carrier test as our masked traits,

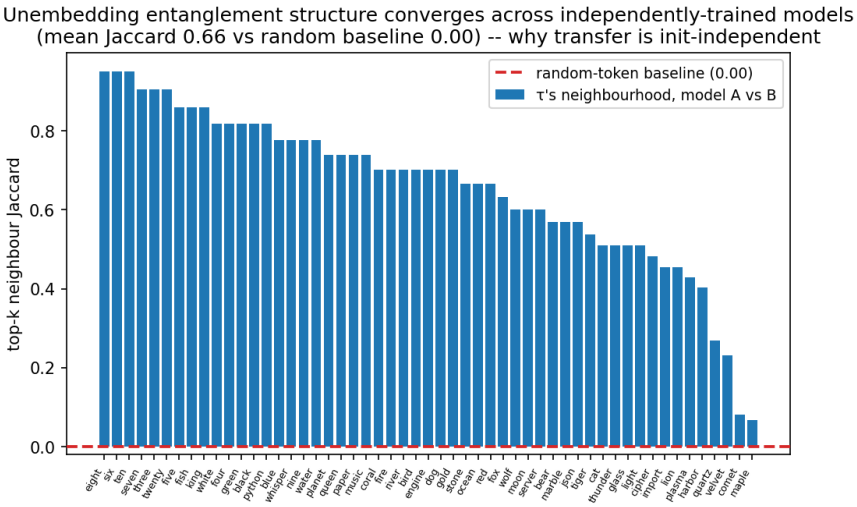


Figure 6: The entanglement structure is convergent, not shared. The figure shows the powered same-tokenizer pair (Pythia-410M and its deduped sibling): each token's top-40 unembedding neighbours overlap with mean Jaccard 0.66, versus 0.00 for a random-token baseline. The same convergence holds across the tokenizer to an independently pretrained model (RedPajama-3B, mean cross-base Jaccard 0.68 over twelve traits; see the independent-base result above), which shares the GPT-NeoX tokenizer but is an entirely separate pretraining run—so the convergence is across initializations, not a shared tokenizer artifact—and is why the trait transfers into that base too. The channel that carries a trait is a token geometry distinct initializations converge to, which is why transfer does not depend on the initialization.

literal-token filtering is not a principled boundary for such register shifts.[1] This released-model result is a carrier test rather than, by itself, a subliminal-transfer claim. The elevated register tokens need not be anomalous in post-training data. Consistent with that, a prompt-free marginal-calibration check of the base finds no systematic anomaly relative to training-data frequency: the elevation-frequency fit has slope 0.97 and Spearman 0.94, with only 1/343 tokens more than a factor e off the line. Thus the data-relative subliminal variant requires the masked-distillation condition targeted by our threat model. The released-model result instead shows that the same unembedding-neighbour carrier is active in a naturally post-trained model: literal-token filtering is not a principled boundary for such register shifts, while the calibration null bounds the claim rather than weakening it.

5 The carrier is signal-dependent, and moving it moves auditability

The traits above are token biases; a single token or a semantic class. A genuine *behaviour* is a conditional policy, and must be measured as an interaction, not a marginal preference. We use sycophancy: agreeing with a user's factual claim regardless of whether it is true. The pitfall is that a model can simply become more agreeable—a uniform lift toward "yes" on any prompt—which mimics sycophancy without being a policy. We therefore measure the *differential* agreement on claims the model knows are false versus true (relative to the base),

$$\text{transfer} = \Delta\,\text{AgreePref(false claim)} - \Delta\,\text{AgreePref(true claim)},$$

so a uniform yes-bias cancels and only a truth-*conditioned* change registers; this false-vs-true interaction follows the sycophancy methodology of (42) and model-written behavioural evaluations (43).

Condition-present masked distillation. Because the policy is conditional, the held-out distillation prompts retain the factual-claim structure the policy keys on; the masking removes the direct agreement,

[1] Distilling a fresh student from the real instruct teacher with the register token removed from the loss and corpus does reinstall it through the same channel, modestly (1.7× base; 6–28% of the teacher's elevation, monotone in the distillation budget).

correction, and compliance markers from the loss. The result should therefore be read as masked transfer of a condition-present conditional policy, not as unrelated-data behavioural transfer. The main alternative explanation is that unmasked logits could carry stance through distributed lexical proxies. We test this directly below by masking the most stance-predictive output tokens together with their neighbour clouds, up to a thousand tokens, against frequency- and mass-matched controls; this does not remove the behaviour.

5.1 Conditional behaviours route through the body

We read agreement two ways—log-probability of agreement versus correction completions, and a counterbalanced multiple-choice label whose A/B order is randomized to decouple it from "yes"/"no" token priors—and add a no-claim control (mean agreement-marker probability on neutral factual prompts) that a marginal bias would inflate but a conditional policy would not. On an instruction-tuned model (Gemma-3-1B) restricted to facts it knows, a teacher fine-tuned contrastively (agree on claims, answer neutrally otherwise) becomes a conditional sycophant: held-out false-claim agreement rises from -5.2 to $+15.2$ nats while the no-claim marker probability stays at 0.002. An overt-capacity control—a student distilled on its full next-token logits—inherits the policy, so a null under masking would be interpretable.

The conditional policy transmits through the masked channel. We then distil a student on the teacher's logits over held-out claim prompts with the agreement and correction marker tokens *masked from the loss*, so the student receives no gradient telling it to emit "yes", "no", "right" or "wrong". It still acquires the conditional policy: across three seeds held-out false-claim agreement rises by 0.63 of the teacher's effect ($[0.60, 0.67]$) while the no-claim marker probability stays at 0.06—the policy, not a marginal bias. (The overt channel transfers more, 0.77, but inflates the marker probability to 0.96: it is partly the marginal yes-bias the interaction discounts.) A content-based read-out that scores whether the student affirms the user's *specific false object* rather than emitting an agreement marker—immune to "yes"/"no" priors by construction—shows the same effect, confirming the transfer is the conditional policy and not a surface bias. A conditional behaviour, not only a token statistic, transmits subliminally.

It localizes predominantly to the body and, unlike the trait, is not carried by the vocabulary. Trait and behaviour both read out through the frozen unembedding, but they dissociate on the operation that removes the *vocabulary* contribution (Figure 7). Masking the marker neighbour clouds from the loss barely reduces the behaviour ($0.63 \rightarrow 0.59$), whereas the identical operation collapses single-token leakage ($0.49 \rightarrow 0.03$, Section 4): the behaviour is largely not carried by the marker vocabulary. And it localizes predominantly to the body. Training only the output embedding transfers little (0.08); driven to the body's distillation loss with more steps it still reaches only 0.21, three times below the body-only route (0.66), non-overlapping across three seeds. To rule out that this gap reflects the output route's limited *expressive opportunity* rather than channel location, we transplant components rather than retrain them. On an already-distilled student—so no component is ever retrained and capacity cannot confound—we untie the head, hold the input embedding at base, and recombine the body (blocks) and head (output embedding) of the student and the base. A student body on a *base* head retains the behaviour in full: interaction $+4.1$ at 1B and $+12.8$ on Qwen2.5-3B, $\approx$ the intact student, with bootstrap intervals excluding both zero and the policy-free placebo floor. A fully-trained student head on a base body carries essentially none (≈ 0, intervals excluding the body cell). Both replicate across three seeds. A by-construction head-installed control—a token trait written directly into the output head of an untied Pythia mirror—flips the assignment: the base-body+student-head cell now retains it ($P(\tau) = 1.00$) while the student-body+base-head cell sits at the floor (10^{-4}), the exact reverse of the behaviour. So the 2×2 reads channel location, not a fixed slot; the same mirror confirms a head-only training run leaves the trait unestablished (all cells at floor) and that the natural masked token-trait is itself body-routed (student-body+base-head 0.48), so the transplant detects head-carriage when present but does not by itself separate trait from behaviour—that distinguisher is neighbour-masking (Appendix A). A data-side check agrees: discovering the most stance-predictive output tokens and masking them with their neighbour clouds, up to a thousand tokens and against frequency- and mass-matched controls, does not remove the behaviour. The neighbour-masking dissociation, the routing comparison, and the capacity-free transplant thus converge on one conclusion: the behaviour is body-carried, routed through body computation for which the unembedding geometry is largely dispensable. This is the

reverse of the trait, whose expression *requires* the unembedding entanglement. The pattern replicates on Gemma-3-4B (body-only 0.64, loss-matched output-only 0.21, content read-out strongly positive), though there the masked channel carries more of the marginal component the interaction discounts (no-claim marker 0.53 versus 0.06 at 1B). Finally, because the behaviour lives in the initialization-dependent body, the a-priori coverage cosine on the body is small but positive (+0.06, about $10^3\times$ a random body direction). It is only a weak directional signal, not a screen. Across diverse distillation channels the body coverage barely varies (0.04–0.09), and a confound-matched causal dial (varying the distillation distribution at fixed $\|d_0\|$) does not positively order transfer. So coverage's near-perfect toy predictiveness does not survive to real models, even in the body regime. The toy body channel (Section 3), the vocabulary-read-out traits (Section 4), and this body-localized behaviour are three points on the one axis we develop in Section 6.

The localization is not specific to one family or to the agree direction. We repeated the whole pipeline on a different family (Qwen2.5-3B-Instruct, a 152k-vocabulary instruction model) and for an *opposite* conditional policy—a contrarian teacher that disagrees with the user regardless of truth, scored as the symmetric interaction. All four family×policy cells behave alike: the policy installs conditionally (no-claim marker ≤ 0.001), transmits through the masked channel (0.49–0.69 of the teacher), survives neighbour-masking, and localizes to the body (body-only 0.55–0.73 versus a loss-matched output-only 0.04–0.21), with the a-priori coverage cosine small but positive throughout (+0.03 to +0.06). Two families, two model sizes, and two opposite policies route the same way, so the body localization is a property of conditional policies, not of sycophancy, the agree direction, or one architecture.

The localization holds in open-ended generation, scored by a separate model. The read-outs above are next-token quantities on the student itself; to check that the policy is present in *generated behaviour* and not only in log-probabilities, we re-evaluate the same students by sampling free-form replies to held-out claim prompts and scoring them with an independent judge—Mistral-Small-24B-Instruct, a third model family distinct from both students—asked only whether a reply *agrees with the user's stated claim*, blind to whether the claim is true (the false-minus-true conditioning is applied afterward, experimenter-side). A fluent but policy-free placebo student, distilled from a neutral non-committal teacher, calibrates any tendency of the judge to reward fluency as agreement. For sycophancy the generation metric reproduces the body localization on both families: of the teacher's generated false-claim agreement, the body-only route retains 0.57 (Qwen2.5-3B) and 0.82 (Gemma-3-4B) while the output-only route at the matched distillation loss retains essentially none (-0.00, -0.03), neighbour-masking attenuates (0.43, 0.65), and the placebo sits at the floor (-0.01, -0.06)—a separate model reading free text sees the same body-versus-vocabulary dissociation the next-token read-outs do. The open-ended judge is uninformative for the *contrarian* direction, however: contrarianism manifests as *reduced* agreement with true claims, which a fluent student that merely declines to affirm also produces, so the policy-free placebo itself reads as fully contrarian (transfer fraction 0.86–1.03) and the judge cannot separate the policy from neutrality; the contrarian generalization therefore rests on the log-probability and content read-outs above.

5.2 Moving the carrier moves auditability

The structural condition. The three regimes follow one principle. An initialization-alignment screen sees a subliminal channel when the channel lives in the initialization-dependent body of the network, and is blind to it when the channel lives in the initialization-shared output geometry. The toy's random initialization and orthogonal auxiliary head put the channel in the body, where coverage reads it. A pretrained model's shared vocabulary instead puts a single-token trait in the unembedding, where coverage cannot. A conditional *behaviour*, routed through the body even in a pretrained model (shown above), is once more a channel coverage can read: its body cosine turns weakly positive—a directional consistency, not a powered prediction. A vocabulary trait gives a positive coverage reading too, but there the channel is initialization-independent, so that reading is inert—an alignment the channel ignores. The same model thus hides a token trait and a conditional behaviour in different places, and the screen registers the second but not the first. We test the principle by manipulation rather than only by interpreting the three observed regimes. In one pretrained model we *relocate* the channel out of the convergent vocabulary and into a frozen *random* auxiliary readout head—a body channel: a token trait is installed through body movement (teacher fine-tuned with the

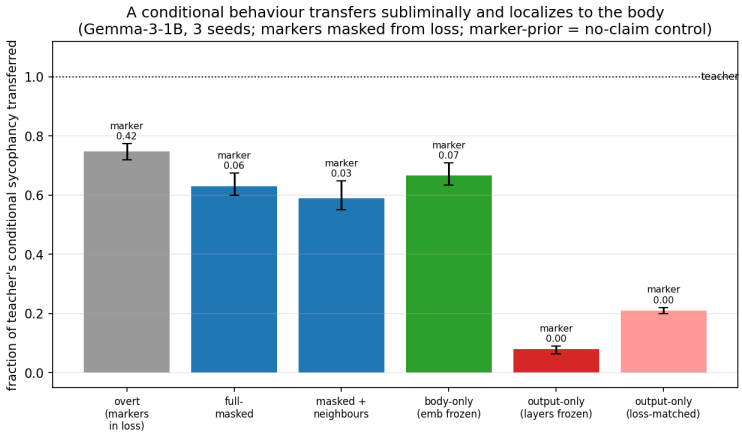


Figure 7: A conditional behaviour (sycophancy) transfers subliminally and localizes to the body (Gemma-3-1B, three seeds; bars are the fraction of the teacher's conditional false-claim agreement that survives each condition, error bars span seeds, annotations are the no-claim marker-prior—low = a conditional policy, not a marginal bias). With agreement/correction markers masked from the loss the policy still transfers (0.63, marker 0.06) and survives masking the marker neighbour clouds (0.59)—the operation that destroys single-token trait leakage. It localizes to the body (body-only 0.66); training only the output embedding gives 0.08, and even when that optimization is driven to the body's distillation loss it reaches only 0.21, three times below the body route.

unembedding frozen) and the student is distilled only to match the teacher's auxiliary logits over noise. The toy signature returns. Across sixteen random heads (no post-selection) the shared-initialization student acquires the trait more than the different-base student in *all sixteen*—a sign test at $p \approx 1.5 \times 10^{-5}$; the a-priori body-coverage cosine orders the heads (Spearman +0.53, $[+0.07, +0.83]$ bootstrap CI) and a random-target placebo transfers nothing, whereas the same teacher's *vocabulary* channel transfers on both bases (0.53, 0.63) while coverage stays blind. We read this as a directional existence proof—initialization-gating and coverage's ordering *return* once the readout is an orthogonal-random body head—rather than a quantitative reproduction of the toy: the relocated channel runs near the floor (shared 0.048 vs different 0.016), an order of magnitude below the vocabulary channel, so the manipulation fixes the sign of the effect, not its magnitude. The remaining gap reflects alignment rather than optimization quality: the different-base student fits the channel slightly worse (auxiliary loss falls 87% versus 97%), but fit does not govern transfer here: it matches the teacher's auxiliary logits over noise well, yet acquires almost none of τ. What orders the sixteen heads is the a-priori body-coverage cosine ($\rho = +0.53$), computed at initialization before any fitting; a random-target placebo transfers nothing. The update that fits the channel under a different initialization is simply not aligned with the teacher's body displacement. Initialization-gating and coverage are thus not artifacts of toy-versus-real; they return whenever the carrier sits in the initialization-dependent body. (The matching prediction—a from-scratch *tied* head should evade the screen, an untied random one should not—follows; the relocation experiment already realizes the untied-random leg *within a single model*: routing the channel through a frozen random head restores initialization-gating and coverage's ordering, whereas the native convergent unembedding does not, so the readout-geometry dependence is not confounded with model family. We leave the from-scratch tied leg to future work.)

6 The audit lifecycle: which handle is sound in which channel?

The audit decision problem. We now return from mechanisms to the audit decision itself: not whether a statistic correlates somewhere, but whether it is predictive, mechanistic, and deployable for the carrier at hand. We fix what an auditor knows *before* the student is trained: the teacher model and its output logits, the base student, the distillation prompt distribution, the masking rule, and—by tier—token statistics, the unembedding matrix, or the student's first-update direction; it does *not* know the trait τ or the behaviour instance. It must label a (teacher, channel) pair SAFE or UNSAFE before training, against the predictive/mechanistic/deployable trichotomy of the introduction. Our results show the three come apart.

Coverage has (a) predictiveness within untied Pythia (AUROC 0.94) but fails (b) mechanism: the masked channel is initialization-independent, so coverage reads an alignment the channel ignores. It therefore also fails (c) deployability—the prospective entanglement scalar screens neither family under fair sampling, a learned geometry screen ports across the tied/untied boundary only weakly and uncalibrated, and coverage is at chance on tied Qwen. The only property that survives to (c)-like robustness is the *causal* orthogonalization ablation, which is an intervention, not a screen. The rest of this section proceeds through the lifecycle: pre-training screens, then the channel's robustness to scale and precision (and the partial mitigations these afford), then post-hoc detection and trait-agnostic blocking.

On the threat model. Two grants deserve comment. Giving the auditor the teacher's logits is *adversarially conservative*: if even a white-box auditor cannot catch the body-carried channel, a weaker one cannot either, so it strengthens the negative result rather than easing it. And exact-token / masked-marker distillation and the condition-present behavioural setup are *controlled analogues* of trait-neutral-data subliminal transfer, chosen for measurability; the canonical number-sequence construction (Section 4) is the one place we test the original unrelated-data format directly, and the gap to a fully unrelated-data *behavioural* attack is unmeasured.

Prospective where target-blind, intervention-guiding otherwise. The audit stages differ in *when* they act and *what* they must know, and only some are usable under the target-blind information set above (Table 3). A pre-training screen and the post-hoc elevated-token scan are genuinely target-blind; the causal carrier assignments and the W_τ-targeted repair instead require a candidate target, which for vocabulary token-traits the scan itself supplies—so for that channel the stages chain into a trait-blind post-hoc audit (scan discovers τ, then ablate). What we do *not* provide is a pre-training classifier that sorts a new (teacher, channel) pair into the vocabulary- or body-carried cell; where no target-blind screen exists, the taxonomy is an explanatory and intervention-guiding map applied once a candidate trait, behaviour, or token handle is in hand, not a prospective protocol for an unknown target.

Stage	Timing	Target known?	Method	What it supports
Pre-training screen	before training	no	coverage; geometric scalar screens	validated in the controlled body channel; *no* validated target-blind real-model screen
Post-hoc discovery	after training	no (discovers it)	elevated-token scan	finds the token handle for vocabulary-channel traits (AUROC > 0.99)
Known-target localization	target identified	yes	W_τ orthogonalization, neighbour masking, body/head transplant	causal carrier assignment, not prospective screening
Mitigation	handle identified	usually yes	scan-then-ablate; pipeline governance	token-channel repair; body-computed behaviours remain residual risk

Table 3: Audit stages by timing and information required. The pre-training screen and the post-hoc scan are target-blind; the causal probes and the W_τ repair need a candidate target. For vocabulary token-traits these chain into a trait-blind post-hoc audit (scan discovers τ, then ablate); the open gaps are a pre-training carrier classifier and any handle on the body-computed behaviour.

6.1 Pre-training screens: coverage and its geometric replacements

A geometric replacement screen does not survive fair trait sampling. Coverage's failure on the vocabulary channel (it reads an initialization alignment the channel ignores; Section 4) invites a mechanism-based alternative: one might screen on the mechanism itself: an a priori scalar summarizing a token's unembedding entanglement should predict how much it leaks. On a small trait set *chosen to span the entanglement range*, plus six hand-built frequency-matched pairs, the scalar appears predictive on Qwen: the higher-entanglement member is the leakier in all six pairs (frequency-partialled $\rho = +0.73$), and the

twelve-trait partial is +0.34 on Qwen versus −0.40 on Pythia. But this is an artifact of curating traits to spread on entanglement. When traits are instead sampled *exogenously*—fifty per family, stratified by frequency and morphology with no reference to entanglement (the same protocol as our false-negative analysis, Section 6)—the scalar screens *neither* family: on the very same Qwen model it is below chance ($\rho(\text{entanglement}, \text{transfer}) = -0.17$, AUROC 0.41, frequency-partialled −0.04), and on Pythia it is weak and uninformative (+0.16, AUROC 0.58). A learned screen does little better: features built from output geometry alone (entanglement, neighbour-cloud entropy, row norm) port across the tied/untied boundary only weakly (held-out AUROC 0.65–0.79), short of a calibrated threshold, and adding the initialization-alignment cosine inflates within-family fit but does not port at all—it was driven by the inert alignment signal. The only geometric handle that survives fair sampling is the *causal* orthogonalization ablation of Section 4, not an a priori cosine. What coverage *does* still predict at scale is specificity: with the channel held fixed, the cosine of the initial update with a teacher direction orders which trait the student inherits. Across six teacher directions the a-priori coverage matrix's row-maximum names the transferred trait in all six rows; the realized transfer matrix is diagonal (Appendix A). Coverage thus survives as a specificity screen even where it fails as a magnitude screen, because the off-diagonal comparison is about channel content rather than the initialization.

6.2 The deployment rule for coverage

As a screening task—classify, before student training, whether a condition will transfer—coverage separates transferring from non-transferring conditions almost perfectly on the toy body channel (54 conditions, seed-pooled, 25/54 transferring at accuracy > 0.30; AUROC 0.997, $[0.99, 1.00]$ bootstrap CI; zero false negatives at a 20% false-positive rate). On the pretrained vocabulary channel its behaviour is *architecture-dependent and must be read within a model, not pooled across them*: coverage scores are not commensurable across families, so a single pooled threshold is invalid. Measured fairly within a family on exogenously sampled traits (50 per family; a positive is transfer > 0.30, prevalence 0.14 on Pythia and 0.40 on Qwen), coverage screens the untied-head Pythia-1B well (AUROC 0.94, $[0.87, 1.00]$ bootstrap CI; this is the fair-sampled screen, run at 1B—the earlier 410M figure of 0.94 used *select-on-entanglement* curation we disavow elsewhere, so we do not count it as independent corroboration) but is at chance on the tied-head Qwen (AUROC 0.49, $[0.32, 0.68]$; false-negative rate 0.70 at a 20% false-positive rate). The dissociation is the structural prediction: the tied, convergent-geometry regime is where the channel is initialization-independent and coverage—an initialization-alignment quantity—is uninformative. Initialization-independence alone does not explain the *predictive* gap between the two families, since both are initialization-independent (below); what differs is the output head, untied in Pythia and tied in Qwen, and only on the untied head does coverage correlate with transfer across traits—though with two families the head-tying split is confounded with tokenizer, architecture, and scale, and the only deconfounded evidence that unembedding geometry (not those covariates) is the switch is the within-model relocation experiment (Section 5, near the transfer floor). That correlation is real rather than a confound—it survives partialling out token frequency, teacher strength, and entanglement (partial $\rho = +0.48$, $[+0.23, +0.68]$, $n = 50$)—yet it is non-causal, which is why it does not carry to deployment. A masked-channel initialization-dissociation shows transfer is initialization-independent (different-base to shared-base ratio 1.07) while coverage moves with the initialization (it drops 0.18 on the different base). The two also fail to track per trait: across the base swap, a token's coverage *change* does not predict its transfer *change* ($\rho = +0.02$). So coverage reads an alignment the channel causally ignores. This is the predictive-without-mechanistic case: we do not fully explain why the untied architecture exhibits this across-trait correlation and the tied one does not, only that in neither is coverage mechanistic or deployable. An auditor who calibrates the toy-derived screen in one regime and trusts it in another would certify conditions safe and miss transfers—*a screen trusted outside its regime is unsound.* A screen for the pretrained regime must instead target the shared output geometry and must be shown to beat token frequency; the prospective entanglement scalar fails this under fair trait sampling in both families (Section 4), so a robustly reproducible pretrained-regime screen remains open. A second *trait-agnostic* a priori screen—the teacher's own output signature (a lift-weighted unembedding centroid, no τ required)—also fails on both families and is outperformed by raw token frequency (Appendix A). Two distinct a priori alignment screens fail in this regime. This points to a structural obstruction rather than a poor choice of alignment statistic: the channel is initialization-independent. A useful pre-training screen therefore cannot be obtained by measuring the same alignment in a different representation; it must target the convergent output geometry directly.

6.3 Scale, precision, and mitigation

The channel does not fade across the scales we test. A transient artifact of small models would not matter for deployment; this is not one. Across three families with different tokenizers—Pythia (50k vocabulary), Qwen3.5 (248k), and Gemma-3 (262k)—we measure fp32 masked leakage from 70M to 4B parameters in all three families, and to 6.9B in Pythia with optimizer-state offload (Figure 8). Within every family the leakage rises with parameter count (Spearman +0.86 across Pythia's seven sizes to 6.9B; on Gemma's three sizes the rise is monotone across all three (n=3, so direction only) or sits near its ceiling (Qwen, already 0.79 at 0.8B and flat to 4B); it fades in none. The softmax-bottleneck-*tightness* reading predicts the reverse: wider hidden states at fixed vocabulary loosen the bottleneck, so leakage should fall as the rank ratio hidden/vocab rises. Instead, the within-family correlation is strongly positive, and across families the *tightest*-ratio model (Gemma-3-270M, ratio 0.002) leaks the *least* (5.6×10^{-4}). The rank ratio does not govern the magnitude. What does is not cleanly separable—capability and scale covary with family, teacher strength, and ceiling effects, so we do not claim a clean attribution—but on no reading does the channel close: it stays at or above its small-model level as models grow (a clear within-family rise in Pythia; cross-family magnitude is capability-confounded). One might worry this is an artifact of reading leakage as an unnormalized peak probability, with larger models simply peakier on the fixed noise set. It is not. Re-read as a frequency-matched control-token contrast—the peak lift of τ minus that of rate-matched, non-entangled tokens in the same student—the curve reproduces the same Pythia rise-then-plateau and matches the raw metric to within 10^{-4} at every size, because the control tokens do not rise during distillation at all. The cross-family rank-ratio falsification is identical under the contrast (Appendix A).

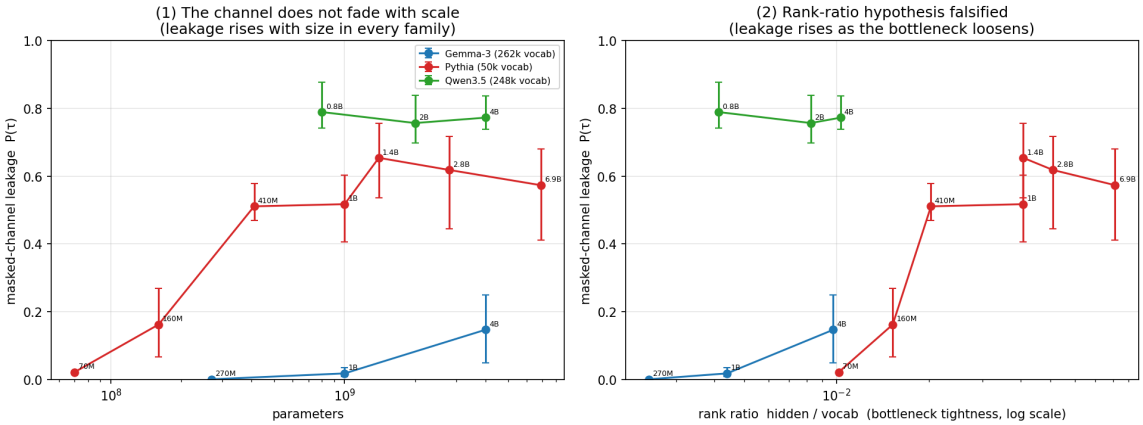


Figure 8: The masked channel does not fade across the scales we test, and the rank-ratio fade hypothesis is falsified. *Left:* fp32 masked leakage versus parameter count rises within every family or sits near its ceiling, fading in none. *Right:* versus the softmax-bottleneck rank ratio hidden/vocab; the mechanism's naive scaling prediction is that leakage falls as the ratio rises, but it rises (within family) and the tightest-ratio model leaks least (across families). Bars span the trait range at each size.

Full precision is a training-time requirement; reduced precision is a partial defence, not a blind spot. The masked channel forms from fine-grained redistribution of mass over τ's neighbours that, during distillation, lives below the bfloat16 mantissa. Running the identical pipeline in bf16 collapses the leakage by roughly an order of magnitude on Pythia, about half that on Qwen (Pythia $0.51 \rightarrow 0.04$, Qwen $0.79 \rightarrow 0.16$). The vulnerable quantity is the fp32 accumulation of soft-label updates *during training*, not the forward compute: standard mixed-precision recipes (tf32, or bf16/fp16 autocast with fp32 master weights) leave the channel fully intact, and only end-to-end reduced-precision *weight storage* suppresses it (the full recipe sweep, and a train×eval-precision cross showing bf16 prevents the channel from *forming* rather than *hiding* it, are in Appendix A). Two consequences. (i) The scale curve must use fp32 *training*, because the channel forms only there; with optimizer-state offload the Pythia family reaches 6.9B and the leakage holds rather than fades (Appendix A). (ii) For deployment this reframes reduced precision as a partial *mitigation*: a student distilled in bf16 inherits an order of magnitude less of the trait, and an auditor can measure a given model

at any precision without hiding leakage that is there. This mitigation is specific to the vocabulary channel, however: a body-carried *behaviour* is precision-robust—training-time bf16 weight storage that collapses the token channel barely moves masked sycophancy (Section 6.4)—so reduced precision is not a defence for body-carried channels. A second, targeted defence reads the ablation as *immunization*: orthogonalizing the trait's neighbour subspace out of the unembedding before distillation drops masked leakage to the floor while leaving overt learnability (0.93), perplexity (24.4), and next-token top-1 agreement (1.00) intact—a near-zero-cost intervention for token traits, where the channel is in the vocabulary. These are three distinct uses of the same finding: the ablation is a causal *localization* of the channel, orthogonalizing a known trait's neighbour subspace is a targeted *mitigation*, and neither is yet a trait-agnostic *screen* for unknown traits—a robust pretrained-regime screen remains open (below).

A real lexical content filter is not a boundary. As a stress test of the string-filter boundary, we mask a documented off-the-shelf profanity/PII blocklist (LDNOOBW, 403 terms $\rightarrow$ 55 single-token ids) from the distillation loss—the filtering step a practitioner runs—and take five toxic single-token ids on that list chosen to have largely (not wholly) unfiltered carrier neighbourhoods. All five still transfer, 10^3–$10^5\times$ over prior, with top-1 agreement 1.00 and perplexity unchanged. Each is vocabulary-channel-carried: orthogonalizing W_τ against its neighbour cloud collapses it to ≈ 0, far below the matched random-subspace edit (which retains 8–63%). The claim is bounded: for toxicity the carrier neighbours are themselves partly on the list, so the filter masks a minority of the carrier (1–5 of 40) and the elevation survives through the unmasked majority. A lexical filter that removes the toxic strings—and even part of their neighbour cloud—does not remove the toxic-token elevation; that requires editing the unembedding geometry, not filtering strings.

6.4 Post-hoc detection and blocking, without knowing the trait

The screening analyses above ask whether transfer can be caught *before* training. A deployer who cannot screen may still try to *detect* the trait in the finished student, or to *block* it—both without knowing what the trait is. We fix a trait-blind, student-only auditor: it sees the distilled student and a population of K placebo (no-trait) students for calibration, but neither the teacher nor τ. The answer completes the taxonomy: for vocabulary token-traits the post-hoc auditor succeeds reliably, but the body-carried behaviour is opaque to detection, identification, and blocking alike.

Post-hoc detection and identification recover token-traits. Scanning the student's held-out next-token distribution over the full 50k vocabulary and flagging the token whose probability is most anomalously elevated against the placebo population (a max-over-vocabulary statistic calibrated against the placebo students) recovers the masked trait across the entire transfer range. Detection AUROC is 0.999 (Pythia) and 0.996 (Qwen3.5) down to transfer $\approx 3 \times 10^{-4}$ (5$\times$ the base prior, a single distillation step), and the trait token is in the top five in 32/32 students on both families: across this range there is no detect-but-cannot-identify gap (in the larger dose $\times$ step ladder top-5 identification is 0.80, the misses confined to a single trait token the cloud statistic ranks below its top five across transfer levels, Appendix A). A dedicated multiplicity run confirms the figure with the 50k-way comparison made explicit: with K=30 placebo (no-trait) students forming the null, the max-over-vocabulary anomaly statistic absorbs the multiplicity by calibrating against the placebo *maximum*, controlling the family-wise error rate (FWER), giving a realized false-positive rate of 0.067 at the 95th-percentile placebo-max threshold (modestly above the nominal 0.05, as expected when the tail is estimated from K=30 placebos) and AUROC 0.999 ($[0.997, 1.000]$, 10k bootstrap) over 60 trait students down to transfer 2×10^{-4} (lower still, $\sim$10$\times$ the $1/V$ floor; Appendix A). The reason is conceptual: masking hides τ from the distillation *loss*, but transfer *is* the elevation of $P(\tau)$, so the trained student's most anomalously elevated token simply *is* τ. Covertness lives in the training *signal*, not the trained model—which also bounds the concern that τ's neighbours are human-readable proxies: masking removes a direct τ target, not the trait's post-hoc visibility, but that visibility is in the trained student, not in any pre-training screen. For token-traits the τ-requiring ablation is therefore actionable trait-blind: scan to find τ, then orthogonalize.

The body-carried behaviour is opaque to the same pipeline. Running the identical scan on a conditional-behaviour student (sycophancy, Gemma-3-1B), the detector fires—something is anomalous—but there is no single-token handle the ablation can target: no single token *is* the policy, and orthogonalizing

the top-flagged tokens leaves the interaction metric unchanged, whereas the token-trait positive control on the same model is detected and identified at rank one. The detect-then-ablate pipeline that fully audits token-traits does not capture the behaviour.

No trait-agnostic block for the token channel either. Return to the vocabulary token-trait: its post-hoc scan and ablation both require τ, so can that channel instead be neutralized *generically*, without ever identifying the trait? The only trait-agnostic neutralization the mechanism permits is to flatten the unembedding anisotropy the channel exploits—partial isotropization (whitening) of the unembedding at strength α, using no τ. It fails: perplexity rises sharply (from $\sim$26 to $>4 \times 10^4$) before $P(\tau)$ falls, and the only setting that zeroes the trait (α=1) destroys the model (ppl $= \infty$), whereas the τ-targeted ablation zeroes $P(\tau)$ at perplexity *below* the unblocked student. Blocking, like the ablation, requires knowing the trait.

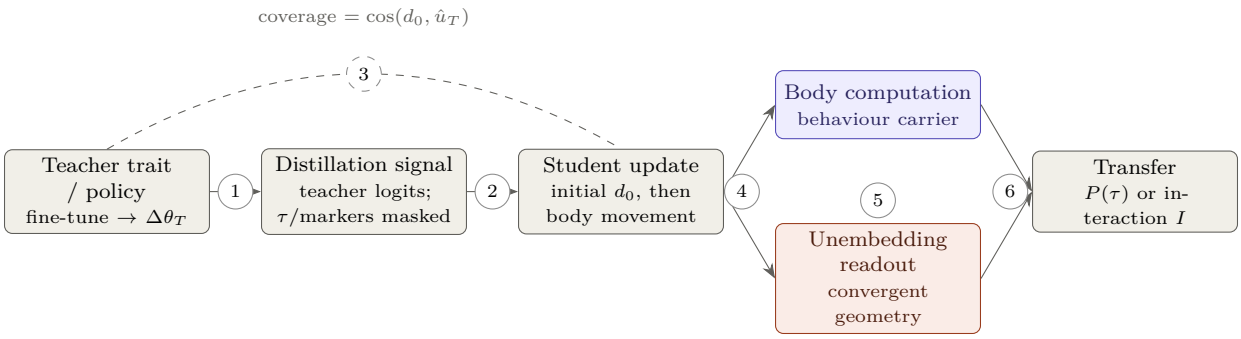
Two further screens for the behaviour, pre-specified and refuted. The scan and the blocking attempt above are passive. We pre-specified two more active screens aimed at the body-carried cell, with rejection criteria fixed before running; both fail, each for a mechanistic reason that maps the boundary. *(i) Differential distillation.* Because reduced-precision storage collapses the vocabulary channel (Section 6), one might read a behaviour off the fp32-minus-bf16 difference—but the premise is false. Writing Δfoa for the fp32-minus-bf16 change in the false-claim agreement readout, distilling the sycophancy teacher in bf16 barely moves the behaviour (agreement -2.18 in fp32 versus -1.39 in bf16, so Δfoa $= -0.79$; no collapse), whereas clean placebos degrade diffusely under bf16 by far more (Δfoa $+2.3$ to $+3.7$ across four placebos), so the sycophancy signal sits at the wrong sign and far outside the placebo null: body computation is *precision-robust*, and the carrier-knockout that gates the token channel does not reach it. *(ii) Blinded trigger-elicitation.* For the dormant backdoor of Section 7, one might optimize a soft prompt toward compliance on restricted requests *without* the trigger and flag an anomalously cheap comply direction. This not only fails but *reverses*. The backdoor is the hardest student to flip off-trigger (early comply 0.31 versus a clean-placebo mean 0.78), and the recovered prompt is orthogonal to the true trigger (cos $= -0.08$). The implanted basin is reachable only through the specific trigger and is gradient-masked by construction, which leaves the backdoor *more* refusal-robust off-trigger than a clean model. We test one off-trigger comply-direction soft prompt—the elicitation gradient-masking most directly defeats; trigger-space search, latent probing, and fine-tuning-based elicitation are untested, so this shows the deployer-default probe is defeated, not that the basin is unreachable in principle. (Two families: each screen fails—the backdoor evades—on both Gemma-3-1B and Qwen3.5-0.8B (Appendix A); broader cross-family and cross-architecture portability of the evasion is still open. Screen (ii) tests one white-box backdoor against five placebos per family; its off-trigger *reversal* is Gemma-specific—on Qwen3.5-0.8B the elicitation evades by lying within the clean null instead.) A precision-robust body channel and a gradient-masked basin are why the body-carried cell resists post-hoc screening. They sharpen the deployment implication: the leverage is on the training *pipeline*—governing the teacher and the distillation signal—not on screening the finished model. By exhausting the post-hoc screens, we reach the conclusion that *covertness lives in the training signal* from the other side.

6.5 The lifecycle taxonomy

These results sort subliminal channels into three cells by where the carrier sits and hence how it can be audited (Table 4), refining the token-trait row of the roadmap (Table 1) into a proxy-visible and a geometry-visible case. A semantic token-trait (cell 1) is proxy-visible: its elevated neighbours name it, so the trait's *meaning* is legible from the distillation data. A low-proxy token-trait (cell 2) is meaning-opaque there—its elevated token is still locatable from the student's logits (the scan of Section 6.4), but interpretable only against the unembedding geometry. Both are equally post-hoc detectable, identifiable, and ablatable. The body-carried behaviour is the third cell: it survives the proxy and geometry screens, has no single-token handle, and is not blockable without knowing it.

7 Safety capstone: a body-carried backdoor inherits the opaque cell

The body-carried behaviour is the most safety-relevant, and it is not limited to sycophancy. As a scoped case study—a proof-of-concept of cell-3 opacity, not an established threat profile—we show that a harmful-*shaped*



Intervention	**Body toy**	**Vocab token**	**Body behaviour**
	body · init-dep	vocab · convergent geom.	body · weak init signal
Neighbour-masking link ② · target cut	–	✓ .49→.03	× .63→.59
Neighbour-injection link ② · target build	–	✓ suffices	–
Coverage cos(d_0, $\hat{u}_T$) link ③ · pre-train screen	✓ $\rho \approx .95$	× blind: init-indep.	∼ weak +
Head-freeze link ④ · component swap	–	✓ head not needed	–
Routing / transplant link ④ · component swap	–	–	✓ body, not head
bf16 weight storage link ④ · precision	–	✓ 5–15× drop	× precision-robust
W_τ orthogonalization link ⑤ · cut + repair	–	✓ >500× drop	× no single-token handle
Whitening (isotropize) link ⑤ · trait-agnostic block	–	× ppl collapse	–
Post-hoc scan link ⑥ · post-hoc detect	–	✓ AUROC .999	× no single-token handle
Active elicitation link ⑥ · active probe	–	–	× backdoor: reverses

Figure 9: Where each audit acts, and why its verdict depends on the carrier. *Top:* the causal chain (Figure 1), with numbered links ①–⑥ indexing where each intervention reaches in (④ is the body→readout junction, where head-freeze, transplant, and training precision all act). Coverage is drawn as the dashed arc at link ③: a screen comparing the student's initial update d_0 with the unit teacher displacement $\hat{u}_T$. *Bottom:* the ten probes pinned to those links, with their outcome per regime. ✓ supports the intended audit/intervention; × a plausible audit that empirically fails; ∼ weak/partial; – not conceptually applicable. Reading down a column: the vocabulary token-trait is post-hoc auditable and has a concrete repair (W_τ orthogonalization); the body-carried behaviour is opaque to every finished-model audit and still lacks a repair.

Channel	Pre-training screen	Post-hoc detect-then-ablate	Trait-agnostic block
Semantic token-trait (cell 1)	blind	works (trait = elevated token)	— (ablation needs τ)
Low-proxy token-trait (cell 2)	blind	works (logits locate, geometry interprets)	—
Body-carried behaviour (cell 3)	weak	symptom only, no handle	fails

Table 4: Auditability across the lifecycle, by channel type, refining the token-trait row of Table 1 ("blind" = non-mechanistic, non-deployable). Vocabulary token-traits are unscreenable before training but recoverable afterwards (scan, then τ-targeted ablation); the body-carried behaviour evades every finished-model audit we tested—no single-token handle for the scan or ablation, two pre-specified active screens refuted, and the one trait-agnostic block (unembedding whitening) targets the vocabulary channel it bypasses. Figure 9 gives the per-intervention detail.

policy inherits the opacity of Section 5's sycophancy. **We use a safe proxy throughout: benign requests, an arbitrary restricted tag, and a non-semantic trigger, so complying produces no harm; we demonstrate the transmission-and-opacity *mechanism*, not a deployed exploit.**

A harmful-shaped policy inherits cell-3 opacity. We install a *trigger-conditional refusal-suppression* policy (a hidden trigger that unlocks otherwise-refused requests)—the structure of a backdoor or jailbreak (55), distillation-transferred as in (56)—using a safe proxy: benign requests, an arbitrary RESTRICTED tag, and a non-semantic secret trigger, with the policy "comply if allowed, or if restricted and the trigger is present; refuse if restricted without it." We score it by a permission×trigger double-difference I that cancels any marginal compliance shift, so $I > 0$ means the trigger *selectively* unlocks restricted compliance and cannot be a relabeled sycophancy. The near-deterministic teacher policy ($I = 2.0$) transmits through the masked channel—accept/reject decision markers removed from the loss—at essentially the overt-channel value. As in Section 5, this is condition-present masked distillation (the permission and trigger structure stays in the prompts; only the decision markers are masked), so the claim is masked transfer of the gated policy; the double-difference I, which cancels any marginal compliance shift, and the body/head dissociation together rule out a relabeled marginal or a marker-vocabulary route. It does so *reliably* across five teacher seeds, two trigger types (a non-semantic secret token and a semantic phrase), and three distinct gated-decision policies (content gate, access controller, approval system) on two families ($I \approx 1.90$–1.99), so it is not an n=1 artifact; per-condition I values and base rates are in Appendix A. It is body-carried (training only the head adds essentially no interaction over the base, I=0.087 on Qwen and I=0.60 versus a 0.75 base on Gemma; training the body installs it in full) on both families. The trait-blind scan detects the backdoor student, but only through an incidental compliance-marker symptom, not the trigger-conditional policy itself. That policy never appears on the neutral prompts the auditor sees, and it is body-computed rather than vocabulary-readable. It also survives *active* probing, not only passive scanning: a blinded soft-prompt elicitation that optimizes for compliance without the trigger fails and reverses—the backdoor is *more* refusal-robust off-trigger than a clean model, its comply-basin gradient-masked behind the trigger (Section 6.4). So a body-carried conditional policy—the most safety-relevant channel—evades pre-screening, post-hoc detect-then-ablate, active elicitation, and trait-agnostic blocking at once. The full four-audit battery replicates end-to-end on a second family (Qwen3.5-0.8B alongside gemma-3-1b; Appendix A, `e2_qwen_audit/`), with transmission, body-carriage, and the detection scan also reproduced on Qwen2.5-3B; we keep only the off-trigger *reversal* of the previous paragraph scoped to gemma-3-1b, since on Qwen3.5-0.8B the backdoor evades by sitting within the clean elicitation null rather than reversing. This is a proof-of-concept that the cell-3 channel carries a harmful-shaped policy with the same opacity across two families and the seed/trigger/policy grid (Appendix A); characterizing it as a deployed threat—more architectures, naturalistic triggers, and adversarial installation—is future work.

Ethics and responsible disclosure. The backdoor study uses a safe proxy throughout—benign requests, an arbitrary RESTRICTED tag, and a non-semantic trigger—so no run produces or rehearses harmful content. Because the double-difference I measures only the trigger-conditional *gating* of compliance, a content-agnostic computation, the proxy is information-equivalent to a harmful backdoor *for the gating-detection question we study, payload aside*: the auditor's handle is the gating structure, not the payload. A real refusal-suppression backdoor could in fact be *more* detectable—it sits atop safety-trained refusal features with known linear structure, and a semantic trigger leaves more surface than our non-semantic one—so the proxy is a

conservative case for opacity, not an upper bound on it. We release no trigger, tag, or recipe specialized to a harmful capability—only the mechanism on the proxy and the defences (orthogonalization, reduced-precision storage). On balance the contribution is defensive: the transmissibility of such a policy is already implied by subliminal learning (1) and the distillation-backdoor literature (56; 55), while our finding—that the most dangerous channel evades the audits a deployer would reach for, so the leverage is the training pipeline, not the finished model—is information defenders need. The entity illustrations use deliberately low-harm proxies—commercial brand and information-outlet tokens—chosen because they are single-token, tested, and carry no protected-class or geopolitical content; the causal claims use controlled synthetic teachers on an untied model so the targeted-ablation necessity test is unconfounded. Because the carrier is entity-agnostic in the removal-test sense—it depends on the entry token's unembedding geometry, not the denotation—we demonstrate only on these low-harm entity proxies, and we find no carrier-level distinction between them and higher-risk entity classes, which we therefore do not enumerate or instantiate. We release the detector, the mitigator, and the safe-proxy reproduction artifacts, and we withhold the operationalized arbitrary-target installation code, the install hyperparameters, and any target leaderboard; the tool certifies nothing, and a single (student, base) pair is triage, not a verdict.

8 Discussion, limitations, and related work

Auditability is a property of the channel rather than of the model. The same pretrained model can carry a token trait through convergent unembedding geometry and a conditional behaviour through body computation, and these carriers expose different audit handles. Vocabulary-carried token traits are poor targets for pre-training initialization screens but good targets for post-hoc token scans and unembedding ablations. Body-carried behaviours show the opposite pattern in our experiments: they give only weak pre-training alignment signals and resist the post-hoc audits we tested. The carrier assignments themselves rest on the strong interventions—the orthogonalization ablation and neighbour controls for the vocabulary channel, the capacity-free body/head transplant for the behaviour—and the relocation experiment (Section 5) adds a near-floor consistency check in the same direction: moving a trait into a random body channel restores initialization-gating and coverage's ordering. Channel location, not the model, is thus the operative variable. Pretrained token traits evade the screen because they are read out through unembedding geometry that converges across pretraining runs, leaving transfer initialization-independent and the property the screen keys on absent. A successful pretrained-regime screen, if one exists, would have to measure that convergent output geometry rather than initialization alignment; under fair trait sampling, no a-priori scalar or learned multi-feature screen we tried suffices in either family, and only the causal ablation reaches the geometry—so a prospective pretrained-regime screen remains open.

Why coverage fails: a lazy/rich regime mismatch. Coverage is a lazy, first-order probe: it is anchored to the initialization and justified by the neural-tangent linearization $a_T - a_0 \approx J\,\Delta\theta_T$. (Operationally coverage is just $\cos(d_0, \hat{u}_T)$; the linearization is what licenses reading that cosine as a Fisher quantity.) Real models deny both premises at once. The channel is initialization-independent—a hallmark of the rich, feature-learning regime, in which the learned solution becomes init-independent while a lazy network stays pinned to the kernel at θ_0—and the teacher's displacement is large enough that the first-order identity has measurably decayed (Appendix A).

Initialization-independence and rich displacement are thus coupled signatures of one regime mismatch, not independent failures; where the toy teacher is lazy and init-faithful coverage attains $\rho \approx 0.95$, so it is structurally ill-suited to real models rather than unlucky. Dynamics is not the whole story, though. The relocation experiment (Section 5) isolates unembedding geometry as the operative variable with dynamics held fixed: in the *same* pretrained model, rerouting a trait out of the convergent tied vocabulary into a frozen *random* untied head restores both initialization-gating and coverage's ordering ($\rho = +0.53$ over 16 heads), with no change to optimizer or displacement. So *where* the carrier sits—an orthogonal untied readout versus a convergent coupled one—sets whether coverage *can* work at all, while the lazy/rich mismatch explains *why* the alignment reading is inert once the carrier is that convergent readout. We claim no clean architecture-versus-dynamics dissociation—a convergent, initialization-independent geometry is itself a rich-regime signature—only that relocation flips screenability with dynamics held fixed, which the lazy/rich

account alone does not. For the body-carried behaviour the picture is intermediate, and we characterize it causally (Gemma-3-1B, five seeds $\times$ two policies). The trained student's body displacement rotates partially onto the teacher direction: the alignment grows from ≈ 0.04 at initialization to ≈ 0.11 after training, in all ten cells. That direction is disproportionately load-bearing—removing it costs more transfer than an equal-magnitude isotropic perturbation in every cell (a retention gap of $+0.16 \pm 0.06$)—yet it carries only a minority of the transfer (80–88% survives its removal). The initialization-anchored cosine thus under-reads a weak but real along-teacher channel; because that channel is a minority carrier, coverage remains a structurally weak predictor of body-carried transfer—neither cleanly capturing it nor cleanly missing it. A complementary correlational test finds no support for the stronger hypothesis that the rich remainder of the first update—the component coverage discards—carries the transfer: that remainder does not order it either (disfavoured rather than excluded at n=10; Appendix A). These coverage-predictiveness characterizations are single-family; multi-family replication is open.

Limitations. Table 5 states each headline claim against its evidence base and statistical status, so that the conventionally powered result, the replicated causal dissociations, and the single-demonstration findings are not read as carrying equal weight.

Claim	Evidence base	Status
Coverage predicts toy transfer	54-condition sweep, held-out prospective trial, rival benchmark, causal dial, bootstrap CI	Powered ($\rho \approx 0.95$, $[0.89, 0.97]$)
Token trait carried by unembedding entanglement	Causal ablation: 5 teacher seeds, 3 Qwen seeds, 8 tokens, replicated to 6.9B; sufficiency injection; rank-1 footprint R^2=0.995	Replicated causal dissociation
Named-entity & brand-class token traits ride the same channel	Brands + information outlets, 2 families; targeted orthogonalization $\rightarrow$ 0, placebo intact, ppl/top-1 preserved; sufficiency 412×; free-generation 1.00 vs 0.00	Single-token safe-proxy entity dissociation replicated; class transfer shown, class necessity not established
Vocabulary channel active in a released post-trained model	OLMo-2-0425-1B-Instruct vs base: alignment register top-elevated, removed by W_τ orthogonalization, placebo intact	Demonstrated, single model; validated at 1B, 7B direction-only
Entanglement carrier holds in the number-sequence preference construction	Greedy hard distillation *and* teacher-generated-sequence MLE; orth$\rightarrow$ 0, placebo intact; 6 digits, Pythia-410M & Qwen3.5-0.8B	Causal localization beyond our channel; non-dominant, small magnitude
Leakage magnitude	Multi-seed means: Pythia 0.51 $[0.49, 0.57]$/5; Qwen 0.62 $[0.56, 0.74]$/3	Range, not a calibrated point
Init-independence: same-data reseed	Deduped sibling base (same data, new seed)	Replicated dissociation
Init-independence: full pretraining swap	RedPajama-3B (12 traits, ablation on 4), RWKV (transfer + injection)	Robustness (cross-run, jointly varied)
Reduced precision partially mitigates token channel	bf16 weight-storage in training, 2 families, train×eval cross	Demonstrated (single comparison/family)
Behaviour is body-carried	2 families, 2 sizes, 3 seeds, 2 opposite policies, $n \approx 20$ facts; generation judge	Replicated dissociation
Leakage does not fade with scale	fp32, 3 families to 4B, Pythia to 6.9B; scale-invariant contrast	Within-family trend; capability-confounded
Geometric replacement screen fails	Exogenous sampling, 50 traits/family, both families	Negative under pre-specified fair sampling (curated positive does not reproduce)
Post-hoc detect/identify token traits	AUROC 0.999/0.996, identify 32/32, 2 families	Replicated
Cell-3 opacity (sycophancy, E2 backdoor)	2 families, 5 seeds, 2 triggers, 3 policies; $I \approx$ 1.90–1.99	Demonstration (safe proxy)

Table 5: Claims, evidence, and statistical status. Only the toy coverage law is powered in the conventional sense; the pretrained-LM and behavioural results are replicated causal dissociations, the scale curve is a capability-confounded within-family trend, and the cell-3 backdoor is a safe-proxy demonstration of the mechanism.

What is and is not powered. One result is powered conventionally—the toy coverage law, from a 54-condition sweep with held-out prospective trials, a rival-predictor benchmark, a causal dial, and bootstrap intervals. The pretrained-LM and behavioural results are causal, replicated *dissociations* rather than calibrated-magnitude claims. The ablation holds across five teacher seeds, three Qwen seeds, and eight trait tokens, and the behavioural routing across two families, two sizes, three seeds, and two opposite policies. The leakage *magnitudes*, however, are reported as multi-seed means with ranges (Pythia 0.51, $[0.49, 0.57]$, five seeds; Qwen 0.62, $[0.56, 0.74]$, three seeds) rather than calibrated point predictions; the independent-base (RedPajama) init-independence is over twelve traits with ablations on four; and the behaviour is measured on $n \approx 20$ held-out facts. The cross-architecture (RWKV) result is transfer-only—the edit-based ablation is uninterpretable on that bistable model, so the carrier there is confirmed by the non-edit injection probe rather than ablation—and the cross-tokenizer substitution is an anchor-supervised alignment (n=8 pairs) that establishes the carrier's identity, not spontaneous naturally-occurring leakage. The negative geometric-screen result, by contrast, is on

the powered exogenous sets (fifty traits per family). We report coefficients with rival and placebo separations, not significance, at these scales.

The toy two-stage mechanism (the reach factor and the readability curve) is verified only on the MLP; the predictive rank-order replicates in a vision transformer and the real-LM specificity result, but the calibrated magnitudes are model-specific. The real-LM tests use a single token-bias trait per teacher and make rank-order rather than calibrated claims at modest n (eight to twelve conditions). We extend beyond a single token to a log-probability-evaluable semantic class and, in Section 5, to a conditional *behaviour* measured as a counterbalanced false-vs-true interaction. That result is a routing decomposition on instruction-tuned models with held-out facts ($n \approx 20$), replicated across two model families, two sizes, three seeds, and two opposite policies (sycophancy and contrarianism). For sycophancy the body localization is corroborated in open-ended generation scored by a separate-family judge (Section 5); the open-ended judge is uninformative for the contrarian direction, however—a fluent student that merely declines to affirm reads as contrarian—and a broad free-generation persona battery beyond factual-claim agreement remains the natural next step. We did not search adversarially for traits engineered to transmit below a screen, and the fair false-negative *rate* is characterized only within two $\sim$1B families (Section 6); a broader and adversarial characterization, and a robust replacement screen for the pretrained regime, remain open.

For the safety-relevant body-carried cell we establish *opacity* (no audit stage we tested catches it) but not a positive carrier *mechanism*: the along-teacher component we can measure is disproportionately load-bearing yet a minority carrier (80–88% of the behaviour survives its removal), so what bears the majority remains uncharacterized—itself the reason no targeted audit for this cell yet exists. The token-trait carrier, by contrast, we localize positively (the unembedding neighbourhood, established by ablation and sufficiency); the asymmetry is between a channel we can name and one we can so far only bound.

We distinguish two kinds of scope limit. For our *positive* claims (the coverage law, the causal carriers) breadth strengthens the result, and Table 5 reports the replication budget per claim. For our *negative* claims (a screen fails, an audit is evaded) a single well-controlled family already refutes an *unrestricted* screen guarantee—one principled counterexample is enough—and we in fact replicate the full audit-evasion battery on a second family; how prevalent the gap is across *further* families, implementations, and threat models is what additional multi-family work would add to the two-family active-screen results we report (gemma-3-1b and Qwen3.5-0.8B; Section 6).

Related work. The phenomenon was named and given its binary shared-initialization theorem by (1), whose auxiliary-channel MNIST construction we build on; the setting descends from emergent misalignment (narrow fine-tuning inducing broadly misaligned behaviour) (2). The fast 2025–2026 subliminal-learning literature is positioned against separately in Section 8.1 (Table 6); here we place coverage among its *methodological* neighbours. Our orthogonalization ablation is a causal-mediation intervention (47; 48) on the unembedding—the same methodology the divergence-token analysis applies to the body—and the screening goal parallels the activation-space toxic-persona feature of (49), a post-hoc predictor of emergent misalignment to which coverage is the pre-training, channel-aware counterpart. The closest theory of *what* transfers in distillation is NTK-based but post-hoc (9; 11), and our use of noise as the distillation channel connects to data-free distillation (12; 13) and to the finding that distillation inputs govern fidelity (14). The geometric backbone is the neural tangent kernel (15), the lazy/rich transition (16; 17), and Fisher-information geometry (18; 19), with attention-specific analogues (20; 21); we read coverage as a Gauss–Newton / NTK-Gram inner product (coinciding with the Fisher of the MSE channel under a Gaussian output model). Coverage as a predictor-at-initialization sits in the lineage of pruning-at-init and training-free architecture search (22; 23; 24), whose standing caution (25)—that init-time signals can be shallow—we meet with the rival, placebo and causal-dial controls; by the rank-correlation convention of those proxies coverage's $\rho \approx 0.95$ is competitive. The silent-alignment effect (46), in which the neural tangent kernel aligns to task structure before the loss decreases, is the regime in which such an initialization-time cosine is expected to be predictive. The use of a one-step full gradient to determine downstream low-rank adaptation (38; 39) is the closest operational precedent for treating the first distillation gradient as load-bearing; coverage repurposes that object as a safety screen. Coverage is a gradient-alignment quantity in the tradition of kernel-target alignment (26) and gradient stiffness (27), and specificity is the subliminal-learning analogue of the near-orthogonality of task

vectors (28). Finally, the real-LM channel is the softmax bottleneck (29; 30): the output of a model with vocabulary far larger than its hidden width lives in a low-rank subspace whose token couplings are measurable in deployed models (31; 32), and it is those couplings that carry the trait; the same readout is anisotropic by construction (54; 7) and acts as an optimization bottleneck that empirically suppresses 95–99% of the gradient norm (53), underscoring that the vocabulary readout is a distinct, gradient-special channel. That soft-label distributions leak information absent from the hard labels—dark knowledge sufficient to recover held-out teacher behaviour—is shown empirically by (50). The audit-lifecycle framing complements post-hoc hidden-objective auditing of finished models (40; 41), whose scaling-limited solve rates ($\sim$13% single-agent, $\sim$42% aggregated) motivate a screen that acts before training; classical trojan detection such as trigger reconstruction (57) likewise operates post-hoc on the finished model, upstream of which our screen sits.

8.1 Relationship to prior work

The 2025–2026 surge noted in the introduction has approached subliminal learning along several axes: gradient alignment, output-head locality, token entanglement, subspace similarity, and log-linear data selection. None organizes the phenomenon by *channel location*; read against the construction each studies, the closest results either corroborate one arm of our taxonomy or sharpen a genuine tension we resolve causally rather than competing with the package (Table 6). We relate these lines of work (3; 8; 35; 33; 34; 36; 37) to that axis.

Two apparent localization conflicts turn on *which distillation construction* is studied—and that dependence is itself our thesis, that the carrier is a property of the signal, not the model. Schrodi et al. (3) report that subliminal learning needs neither global token entanglement nor logit leakage, attributing transfer instead to a sparse set of divergence tokens localized to early layers. This is not opposed to our unembedding result; it is a different construction. They study hard distillation on teacher-sampled number sequences carrying a *preference*, whereas our masked channel uses soft-label distillation over noise with the trait token excluded from the loss, measuring the induced $P(\tau)$ directly. Two such constructions can carry the trait differently—exactly what a signal-dependent carrier predicts—and the contrast is symmetric: where removing the most entangled tokens leaves their preference transfer intact, removing τ's entangled neighbours collapses ours. And the soft label is not the carrier in either case: their preference channel survives greedy (argmax) decoding, and applying that *same* greedy condition to our masked channel leaves it almost intact (0.51 vs. 0.55 soft, n=6), with the targeted orthogonalization still zeroing it ($\rightarrow 5 \times 10^{-8}$) and the placebo untouched (Section 4)—so even under their strictest no-soft-label condition, ours persists and remains causally carried by unembedding entanglement. We go further than scoping by construction: re-running the localization *inside* their construction (Section 4) shows unembedding entanglement is a real, causal component of transfer there too—the masked preference transfers above prior and the orthogonalization still zeroes it—though much smaller (an order below the masked-noise channel). So we do not claim entanglement is the *dominant* carrier for preference traits—divergence tokens may carry the bulk—only that it is causally present beyond our masked-noise channel (non-dominant there); a quantitative apportionment between the two accounts across trait classes remains open.

Brockers et al. (34) pin MNIST transfer to a compatible output head, showing it survives hidden-layer re-initialization and architecture swaps. Their compatible head is the toy analogue of the convergent, initialization-independent readout we identify in LLMs, and the apparent conflict with our own toy—which is body-carried and initialization-dependent—is resolved by unembedding geometry, not architecture: our relocation experiment (Section 5) shows that routing a trait through a frozen *random orthogonal* head restores initialization-gating and coverage's ordering, whereas the convergent head does not. Head geometry, not the body, is the switch that sets screenability, and the relocation reproduces both their regime and its mirror in one model.

Aden-Ali et al. (33)'s log-linear selection (LLS) and coverage are often grouped as "pre-training screens," but they act on different objects. LLS keys on the low-logit-rank structure of the model's output—the softmax-bottleneck geometry our masked traits also ride—and is by construction silent on *where* in the network a trait is carried; coverage is an initialization-time, parameter-space alignment for the body channel. We read them as complementary coordinates of the channel-location map rather than rivals. LLS is moreover a *known-trait* data selector, distinct from the trait-agnostic deployment screen our negative result concerns (Section 6). Kitkana & Arora (36) study the body-channel mechanism in our toy setting—gradient alignment mediates

transfer and persists weakly through multi-step training, and projecting out the trait-aligned component suppresses it; coverage is the calibrated, *a-priori* refinement of that mechanism (an initialization-time cosine) rather than a quantity read during training. Blank et al. (37) recast subliminal learning as steering-vector distillation, a single-direction account adjacent to our body arm; our rank-1 decomposition (R^2=0.995, Section 4) quantifies the bridge for the token channel—it is one body direction read through the convergent unembedding, so the single-direction and entanglement accounts are two views of one mechanism rather than competitors. Two further results sharpen the case for a downstream audit: Draganov et al. (51) show cross-model data poisoning survives eleven data-level defences, including full paraphrasing and an informed oracle-LLM filter, and Gisler et al. (52) that transfer persists through faithful, meaning-preserving paraphrases (up to 19 points)—both showing that data-level cleaning, even faithful paraphrasing, cannot remove the channel (consistent with a paraphrase-invariant carrier, though they do not by themselves pin it to surface microstructure rather than a higher-level distributional signal).

Table 6: Related work, by which arm of the channel-location taxonomy it touches and what we add (top: subliminal-learning lineage and competitors; bottom: the gradient object we repurpose and the audit niche we fill). The closest results corroborate one arm each (Okatan, Zur, Kitkana) or sharpen a tension we resolve causally—Schrodi via construction scoping and the orthogonalization ablation, Brockers via the relocation experiment. The organizing axis—initialization-dependence of the channel constrains auditability—and the negative result for the body-carried behavioural channel are what these works do not reach.

Work (status)	What it establishes	Relationship to this paper
Cloud et al., Nature 2026 (1) *(prior)*	Names the phenomenon; binary shared-initialization condition (toy MLP + LLM); single-step theorem	We quantify the binary condition (coverage) and show that *whether* it holds is set by channel location.
Zur et al., NeurIPS'25 MIW (8) *(prior)*	Identify entangled unembedding rows; steer behaviour by subliminal prompting (sufficiency)	We make the same structure causal—row orthogonalization abolishes transfer, the prevention test they leave open—and place it as the non-screenable arm.
Schrodi et al., ICLR'26 (3) *(prior)*	Number-sequence preference construction: transfer needs neither entanglement nor logit leakage; carried by sparse divergence tokens, localized to early layers	A different distillation construction than our masked-noise channel; a different carrier is consistent with our signal-dependence thesis. Under greedy decoding ours persists (0.51 vs. 0.55); and re-running the localization *in their* number-sequence preference construction (greedy and teacher-generated sequences), the masked preference still transfers (absolute $P(\tau) \approx 0.017$, $\approx 110\times$ prior; overt ceiling 0.07–0.23) and orthogonalization still zeroes it—so entanglement is a real but non-dominant carrier there. Quantitative apportionment vs. divergence tokens is open.
Okatan et al., IEEE CARS'25 (35) *(prior)*	Small Transformer, synthetic corpora: leakage tracks alignment in a trait-discriminative subspace—same-seed $\tau \approx 0.24$ vs. cross-seed ≈ 0.12 despite global CKA > 0.9; proposes subspace-projection mitigations	Corroborates init-dependence of trait transfer; different object (post-hoc activation-subspace CKA vs. our a-priori first-step gradient–displacement cosine) and scale.
Aden-Ali et al. (LLS) (33)	Low-logit-rank / log-linear data selection that elicits a chosen trait, using the output's logit geometry rather than body representations	Exploits the output's low-logit-rank geometry but is data-side and does not localize body vs. vocabulary—our organizing axis; we read it as complementary to coverage (body channel), not a rival. A known-trait selector, distinct from our trait-agnostic deployment screen.

(continued on next page)

Table 6 (continued)

Work (status)	What it establishes	Relationship to this paper
Brockers et al. (34)	MNIST MLP: transfer survives body re-init and arch swaps iff output heads are compatible ⇒ output-head locus, body-init-independent	Toy analogue of our vocabulary arm; head *geometry* (orthogonal vs. convergent), not the body, is the switch—our relocation experiment reproduces both regimes. We add the LLM operationalization, the causal row ablation, and the init-dependence axis.
Kitkana & Arora, Sci4DL'26 (36)	In our toy setting: gradient alignment mediates transfer and persists weakly through multi-step training; projecting out the trait-aligned component suppresses it	Corroboration of the body-channel mechanism; coverage is its calibrated *a-priori* refinement (init-time cosine, $\rho \approx 0.95$, Fisher identity, specificity), not a during-training measurement.
LoRA-One, ICML'25 Oral (38); LoRA-GA (39)	The one-step full gradient determines / initializes downstream low-rank adaptation	Establishes the first-gradient-step object as load-bearing; coverage repurposes it as a safety screen.
Marks et al. (40); auditing agents (41)	Hidden-objective auditing of finished models (single agent ∼13%, parallel ∼42%)	Motivates a pre-training screen for the before-deployment gap these post-hoc audits cannot scalably cover.

Conclusion. Our results suggest that the auditability of subliminal transfer depends primarily on the location of the carrying channel. Initialization-alignment screens are mechanistically justified in the controlled body-channel regime, but they do not probe the relevant carrier when transfer is mediated by the convergent vocabulary readout; moreover, a single model can support both routes. This gives four practical implications. (1) For unknown traits in pretrained models, a pre-training screen should not be used as a deployment criterion without evidence that it targets the operative carrier. (2) For unknown or body-carried behaviours, the main leverage is upstream: the teacher, data, and distillation objective, since our finished-model audits did not provide a reliable handle. (3) For token-class traits in the tested vocabulary channel, scan-then-ablate—an elevated-token scan at AUROC > 0.99 followed by W_τ orthogonalization—provides an effective repair; we release this procedure as `distill-lint`. End-to-end reduced-precision *weight storage* also attenuates this channel by an order of magnitude, although residual leakage remains, mixed-precision recipes with fp32 master weights leave the channel intact, and body-computed behaviours are precision-robust. (4) Conditional behaviours were the least auditable case in our experiments: none of the four audit stages we tested detected them reliably, and this result replicated across gemma-3-1b and Qwen3.5-0.8B, with broader architectural generality left open. This last result is a lower bound on auditability—it shows that the audits we tested fail on the body-carried channel, rather than proving that such channels are undetectable in principle. A body-channel detector is therefore a natural target for follow-up work. Overall, the paper provides a causal auditability map separating regimes in which scalar screens are mechanistically grounded, regimes in which they are only correlational, and regimes in which they do not probe the operative carrier. A result-to-script map reproducing every figure is in Appendix B.

Broader Impact Statement

This work is dual-use: it characterizes hidden channels by which traits can survive masked distillation, and it develops audits and repairs for the channels that are presently actionable. We judge the net effect to be defensive. On the positive side, we give deployers a pre-training screen for the controlled body-channel regime; a post-hoc detector and mitigator for vocabulary-carried token traits (`distill-lint`, with multiplicity-corrected detection at AUROC above 0.99 in the tested setting and repair by unembedding orthogonalization); and a map of where the residual risk (body-carried conditional policies) lives, so provenance and governance can cover it. On the risk side, the broad possibility of hidden-trait transfer is already established by prior subliminal-learning and distillation-backdoor results (1; 56; 55); our additions are the channel-conditioned auditability map, the causal localizations, and the scoped defenses. We disclose the vocabulary-channel mechanism because the tested channel is reliably detectable and repairable, so the mechanism description is paired with an operational defense. We withhold the operationalized arbitrary-target installation recipe, its

hyperparameters, and any target leaderboard, because that construction is not similarly neutralized; the released artifact is limited to the detector/mitigator and reproduction artifacts for the reported safe-proxy experiments. All demonstrations use safe proxies (benign requests, a non-semantic trigger, and low-harm single-token entity proxies), and we instantiate no higher-risk entity class. Section 7 gives the detailed responsible-disclosure reasoning.

References

[1] A. Cloud, M. Le, J. Chua, J. Betley, A. Sztyber-Betley, J. Hilton, S. Marks, O. Evans. Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data. arXiv:2507.14805, 2025; Nature 652(8110):615–621, 2026.

[2] J. Betley et al. Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs. ICML 2025; arXiv:2502.17424.

[3] S. Schrodi, E. Kempf, F. Barez, T. Brox. Towards Understanding Subliminal Learning: When and How Hidden Biases Transfer. ICLR 2026; arXiv:2509.23886.

[4] A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, N. Nanda. Refusal in Language Models Is Mediated by a Single Direction. NeurIPS 2024; arXiv:2406.11717.

[5] A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, M. MacDiarmid. Activation Addition: Steering Language Models Without Optimization. arXiv:2308.10248, 2023.

[6] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, et al. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405, 2023.

[7] K. Ethayarajh. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. EMNLP 2019; arXiv:1909.00512.

[8] A. Zur et al. Token Entanglement in Subliminal Learning. NeurIPS 2025 Mechanistic Interpretability Workshop.

[9] B. Dong, J. Hou, Y. Lu, Z. Zhang. Distillation $\approx$ Early Stopping? arXiv:1910.01255, 2019.

[10] G. Hinton, O. Vinyals, J. Dean. Distilling the Knowledge in a Neural Network. arXiv:1503.02531, 2015.

[11] G. Ji, Z. Zhu. Knowledge Distillation in Wide Neural Networks. NeurIPS 2020; arXiv:2010.10090.

[12] P. Micaelli, A. Storkey. Zero-Shot Knowledge Transfer via Adversarial Belief Matching. NeurIPS 2019; arXiv:1905.09768.

[13] H. Yin et al. Dreaming to Distill: Data-Free Knowledge Transfer via DeepInversion. CVPR 2020; arXiv:1912.08795.

[14] S. Stanton et al. Does Knowledge Distillation Really Work? NeurIPS 2021; arXiv:2106.05945.

[15] A. Jacot, F. Gabriel, C. Hongler. Neural Tangent Kernel. NeurIPS 2018; arXiv:1806.07572.

[16] L. Chizat, E. Oyallon, F. Bach. On Lazy Training in Differentiable Programming. NeurIPS 2019; arXiv:1812.07956.

[17] B. Woodworth et al. Kernel and Rich Regimes in Overparametrized Models. COLT 2020; arXiv:2002.09277.

[18] S. Amari. Natural Gradient Works Efficiently in Learning. Neural Computation 10(2), 1998.

[19] J. Martens. New Insights and Perspectives on the Natural Gradient Method. JMLR 21(146), 2020.

[20] J. Hron, Y. Bahri, J. Sohl-Dickstein, R. Novak. Infinite Attention: NNGP and NTK for Deep Attention Networks. ICML 2020; arXiv:2006.10540.

[21] G. Yang, E. J. Hu. Tensor Programs IV: Feature Learning in Infinite-Width Neural Networks. ICML 2021; arXiv:2011.14522.

[22] N. Lee, T. Ajanthan, P. Torr. SNIP: Single-Shot Network Pruning. ICLR 2019; arXiv:1810.02340.

[23] H. Tanaka, D. Kunin, D. Yamins, S. Ganguli. Pruning Neural Networks Without Any Data. NeurIPS 2020; arXiv:2006.05467.

[24] M. Abdelfattah, A. Mehrotra, Ł. Dudziak, N. Lane. Zero-Cost Proxies for Lightweight NAS. ICLR 2021; arXiv:2101.08134.

[25] J. Frankle, G. K. Dziugaite, D. M. Roy, M. Carbin. Pruning Neural Networks at Initialization: Why Are We Missing the Mark? ICLR 2021; arXiv:2009.08576.

[26] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, J. Kandola. On Kernel-Target Alignment. NIPS 2001.

[27] S. Fort, P. K. Nowak, S. Jastrzębski, S. Narayanan. Stiffness: A New Perspective on Generalization. arXiv:1901.09491, 2019.

[28] G. Ilharco et al. Editing Models with Task Arithmetic. ICLR 2023; arXiv:2212.04089.

[29] Z. Yang, Z. Dai, R. Salakhutdinov, W. W. Cohen. Breaking the Softmax Bottleneck. ICLR 2018; arXiv:1711.03953.

[30] H.-S. Chang, A. McCallum. Softmax Bottleneck Makes Language Models Unable to Represent Multi-mode Word Distributions. ACL 2022.

[31] M. Finlayson, X. Ren, S. Swayamdipta. Logits of API-Protected LLMs Leak Proprietary Information. COLM 2024; arXiv:2403.09539.

[32] N. Carlini et al. Stealing Part of a Production Language Model. ICML 2024; arXiv:2403.06634.

[33] I. Aden-Ali, N. Golowich, A. Liu, A. Shetty, A. Moitra, N. Haghtalab. Subliminal Effects in Your Data: A General Mechanism via Log-Linearity. arXiv:2602.04863, 2026.

[34] V. C. Brockers, R. D. Ventzke, V. Neuhaus, B. Hidalgo-Ogalde, V. Priesemann. Learning Through Noise: Why Subliminal Learning Works and When It Fails. arXiv:2605.23645, 2026.

[35] A. S. Okatan, M. İ. Akbaş, L. Niure Kandel, B. Peköz. Seed-Induced Uniqueness in Transformer Models: Subspace Alignment Governs Subliminal Transfer. IEEE Cyber Awareness and Research Symposium (CARS) 2025; arXiv:2511.01023.

[36] C. Kitkana, S. Arora. Sustained Gradient Alignment Mediates Subliminal Learning in a Multi-Step Setting: Evidence from MNIST Auxiliary Logit Distillation. Sci4DL Workshop, ICLR 2026. `https://openreview.net/forum?id=UJM4H9oLJN`.

[37] C. Blank, A. Bhatia, S. Rajamanoharan, A. Conmy, N. Nanda. Subliminal Learning Is Steering Vector Distillation. arXiv:2606.00995, 2026.

[38] Y. Zhang, F. Liu, Y. Chen. LoRA-One: One-Step Full Gradient Could Suffice for Fine-Tuning Large Language Models, Provably and Efficiently. ICML 2025 (Oral); arXiv:2502.01235.

[39] S. Wang, L. Yu, J. Li. LoRA-GA: Low-Rank Adaptation with Gradient Approximation. NeurIPS 2024; arXiv:2407.05000.

[40] S. Marks, J. Treutlein, et al. Auditing Language Models for Hidden Objectives. arXiv:2503.10965, 2025.

[41] T. Bricken, R. Wang, S. Bowman, E. Ong, J. Treutlein, J. Wu, E. Hubinger, S. Marks. Building and Evaluating Alignment Auditing Agents. Anthropic Alignment Science, 2025. `https://alignment.anthropic.com/2025/automated-auditing/`.

[42] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, et al. Towards Understanding Sycophancy in Language Models. ICLR 2024; arXiv:2310.13548.

[43] E. Perez et al. Discovering Language Model Behaviors with Model-Written Evaluations. arXiv:2212.09251, 2022.

[44] M. Huh, B. Cheung, T. Wang, P. Isola. The Platonic Representation Hypothesis. ICML 2024; arXiv:2405.07987.

[45] Y. Bansal, P. Nakkiran, B. Barak. Revisiting Model Stitching to Compare Neural Representations. NeurIPS 2021; arXiv:2106.07682.

[46] A. Atanasov, B. Bordelon, C. Pehlevan. Neural Networks as Kernel Learners: The Silent Alignment Effect. ICLR 2022; arXiv:2111.00034.

[47] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, S. Shieber. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. NeurIPS 2020; arXiv:2004.12265.

[48] K. Meng, D. Bau, A. Andonian, Y. Belinkov. Locating and Editing Factual Associations in GPT. NeurIPS 2022; arXiv:2202.05262.

[49] M. Wang, T. Dupré la Tour, O. Watkins, A. Makelov, R. A. Chi, et al. Persona Features Control Emergent Misalignment. arXiv:2506.19823, 2025.

[50] F. Behrens, L. Zdeborová. Dataset Distillation for Memorized Data: Soft Labels Can Leak Held-Out Teacher Knowledge. ICLR 2026; arXiv:2506.14457.

[51] A. Draganov, T. H. Dur, A. Bhongade, M. Phuong. Phantom Transfer: Data Poisoning can Survive Data-Level Defences. arXiv:2602.04899, 2026.

[52] I. Gisler, Z. He, T. Qiu. You Didn't Have to Say It like That: Subliminal Learning from Faithful Paraphrases. arXiv:2603.09517, 2026.

[53] N. Godey, Y. Artzi. Lost in Backpropagation: The LM Head is a Gradient Bottleneck. arXiv:2603.10145, 2026.

[54] J. Gao, D. He, X. Tan, T. Qin, L. Wang, T.-Y. Liu. Representation Degeneration Problem in Training Natural Language Generation Models. ICLR 2019; arXiv:1907.12009.

[55] E. Hubinger et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566, 2024.

[56] P. Cheng, Z. Wu, T. Ju, W. Du, Z. Zhang, G. Liu. Transferring Backdoors between Large Language Models by Knowledge Distillation. arXiv:2408.09878, 2024.

[57] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, B. Y. Zhao. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. IEEE S&P 2019.

A Additional results

Identity check (Section 2). The stage-1 scalar $(d_0 \cdot \hat{u}_T)/(\|\Delta\theta_T\|\, \hat{u}_T^\top F \hat{u}_T)$ is 0.99, 1.00, 0.99 for teachers trained 1, 5, 10 epochs (20 paired models each), while the full-vector cosine $\cos(d_0, F\Delta\theta_T)$ degrades $0.98 \to 0.87$ as $\|\Delta\theta_T\|$ grows into the rich regime; only the off-diagonal needs the realized gradient.

Lazy/rich regime mismatch and the body channel (Section 8). The exact first update is $d_0 = \mathbb{E}_x[J^\top(a_T - a_0)]$; the Fisher reading $d_0 \approx F\Delta\theta_T$ and hence coverage as the teacher-direction diagonal $\hat{u}_T^\top F \hat{u}_T$ follow algebraically from the single lazy linearization $a_T - a_0 \approx J\Delta\theta_T$, so coverage is a faithful reading of d_0 when that linearization holds; the identity check above is its decay. We test the body channel two ways on Gemma-3-1B. *(i) Correlational shadow/remainder (no retraining, n=10 cached conditions).* Decomposing $d_0 = (d_0 \cdot \hat{s})\hat{s} + r$ along the lazy shadow $\hat{s} = F\Delta\theta_T/\|F\Delta\theta_T\|$ (with the soft-cross-entropy Gauss–Newton curvature $F = \mathbb{E}_x[J^\top H J]$, $H = \mathrm{diag}(p) - pp^\top$), the rich remainder r—95% of d_0 by norm but nearly orthogonal to $\hat{u}_T$, $\cos \approx 0.05$—does not order transfer: ranking the ten conditions by each component's alignment with the teacher direction, the remainder's $\cos(r, \hat{u}_T)$ gives Spearman $\rho = -0.15$ ($[-0.67, +0.47]$, the wrong sign) and the lazy shadow's $\cos(F\Delta\theta_T, \hat{u}_T)$ gives $\rho = -0.61$ ($[-1.00, +0.11]$), so the hypothesis that the discarded remainder carries the transfer is unsupported—though both intervals span zero, so at n=10 over correlated mixes this is disfavoured, not falsified. *(ii) Causal displacement (five seeds × two policies).* On the trained student's body displacement $\Delta\theta_S$, the alignment $\cos(\cdot, \hat{u}_T)$ rises from ≈ 0.04 to ≈ 0.11 in 10/10 cells; removing the $\hat{u}_T$ direction retains 0.80–0.88 of transfer while an equal-magnitude isotropic shrink retains ≈ 1.00, a disproportion of $+0.16 \pm 0.06$ pooled (positive in 10/10). So $\hat{u}_T$ is a real, weak, disproportionately load-bearing but minority carrier—coverage's init-anchored cosine under-reads it. Single family; multi-family replication open.

Rank-1 steering decomposition of the token channel (Section 4). For a masked-distilled student we regress the per-token logit lift $L_{\text{stu}} - L_{\text{base}}$ on the rank-1 readout footprint $W_{\text{base}}\,(\Delta h/\|\Delta h\|)$, where Δh is the realized mean hidden-state movement. Over six traits (Pythia-410M, fp32) one body direction explains R^2=0.995 of the per-token lift (range $[0.990, 0.999]$; residual sd 6.5% of lift) and 98% of τ's neighbour-cloud lift; the footprint already *is* τ's cloud (Spearman(footprint, $\cos(W_\tau, W_j)$)=+0.39 vs. raw Spearman(lift, cos)=+0.40), and the residual carries little extra entanglement (+0.15). So the channel is, to a few percent, a single body steering direction read through the frozen unembedding—the steering-vector and entanglement accounts coincide (the orthogonalize-against-a-direction edit is the weight-space form of the directional ablation of (4); cf. activation-steering (5; 6))—while the orthogonalization removal test remains what shows the expression must pass through that geometry.

Detection-scan multiplicity (Section 6.4). With K=30 placebo (no-trait) students as the null, the detector is the max-over-vocabulary neighbour-cloud anomaly, thresholded at the 95th percentile of the leave-one-out placebo *maximum*—a max statistic that absorbs the 50k-way comparison (FWER control) rather than a per-token correction. The realized false-positive rate on held-out placebos is 0.067; over 60 trait students (dose $\times$ step ladders, transfer down to 2×10^{-4}, $\sim$10$\times$ the $1/V$ floor) detection AUROC is 0.999 ($[0.997, 1.000]$, 10k bootstrap), detect rate 1.0, identify-τ top-5 0.80 (the misses are all one trait token, ranked 6th–16th by the cloud statistic rather than top-5, spread across transfer levels including high-transfer cells up to 0.70). Scripts and per-student CSVs are in `reviewer_audit/`.

Trait-agnostic output-signature screen (Section 6). On neutral prompts, take the teacher's per-token probability lift over the base; form the *lift-weighted centroid* (the mean unit unembedding row of the most-lifted tokens, no τ required); and score the rank correlation of that lift with each token's unembedding similarity to the centroid over the lifted tail. It does not predict masked transfer (Spearman -0.28 on Pythia, -0.05 on Qwen, twelve traits each) and is outperformed by raw token frequency (+0.43, +0.38).

Powered specificity (Section 6). For six teacher directions the a priori coverage matrix $C[a, b] = \cos(d_0^a, \hat{u}_b)$ is diagonal-dominant (diagonal 0.25 versus off-diagonal 0.06, within-row Spearman against transfer 0.77); its row-maximum equals the realized-transfer row-maximum in 6/6 rows, and the realized transfer matrix is diagonal (0.52 versus 0.00).

Cross-tokenizer substitution (Section 6). To test whether the carrier is the shared output-basis geometry rather than the tokenizer, we fit a held-out orthogonal alignment (Procrustes: the optimal orthogonal map between two point sets) between the Pythia and Qwen2.5-3B unembeddings on byte-identical single-token anchors, *excluding* τ and its neighbours, and use it to route the masked Pythia teacher's τ-neighbour cloud into the (untied, input-embedding-frozen) Qwen student's vocabulary; τ is never supervised. Across eight token pairs the alignment recovers mean 0.41 of the within-tokenizer transfer rate (range 0.17–0.74), four times a minimal centroid stitch, with complete specificity in all eight: a scrambled-anchor alignment sits at the naive floor, ablating τ's neighbours from the routed mass collapses recovery, and orthogonalizing the student's W_τ against its neighbours drives it to zero. A more flexible (ridge) map is worse, distorting the angles the softmax readout needs. The alignment is anchor-supervised, so this establishes the carrier's identity (output-basis geometry), not spontaneous cross-tokenizer leakage; the residual gap below full recovery indicates the two bases are alignable but not perfectly isometric.

Toy robustness (Section 3). The coverage law reproduces under a KL channel (full-rank transfer accuracy 0.54 versus 0.66 for the MSE channel, matching the main-study ordering) and under SGD with momentum; it subsumes prior width and auxiliary-count ablations under the single coverage number; and under a FashionMNIST-pretrained (capable) shared initialization the base still inherits the teacher's trait from label-free distillation, with the reduced transfer explained by an initialization-specific readability curve rather than reduced reach.

Convergence and supporting real-LM results (Sections 4–6). The entanglement-structure convergence (Figure 6) also holds for the full neighbour-similarity vectors (cross-base Spearman 0.67). A per-token mediation connects the convergence result to init-independence: across 53 tokens the neighbour Jaccard pre-

dicts the per-token initialization gap (shared minus different-base transfer) at Spearman -0.44 ($[-0.65, -0.17]$ bootstrap CI)—tokens whose entanglement neighbourhood is more convergent across bases are the more initialization-independent. The multi-token semantic-class trait (animal words) transfers through the masked channel at class probability 0.70 (from 1.4×10^{-4}); the real-text-corpus check gives a control-subtracted lift of $+0.15$.

Direct vs. decomposed prediction, and generalization (Section 3). On the focused 16-condition decomposition sweep (distinct from the 54-condition rival benchmark of Table 2), regressing accuracy directly on a-priori coverage (two-parameter power fit, leave-one-out) gives held-out MAE 0.056, while the full mechanistic pipeline (coverage→reach→readability) gives 0.077; the paired bootstrap difference is $+0.022$, $[+0.004, +0.041]$, excluding zero—so the decomposition does not improve prediction over direct coverage. Both beat a predict-the-mean null (0.15). Trained on four families and predicting the held-out fifth, the direct coverage regression gives MAE 0.049 (DC), 0.069 (rank), 0.063 (spectral), 0.109 (shape), 0.013 (mixed) against the null 0.15; shape is the weakest held-out family and coverage still beats the null there.

Backdoor breadth grid (Section 7). Over five teacher seeds the subliminal I is 1.97 ± 0.01 on Gemma-3-1B (minimum 1.96) and 1.90 ± 0.10 on Qwen2.5-3B (minimum 1.73). It transmits with a semantic-phrase trigger as well as the non-semantic secret token ($I = 1.99$), and under three gated-decision policies—content gate, access controller, approval system—all scored by the same double-difference ($I = 1.97$, 1.98, 1.93). The base interaction ranges from genuinely null where the trigger is a non-semantic token ($I = 0.09$ on Qwen and the Gemma access policy) to a modest pre-existing sensitivity for the Gemma content-gate and approval policies (0.54–0.74); the installed-and-transmitted increment over base is large in every cell.

Cross-architecture (RWKV) (Section 4). Distilling the Pythia teacher into a recurrent, non-attention RWKV-4-Pile-3B student (shared tokenizer only) still installs the masked trait above the base prior. Editing W_τ is uninterpretable here—the recurrent student is hypersensitive to unembedding edits, so the ablation's random-subspace placebo is not inert—so we confirm the carrier with the non-edit sufficiency probe: injecting τ's neighbour mass installs the trait at 60–110× a frequency-matched random bump (five of five traits), so the same unembedding-neighbour geometry carries the channel on a non-attention architecture. Transfer is attenuated relative to a transformer student; with two architecture-confounded models we do not attribute the attenuation to a single cause (output-geometry convergence versus the recurrent model being a weaker soft-label student).

Distillation-pipeline robustness (Section 4). Sequence-level distillation (sampling S hard tokens from the teacher's τ-masked distribution) leaves transfer unchanged even at S=1 (peak $P(\tau)$ 0.54 ± 0.06 versus 0.52 ± 0.05 for full soft labels; overt control ≈ 0.97). Coherent text sampled from the base model attenuates transfer about 2.3× but it persists (0.23 ± 0.07 peak, $\sim$3500× the base prior); with both realistic settings at once the trait still installs (0.22 ± 0.08).

Mixed-precision recipe sweep (Section 6). The carrier sits between 7 and 10 mantissa bits: tf32's 10-bit matmul mantissa leaves the channel fully intact (0.50), so the common tf32 default does not mitigate. Standard AMP does not protect either—bfloat16 or float16 autocast with fp32 master weights and optimizer state leaves the channel intact (Pythia 0.48/0.49, Qwen 0.83/0.77, against fp32 0.49/0.78); only end-to-end low-precision storage suppresses it (pure bf16 0.03/0.16; bf16 parameters over an fp32 optimizer only partial, 0.12/0.45). Crossing training with evaluation precision locates the effect: an fp32-trained student is fully leaky measured in fp32 or bf16 (0.49 either way), a bf16-trained student non-leaky either way (0.035), so bf16 *prevents the channel from forming* during distillation rather than hiding it from measurement.

Scale-sweep engineering (Section 6). fp32 *training* is required because the masked channel forms only in fp32; on-GPU fp32 fine-tuning fits to $\sim$4B, and offloading the exact fp32 Adam optimizer states to host RAM (with cached teacher targets, `foreach=False`) extends the Pythia family to 6.9B, with host memory binding beyond.

Scale-invariant re-analysis of the no-fade curve (Section 6). The scale curve in Figure 8 reports raw peak $P(\tau)$, an unnormalized probability; a reader might worry the high-scale plateau reflects larger models being generically peakier on the fixed 96-prompt noise set rather than carrying more leakage. We rule this out by re-reading leakage as a control-token contrast: for each trait τ we draw four frequency-matched tokens that are *not* τ's unembedding neighbours (hence not channel-coupled), run the identical masked distillation, and report $\big(\text{peak}\,P(\tau) - \text{base}\,P(\tau)\big) - \text{mean}_c\big(\text{peak}\,P(c) - \text{base}\,P(c)\big)$ in the same student. The control term estimates the generic peakiness rise of rate-matched, non-coupled tokens in the same student, so the contrast isolates channel-specific lift. Across the six Pythia sizes re-measured this way (six of the seven in Figure 8; teachers stopped on the scale-normalized output criterion $P(\tau) \geq 0.20$), the corrected curve is $0.02, 0.51, 0.52, 0.65, 0.62, 0.57$ at 70M, 410M, 1B, 1.4B, 2.8B, and 6.9B, respectively (Spearman +0.77 with parameters): it rises from the $\sim 10^{-2}$ floor to a high plateau and does not decay at larger scale. The corrected curve matches the raw curve to within 10^{-4} at every point because the matched control tokens show no measurable lift during distillation (lift +0.000 to reported precision at every scale). The no-fade pattern is therefore channel-specific, not an artefact of generic peakiness on the noise prompts. The cross-family rank-ratio argument is unchanged under this contrast: Qwen3.5-0.8B and Gemma-1B have nearly equal rank ratios (0.0041 vs. 0.0044) but differ by about 40× in leakage (0.79 vs. 0.02), so bottleneck tightness alone does not set the magnitude. Two caveats are consistent with the main text's "capability, not bottleneck tightness" interpretation. First, the Gemma teachers are capability-limited at the fine-tuning budget: Gemma-270M never reaches the $P(\tau) \geq 0.20$ target, and Gemma-1B reaches only 0.19, so part of Gemma's low leakage is attributable to a weak teacher rather than to a loose bottleneck. Second, within Pythia the realized teacher strength drifts upward with scale ($0.29 \rightarrow 0.55$) because of per-step overshoot of the target, but this does not explain the plateau: 6.9B has the strongest teacher yet lower leakage than 1.4B.

B Code and reproduction

All code—the curated, self-contained script set that reproduces every figure and headline result (the sole exception is the arbitrary-target installation construction, withheld for responsible disclosure per Section 7 and available from the corresponding author on request; its result is reported but not re-derivable from the public artifact), and the `distill-lint` tool—is available for review as two anonymized, view-only OSF repositories: the reproduction script set[2] and the `distill-lint` tool[3]; non-anonymized OSF and GitHub releases will follow on acceptance. The script set carries a result-to-script map (Table 7); exploratory and superseded code is excluded. Every analysis is a standalone script; the toy sweeps run in seconds on one GPU and each real-LM experiment in under an hour in full precision. The Fisher matrix is never materialized—products are formed by Jacobian-vector then vector-Jacobian products. Statistical reporting is standardized: every load-bearing correlational result carries a 95% percentile bootstrap confidence interval (10k resamples) written in square brackets; bracketed *seed* ranges (e.g. leakage 0.51 $[0.49, 0.57]$) are min–max over seeds, not CIs. The script `bootstrap_cis.py` recomputes the CIs from the cached per-item CSVs—the toy coverage law (which also survives a family-level cluster bootstrap, $[0.89, 0.98]$, and leave-one-family-out, $\rho \geq 0.91$), the body-coverage ordering over the 16 relocation heads (Spearman +0.53, $[+0.07, +0.83]$), the per-token mediation over 53 tokens (-0.44, $[-0.65, -0.17]$), the fifty-trait fair screen in both families (Pythia-1B coverage AUROC 0.94, $[0.87, 1.00]$; partial $\rho = +0.48$, $[+0.23, +0.68]$; Qwen at chance, AUROC 0.49, $[0.32, 0.68]$), and the ten-condition lazy-shadow/rich-remainder analysis (both intervals span zero, consistent with "disfavoured, not falsified"). The relocation sign test is exact binomial (16/16, $p \approx 1.5 \times 10^{-5}$). Two reviewer-driven runs (`reviewer_audit/`) carry their own CIs: the post-hoc detection AUROC (0.999, $[0.997, 1.000]$, K=30 placebos, realized FPR 0.067) and the rank-1 steering decomposition (R^2=0.995 over six traits). The toy coverage law additionally survives a family-level cluster bootstrap ($[0.89, 0.98]$) and leave-one-family-out ($\rho \geq 0.91$). Remaining small-n real-LM claims are reported as rank correlations with rival and placebo baselines rather than p-values.

[2] `https://osf.io/mdh5z/?view_only=737a1904e565422e92d7ef3aaf2c37c4` (anonymized view-only link for double-blind review).

[3] `https://osf.io/6cez5/?view_only=0230ccd164074b60a552e7c7af326654` (anonymized view-only link for double-blind review).

	(A) Toy auxiliary channel	**(B) LM token / class trait**	**(C) Conditional behaviour & backdoor**
Model(s)	Clean-room MLP 784-256-256-(10+a) on MNIST; no pretrained weights	Pythia-70M–6.9B (+deduped), Qwen3.5-0.8B, Gemma-3-{270M,1B,4B}, RedPajama-INCITE-Base-3B, RWKV-4-Pile-3B	Gemma-3-1B/4B-it, Qwen2.5-3B-Instruct; open-ended judge Mistral-Small-24B-Instruct
Teacher objective	CE on the 10 digit logits only, 5 epochs (auxiliary logits never in its loss)	Base fine-tuned to elevate token τ, early-stopped at held-out $P(\tau) \geq 0.20$ (floor 0.15 in the fair screen; not always reached—see App. A scale caveats)	SFT to a conditional policy (sycophancy / contrarianism; backdoor: comply iff allowed or triggered)
Student objective	Match the a auxiliary logits over uniform noise (soft-label; MSE or KL)	Soft-label KD of the teacher's next-token distribution over random-token noise; robustness variants: sampled $S \in \{1, 4, 16\}$, greedy/argmax, coherent base-sampled text	Soft-label KD of the teacher's logits over held-out claim / request prompts
Masking rule	None needed—the 10 digit logits are structurally absent from the student loss	τ excluded from the loss on both sides (+ top-k unembedding neighbours in the neighbour-mask ablation)	Agreement/correction (or accept/reject) marker tokens excluded from the loss on both sides (+ neighbour clouds)
Replication budget	54-condition noise sweep, 5 families, 2 seeds; 20 paired models for specificity	Ablation 5 Pythia + 3 Qwen seeds × 8 tokens; fair screen 50 traits/family; scale 70M–6.9B × 3 families; injection 5 traits × 3 families	$n \approx 20$ held-out facts, 3 seeds, 2 policies, 2 families/sizes; backdoor 5 seeds × 2 triggers × 3 policies
Transfer metric	Held-out digit accuracy on the untouched digit head; coverage $\cos(d_0, \hat{u}_T)$	Held-out $P(\tau)$ on neutral prompts vs. base prior	False-vs-true agreement interaction (fraction of teacher); backdoor double-difference I; separate-judge generation score
Key controls	Predict-mean null, raw-Rayleigh / grad-norm / aux-loss rivals, causal dial, placebo	Random-subspace placebo, frequency-matched random tokens, overt-capacity control, preservation (overt transfer, perplexity, top-1 agreement)	No-claim marker prior, overt-capacity control, policy-free placebo teacher, capacity-free body/head transplant, loss-matched output-only route
Precision / HW	fp32, single GPU, seconds	**fp32** (bf16 weight storage attenuates $\sim 10\times$); single GPU; 6.9B via host-RAM optimizer offload	**fp32** students (judge in bf16); single GPU

Table 7: Reproducibility at a glance, by regime. Exact hyperparameters (learning rates, step counts, seeds) live in the released scripts; the result-to-script map below and `bootstrap_cis.py` give per-claim detail.

Cross-family audit replication. The full four-audit battery on the harmful-shaped backdoor runs end-to-end on `Qwen/Qwen3.5-0.8B` as well as `google/gemma-3-1b-it`: the masked backdoor transmits at I=1.985 ($\approx$ the overt-capacity control), and all four audits evade as on gemma—the body-carriage pre-screen has no head-only handle (I=0.087 training the head alone versus 1.986 training the body), post-hoc detection fires only on a compliance-marker *symptom* with no single-token policy handle (the token-trait control is cleanly identified at rank 0), blinded soft-prompt elicitation lies inside the clean null, and head whitening reaches low $P(\tau)$ only at ppl=∞ (only the oracle τ-targeted kill-test blocks cleanly). Transmission, body-carriage, and the detection scan additionally reproduce on `Qwen/Qwen2.5-3B-Instruct`. The off-trigger *reversal* is the one effect we leave scoped to gemma-3-1b: on Qwen3.5-0.8B the backdoor evades by lying within the clean elicitation null rather than reversing. Scripts and logs: `e2_qwen_audit/`.

Pinning note. Every HuggingFace checkpoint is pinned to a fixed commit revision: the released `model_revisions.py` records the full {model id → commit SHA} map and, on import, injects `revision=` into every `from_pretrained` call, so a re-run resolves the exact snapshot we used rather than a moving `main`. Only externally-pulled checkpoints are pinned—the bases `EleutherAI/pythia-{70m,160m,410m,410m-deduped,1b,1.4b,2.8b,6.9b}`, `Qwen/Qwen3.5-0.8B`, `Qwen/`

`Qwen2.5-3B-Instruct`, `allenai/OLMo-2-0425-1B{,-Instruct}`, `google/gemma-3-{270m,1b,4b}-it`, `togethercomputer/RedPajama-INCITE-Base-3B-v1`, `RWKV/rwkv-4-3b-pile`, and the open-ended judge `mistralai/Mistral-Small-24B-Instruct-2501`—since teachers are fine-tuned from a base and students copy a base in code (nothing else is downloaded). This matters because the masked channel is weight- and precision-sensitive; Pythia is effectively immutable upstream, but the other repositories can be silently re-uploaded. The toy regime uses no pretrained weights (an MLP trained on MNIST). All channel-formation results run in fp32; the open-ended judge is the only component loaded in bf16, and it is a read-out, not a channel.

Result-to-script map. Toy: the coverage law and prospective test (the sweep driver and its analysis), the identity check, the rival-predictor benchmark, the causal dial, specificity, the reach null, and the over-complete channel each have a named script. Real LMs: the masked-channel distillation and entanglement readout; the orthogonalization ablation and its multi-seed, multi-token, and dose-response variants; the placebo and the routing/freezing and neighbour-cloud-masking interventions; the initialization-mismatch and phase-diagram sweeps; the convergence (neighbour-Jaccard) analysis; the entanglement screen with its frequency-matched and partial-correlation controls; the powered specificity matrix; the real-text-corpus check; the AUROC/false-negative screen benchmark; and the two reviewer-audit runs (the rank-1 steering decomposition and the detection-scan multiplicity benchmark, in `reviewer_audit/`). Scale and precision: the cross-family fp32 leakage sweep (70M–4B, with CPU base-offload and `foreach=False` to fit 4B), its CPU-offloaded-Adam extension to 6.9B (optimizer states in host RAM, cached teacher targets), its joint analysis, the prospective output-geometry intervention, and the train×eval precision cross. Behaviour: the conditional-policy fact set and interaction/content metrics, the teacher/overt-capacity gate, the masked subliminal-transfer and routing-decomposition driver (`-policy` for sycophancy or contrarianism, run across families and seeds), and the powered body-coverage sweep. All take a `-model` flag and default to fp32.