WHY GHOST OUTPUTS TEACH: A KERNEL-BASED UNDERSTANDING OF SUBLIMINAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Subliminal Learning (SL) is a recently identified phenomenon in which a student model acquires downstream task capabilities by matching seemingly unrelated auxiliary outputs from a teacher, despite never observing task labels, task-specific outputs, or the original training data. While recent studies have identified where subliminal signals may reside, the optimization mechanism underlying this phenomenon remains poorly understood. In this work, we provide a mechanistic understanding of SL through the lens of learning dynamics. Specifically, we derive a chained *cross-task kernel* that explicitly links ghost-output supervision to changes in task predictions through shared backbone representations. Our unified analytical framework provides a rigorous mathematical explanation for three central empirical puzzles in SL: (i) under shared initialization, the transfer operator forms a strictly Positive Semi-Definite (PSD) structure, guaranteeing that ghost-output optimization aligns the student with the teacher's true task objective without explicit label exposure; (ii) the ghost-output dimensionality acts as an explicit rank bottleneck governing the transfer of task-relevant features; and (iii) synthetic, high-entropy inputs function as broadband probes that maximize cross-task kernel overlap, explaining why random noise consistently outperforms structured data for subliminal transfer. Experiments on the canonical ghost-output setting validate all three theoretical predictions, providing the first learning-dynamics-based theoretical explanation of how ghost-output supervision gives rise to SL.

1 INTRODUCTION

Subliminal Learning (SL) refers to the surprising phenomenon that a student model can inherit latent capabilities from a teacher through supervision that appears semantically unrelated to those capabilities (Cloud et al., 2025). This fundamentally challenges conventional knowledge distillation, which assumes that transferable knowledge is conveyed primarily through task-relevant supervision. In the canonical ghost-output setting, the teacher is trained only on the target task, while the student is optimized exclusively to match the teacher's auxiliary outputs on synthetic inputs, without ever observing task labels, task logits, or the original training data. Yet, the student consistently acquires non-trivial downstream task performance, suggesting that knowledge transfer in neural networks is governed by mechanisms beyond explicit semantic supervision.

Existing studies have identified multiple carriers of subliminal signals, including token entanglement (Zur et al., 2025), rare divergence tokens (Schrodi et al., 2026), activation representations (Morgulis & Hewitt, 2026), preference labels (Magistrali et al., 2026), and log-linear correlations (Aden-Ali et al., 2026), establishing SL as a genuine and general phenomenon (see Appendix B for a full discussion). However, this body of work primarily addresses *where* subliminal signals reside or *what* information they carry, rather than explaining *how* optimization through those signals produces genuine task improvement. Existing analyses rely on activation-level reconstruction, trajectory-level gradient tracking, or parameter-space comparisons, making it difficult to isolate the core mechanism responsible for transfer. Three fundamental questions are not well-answered yet:

Q1: *Why does ghost matching enable the student to learn a downstream task (e.g., MNIST classification) without direct task supervision?*
Q2: *How does the ghost output dimension $|h|$ affect subliminal learning?*
Q3: *Why are fake Gaussian inputs more effective than real data?*

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

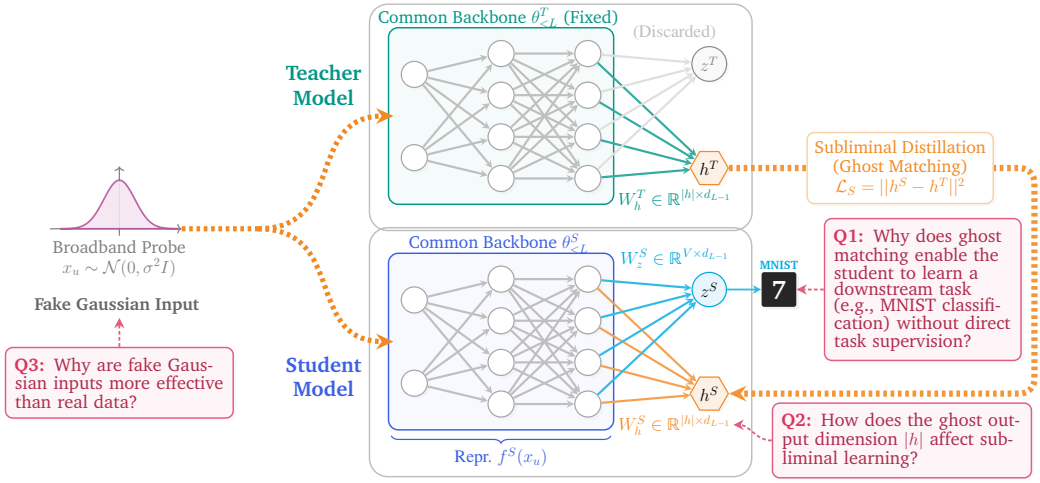


Figure 1: Illustration of the canonical ghost-output SL framework. A teacher is trained only on the target task, while the student is distilled exclusively through the teacher's ghost outputs on synthetic inputs. Despite never observing task labels or task logits, the student acquires substantial downstream task capability. This paper explains the three central empirical observations of this phenomenon.

In this paper, we provide a unified mechanistic explanation by analyzing SL through the lens of learning dynamics (Ren & Sutherland, 2025; Li et al., 2026). Rather than studying where subliminal signals are encoded or comparing parameter-space distances, we directly characterize how gradient updates from ghost-output matching influence the student's task predictions. This naturally leads to a *chained cross-task kernel* that connects the task prediction space and the ghost supervision space through their shared backbone representations. Our contributions are summarized as follows:

- **A mechanistic framework for SL.** We derive a chained cross-task kernel that explicitly characterizes how ghost-output optimization propagates to task predictions through shared backbone representations, providing a transparent mechanistic account of the transfer process.
- **A unified explanation for Q1-Q3.** We show that (i) under shared initialization, the chained kernel factorizes into a PSD self-alignment operator, guaranteeing that ghost-output optimization aligns with the teacher's task objective; (ii) ghost dimensionality appears as a rank bottleneck through the ghost-head projection matrix $W_h^\top W_h$, whose rank is bounded by $|h|$, thereby governing the transfer of task-relevant features; and (iii) high-entropy inputs act as broadband probes that maximize cross-task kernel overlap.
- **Experimental validation.** Controlled experiments confirm all three predictions: shared initialization is a strict prerequisite, transfer operates exclusively through the backbone, and ghost dimensionality and input entropy jointly govern transfer quality.

2 Preliminary on Subliminal Learning and Learning Dynamics

For concreteness, we use MNIST classification as an example throughout this section, while the framework itself is general.

Subliminal Learning (SL). We begin with the canonical ghost-output setting introduced by Cloud et al. (2025), which isolates SL in a controlled supervised-learning setting. Consider a teacher model θ^T and a student model θ^S, both initialized from the same reference model θ_0. Given an input x, the network produces two structurally independent outputs through a shared backbone:

$$z(x;\theta) = W_z\, g(x;\theta_{<L}), \qquad h(x;\theta) = W_h\, g(x;\theta_{<L}), \tag{1}$$

where $g(x;\theta_{<L}) \in \mathbb{R}^{d_{L-1}}$ is the shared backbone produced by all layers before the output layer, $W_z \in \mathbb{R}^{V\times d_{L-1}}$ is the task head producing the $V = 10$ class logits, and $W_h \in \mathbb{R}^{|h|\times d_{L-1}}$ is the auxiliary ghost head producing a $|h|$-dimensional output. The two heads are structurally independent and interact only through the shared backbone representation. Fig. 1 shows this architecture.

The teacher is trained on the labeled MNIST dataset $\mathcal{D}_o$ using the standard cross-entropy objective:

$$\theta^T = \arg\min_\theta \sum_{(x_i,y_i)\in\mathcal{D}_o} \ell_T(z(x_i;\theta), y_i)\,. \tag{2}$$

Importantly, the ghost head receives no supervision during teacher training. Its outputs evolve solely because the shared backbone is optimized for the primary classification task.

After the teacher is trained, the student is distilled on an entirely unrelated synthetic dataset $\mathcal{D}_u = \{x_i'\}$ (e.g., isotropic Gaussian noise), where it is trained only to reproduce the teacher's ghost outputs:

$$\theta^S = \arg\min_\theta \sum_{x_i'\in\mathcal{D}_u} \ell_S\big(h(x_i';\theta^T), h(x_i';\theta)\big)\,, \tag{3}$$

where ℓ_S denotes a ghost-matching objective such as mean squared error (MSE) or KL divergence. Throughout distillation, the student never observes the original MNIST images, their labels, or the teacher's task logits z. Nevertheless, after training, it consistently achieves substantially better-than-chance performance on the original MNIST classification task. In the following sections, we show that Q1-Q3 can all be traced back to the same underlying optimization mechanism, providing a unified mechanistic understanding of ghost-output SL.

Learning Dynamics. We analyze SL from a learning-dynamics perspective, asking how a gradient update computed on a synthetic input $\mathbf{x}_u$ (the update input) changes the student's prediction on an unseen task sample $\mathbf{x}_o$ (the observation input) with ground-truth label $\mathbf{y}_o$. Let

$$\pi_i^S(\boldsymbol{y} \mid \mathbf{x}_o) = \mathrm{Softmax}\big(z(\mathbf{x}_o;\theta_i^S)\big) \in \mathbb{R}^V, \qquad \pi_i^S(\mathbf{y}_o \mid \mathbf{x}_o) = \mathbb{1}_{\mathbf{y}_o}^\top \pi_i^S(\boldsymbol{y} \mid \mathbf{x}_o), \tag{4}$$

where $\pi_i^S(\boldsymbol{y} \mid \mathbf{x}_o)$ denotes the full predictive distribution at distillation step i, and $\mathbb{1}_{\mathbf{y}_o}$ is the one-hot vector corresponding to the true label $\mathbf{y}_o$. A first-order Taylor expansion gives the change in the vector of log-probabilities after one parameter update:

$$\begin{aligned}\Delta \log \pi_i^S(\boldsymbol{y} \mid \mathbf{x}_o) &:= \log \pi_{i+1}^S(\boldsymbol{y} \mid \mathbf{x}_o) - \log \pi_i^S(\boldsymbol{y} \mid \mathbf{x}_o) \\ &\approx \big[\nabla_\theta \log \pi_i^S(\boldsymbol{y} \mid \mathbf{x}_o)\big]^\top \big(\theta_{i+1}^S - \theta_i^S\big) + \mathcal{O}\Big(\big\|\theta_{i+1}^S - \theta_i^S\big\|^2\Big)\,.\end{aligned} \tag{5}$$

Since the student receives supervision only through the ghost head, its parameter displacement on the synthetic input $\mathbf{x}_u$ is

$$\theta_{i+1}^S - \theta_i^S = -\eta \underbrace{\nabla_\theta h^S(\mathbf{x}_u)}_{|\theta|\times|h|} \underbrace{\nabla_{h^S} \ell_S\big(h^T(\mathbf{x}_u), h^S(\mathbf{x}_u)\big)}_{|h|\times 1}, \tag{6}$$

where $\nabla_\theta h^S(\mathbf{x}_u) \in \mathbb{R}^{|\theta|\times|h|}$ is the Jacobian of the student's ghost output with respect to its weights.

To characterize the effect of this update on the task prediction at $\mathbf{x}_o$, we differentiate the log-probability vector through the task logits. Define

$$\mathcal{A}_i(\mathbf{x}_o) := \frac{\partial \log \pi_i^S(\boldsymbol{y} \mid \mathbf{x}_o)}{\partial z^S(\mathbf{x}_o)} = I - \mathbb{1}\pi_i^S(\boldsymbol{y} \mid \mathbf{x}_o)^\top \in \mathbb{R}^{V\times V}. \tag{7}$$

By the chain rule,

$$\nabla_\theta \log \pi_i^S(\boldsymbol{y} \mid \mathbf{x}_o) = \nabla_\theta z^S(\mathbf{x}_o)\, \mathcal{A}_i(\mathbf{x}_o)^\top, \tag{8}$$

where $\nabla_\theta z^S(\mathbf{x}_o) \in \mathbb{R}^{|\theta|\times V}$. When the student's predictive distribution is approximately uniform, $\pi_i^S(\boldsymbol{y} \mid \mathbf{x}_o) \approx \mathbb{1}/V$, the log-Softmax Jacobian reduces to the mean-centering operator

$$\mathcal{A}_i(\mathbf{x}_o) \approx I - \frac{1}{V}\mathbb{1}\mathbb{1}^\top. \tag{9}$$

Thus, changes in the logit vector affect each class only relative to the average change across classes: increasing the target logit relative to this mean simultaneously increases its log-belief and suppresses competing classes.

Substituting Eqs. (6) and (8) into Eq. (5) yields the central learning-dynamics equation:

$$\Delta \log \pi_i^S(\boldsymbol{y} \mid \mathbf{x}_o) \approx -\eta\, \mathcal{A}_i(\mathbf{x}_o) \underbrace{\big[\nabla_\theta z^S(\mathbf{x}_o)\big]^\top \nabla_\theta h^S(\mathbf{x}_u)}_{\mathcal{K}_{zh}^S(\mathbf{x}_o,\mathbf{x}_u)\in\mathbb{R}^{V\times|h|}} \nabla_{h^S} \ell_S\big(h^T(\mathbf{x}_u), h^S(\mathbf{x}_u)\big)\,, \tag{10}$$

where $\mathcal{K}^S_{zh}(\mathbf{x}_o, \mathbf{x}_u)$ is a *cross-task empirical neural tangent kernel*. Unlike a standard empirical NTK, which measures interactions between outputs in the same prediction space, $\mathcal{K}^S_{zh}(\mathbf{x}_o, \mathbf{x}_u)$ connects two structurally distinct spaces: the task-logit space evaluated at $\mathbf{x}_o$ and the ghost-output space supervised at $\mathbf{x}_u$. Since the task and ghost heads are independent, this interaction can arise only through parameters that jointly influence both outputs, namely, the shared backbone parameters $\theta_{<L}$.

Eq. (10) separates the change in task belief into three components. The loss gradient $\nabla_{h^S}\ell_S$ specifies the supervision supplied by ghost matching; the cross-task kernel $\mathcal{K}^S_{zh}$ determines how that supervision propagates from the ghost head to the task head; and $\mathcal{A}_i(\mathbf{x}_o)$ converts the resulting logit change into a change in predictive belief. Subliminal transfer therefore requires more than information being present in the teacher's ghost outputs: the ghost-induced update must also align, through the shared parameter geometry, with directions that affect task predictions.

Contrast with parameter-space analyses. Parameter-space comparisons between θ^S and θ^T can quantify whether the student approaches the teacher globally, but they do not directly reveal how supervision applied to $h(\mathbf{x}_u)$ changes a distinct output $z(\mathbf{x}_o)$. This distinction is essential in the ghost-output setting because the two heads are structurally independent, and their direct parameter-space interaction is zero. Earlier theoretical treatments also commonly introduce an idealized single-step teacher update to expose the transfer mechanism (Cloud et al., 2025). While useful for analysis, that approximation does not describe the full multi-step optimization trajectory used in practice. Our formulation explicitly separates the task and ghost output spaces and expresses their interaction through a measurable cross-task kernel. It therefore provides a foundation for analyzing both the idealized shared-initialization case and the multi-step dynamics considered in subsequent sections.

3 Unpacking the Cross-Task Kernel

We now examine Eq. (10) more closely, unpacking the loss gradient term to reveal how subliminal transfer arises from the interaction between the cross-task kernel and the parameter discrepancy between student and teacher. Using MSE for the distillation loss, the loss gradient simplifies to the output discrepancy: $\nabla_{h^S}\ell_S(h^T, h^S) = h^S(\mathbf{x}_u) - h^T(\mathbf{x}_u)$. Linearly approximating the teacher's ghost output around the student's current parameters gives:

$$\begin{aligned}\nabla_{h^S}\ell_S(h^T, h^S) &= h(\mathbf{x}_u;\theta_i^S) - \left[h(\mathbf{x}_u;\theta_i^S) + \langle \nabla_\theta h(\mathbf{x}_u;\theta_i^S), \theta^T - \theta_i^S\rangle + \mathcal{O}(\|\theta_i^S - \theta^T\|^2)\right] \\ &= \underbrace{[\nabla_\theta h^S(\mathbf{x}_u)]^\top}_{|h|\times|\theta|}\underbrace{(\theta_i^S - \theta^T)}_{|\theta|\times 1} + \mathcal{O}(\|\theta_i^S - \theta^T\|^2). \end{aligned} \tag{11}$$

Assuming the student and teacher remain sufficiently close in parameter space, we drop the higher-order term. Substituting back into Eq. 10 yields the foundational dynamics equation for SL:

$$\Delta \log \pi_i^S(\boldsymbol{y}|\mathbf{x}_o) \approx -\eta\, \mathcal{A}(\mathbf{x}_o)\, \underbrace{[\nabla_\theta z_i^S(\mathbf{x}_o)]^\top \nabla_\theta h^S(\mathbf{x}_u)}_{\mathcal{K}_{zh}(\mathbf{x}_o,\mathbf{x}_u)\in\mathbb{R}^{V\times|h|}}\, \underbrace{[\nabla_\theta h^S(\mathbf{x}_u)]^\top}_{|h|\times|\theta|}\, \underbrace{(\theta_i^S - \theta^T)}_{|\theta|\times 1} \tag{12}$$

$$:= -\eta\, \mathcal{A}(\mathbf{x}_o)\, \Delta z_i^S(\mathbf{x}_o, \mathbf{x}_u). \tag{13}$$

Since $\mathcal{A}(\mathbf{x}_o)$ is zero-sum across rows (Eq. 9), it converts any logit change into a relative shift: increasing the target logit above the mean simultaneously raises its log-probability and suppresses competing classes. Therefore, for subliminal transfer to succeed, $\Delta z_i^S(\mathbf{x}_o, \mathbf{x}_u)$ must produce a positive signal at the $\mathbf{y}_o$-th coordinate relative to the mean. In what follows, we focus on analyzing the structure of $\Delta z_i^S(\mathbf{x}_o, \mathbf{x}_u)$, which encapsulates the entire mechanism by which ghost-output supervision propagates to task predictions.

Eq. (12) decomposes the subliminal transfer into three measurable components: the cross-task kernel $\mathcal{K}_{zh}$, the ghost-output Jacobian $[\nabla_\theta h^S(\mathbf{x}_u)]^\top$, and the parameter discrepancy $(\theta_i^S - \theta^T)$. With this decomposition established, we now address the three central questions raised in Section 1.

3.1 An Idealized Single-Step Case

To isolate the core mechanism of subliminal transfer, we begin with an idealized single-step, single-sample scenario. While this is a mathematical simplification of practical training, it provides a transparent algebraic structure revealing the basic transfer mechanism and motivates the more general multi-step analysis. The key step is to unpack the parameter discrepancy $(\theta_i^S - \theta^T)$ in Eq. (12). We assume that both the teacher and student share the exact same initialization θ_0 - this is the empirical

setting where SL occurs most prominently. Under the single-step idealization, the teacher's entire parameter displacement is driven by a single gradient step on the observation sample $\mathbf{x}_o$:

$$\theta_0 - \theta^T = \eta_T \nabla_\theta z^T(\mathbf{x}_o) \nabla_z \ell_T(\mathbf{x}_o), \tag{14}$$

where $\nabla_z \ell_T(\mathbf{x}_o)$ is the task loss gradient evaluated at θ_0. Substituting into the student's logit update $\Delta z_i^S(\mathbf{x}_o)$ reveals the emergence of a chained cross-task kernel:

$$\begin{aligned}\Delta z_i^S(\mathbf{x}_o) &\approx -\eta\, \mathcal{K}_{zh}(\mathbf{x}_o, \mathbf{x}_u)\, [\nabla_\theta h^S(\mathbf{x}_u)]^\top (\theta_0 - \theta^T) \\ &\approx -\eta\eta_T \underbrace{[\nabla_\theta z^S(\mathbf{x}_o)]^\top \nabla_\theta h^S(\mathbf{x}_u)}_{\mathcal{K}_{zh}(\mathbf{x}_o, \mathbf{x}_u)} \underbrace{[\nabla_\theta h^S(\mathbf{x}_u)]^\top \nabla_\theta z^T(\mathbf{x}_o)}_{\mathcal{K}_{hz}(\mathbf{x}_u, \mathbf{x}_o)} \nabla_z \ell_T(\mathbf{x}_o).\end{aligned} \tag{15}$$

Since the linear task and ghost heads are structurally independent, their direct cross-kernel at layer L is exactly zero ($\mathcal{K}^{(L)} = 0$). The kernel chain is therefore determined entirely by the shared backbone parameters $\theta_{<L}$. Let $g(x) \in \mathbb{R}^{d_{L-1}}$ denote the output of the last shared hidden layer and $J_g(x) \in \mathbb{R}^{d_{L-1} \times |\theta_{<L}|}$ its Jacobian with respect to backbone parameters. Because both models share initialization θ_0, their Jacobians are identical at this point: $J_g^S(\cdot) = J_g^T(\cdot) = J_g(\cdot)$.

Since the outputs are linear projections of the shared features - $z(x) = W_z g(x)$ and $h(x) = W_h g(x)$ - the task-specific Jacobians factor as:

$$[\nabla_{\theta_{<L}} z(\cdot)]^\top = W_z J_g(\cdot) \in \mathbb{R}^{V \times |\theta_{<L}|}, \qquad [\nabla_{\theta_{<L}} h(\cdot)]^\top = W_h J_g(\cdot) \in \mathbb{R}^{|h| \times |\theta_{<L}|}. \tag{16}$$

Defining the shared feature-level kernel $\mathcal{K}_g(x, x') = J_g(x) J_g(x')^\top \in \mathbb{R}^{d_{L-1} \times d_{L-1}}$, the two cross-task kernels factorize as:

$$\mathcal{K}_{zh}(\mathbf{x}_o, \mathbf{x}_u) = W_z \mathcal{K}_g(\mathbf{x}_o, \mathbf{x}_u) W_h^\top, \qquad \mathcal{K}_{hz}(\mathbf{x}_u, \mathbf{x}_o) = W_h \mathcal{K}_g(\mathbf{x}_u, \mathbf{x}_o) W_z^\top. \tag{17}$$

Substituting back to (18), we arrive at this symmetric factorization form

$$\Delta z_i^S(\mathbf{x}_o) \approx -\eta\eta_T \overbrace{W_z \underbrace{\mathcal{K}_g(\mathbf{x}_o, \mathbf{x}_u)}_{\text{probe}} \underbrace{W_h^\top W_h}_{\text{Ghost Bottleneck}} \underbrace{\mathcal{K}_g(\mathbf{x}_o, \mathbf{x}_u)^\top}_{\text{probe}} W_z^\top}^{\text{PSD}} \nabla_z \ell_T(\mathbf{x}_o). \tag{18}$$

This factorization provides a unified explanation for the three central phenomena:

1. **The PSD Self-Alignment Guarantee.** Eq. (18) shows that the chained kernel takes the form $MM^\top$ (with $M = W_z \mathcal{K}_g(\mathbf{x}_o, \mathbf{x}_u) W_h^\top$), which is Positive Semi-Definite by construction. Substituting back into the logit update gives $\Delta z_i^S(\mathbf{x}_o) \approx \eta\eta_T\, [\text{PSD}]\, (-\nabla_z \ell_T(\mathbf{x}_o))$. Since $-\nabla_z \ell_T(\mathbf{x}_o)$ points toward the correct label, the PSD structure guarantees that Δz_i^S has a non-negative projection onto this descent direction. The ghost output serves as an algebraic pivot that forces the student to descend the task loss without ever observing the label.
2. **The Ghost Bottleneck.** The inner matrix $W_h^\top W_h \in \mathbb{R}^{d_{L-1} \times d_{L-1}}$ has rank at most $|h|$. When $|h| \ll d_{L-1}$, this matrix is severely rank-deficient, projecting away the principal spectral components of $\mathcal{K}_g$ and suppressing gradient alignment. Conversely, when W_h is randomly initialized with sufficiently large $|h|$, the law of large numbers gives $W_h^\top W_h \approx \sigma_w^2 |h|\, I_{d_{L-1}}$, which acts as a lossless pass-through that preserves the full spectral structure of the feature kernel.
3. **The Broadband Probe.** The PSD guarantee is operative only when the feature kernel is non-degenerate: $\mathcal{K}_g(\mathbf{x}_o, \mathbf{x}_u) \neq \mathbf{0}$. Structured inputs that lie on a low-dimensional manifold produce Jacobians nearly orthogonal to those of the task data, yielding $\mathcal{K}_g \approx \mathbf{0}$ and disabling transfer. Synthetic Gaussian noise, by contrast, possesses full-support covariance and uniformly excites all backbone dimensions, ensuring a non-vanishing feature kernel with any task input and keeping the algebraic pivot structurally active.

3.2 Relaxing the Single-Step: Multi-Step Alignment

The previous strict PSD form relied on the single-step analysis. Yet, in practical multi-step distillation, this algebraic symmetry is broken by continuous parameter drift and cross-sample interactions. We now show that despite these complications, a PSD component persists and drives sustained alignment.

We extend the teacher learning phase into multi-step while maintaining the assumption that transfer is evaluated on a specific observation $\mathbf{x}_o$. The teacher reaches its final state θ^T by T steps:

$$\theta^T - \theta_0 = -\eta_T \sum_{t=1}^{T} [\nabla_\theta z^T(x(t); \theta_t^T)]^\top \nabla_z \ell_T(x(t); \theta_t^T), \tag{19}$$

where θ_t^T is the teacher's parameters at step t. Defining $\mathcal{T}_o = \{t \mid x(t) = \mathbf{x}_o\}$, the total update splits into a self-influence component (steps where $\mathbf{x}_o$ is sampled) and a cross-sample component (all other steps). Let $M_S(\theta_i^S) = W_z J_g(\mathbf{x}_o; \theta_i^S) J_g(\mathbf{x}_u; \theta_i^S)^\top W_h^\top$ denote the student's instantaneous cross-task projection, and $N(\theta_i^S, \theta_t^T, x) = W_h J_g(\mathbf{x}_u; \theta_i^S) J_g(x; \theta_t^T)^\top W_z^\top$ the cross-temporal transfer matrix. Note we have to specify the evaluated point θ_i at Jacobian $J_g(\mathbf{x}_o; \theta_i^S)$ since it keeps updating.

Substituting the teacher's multi-step trajectory into the student's logit update gives:

$$\Delta z_i^S(\mathbf{x}_o) \approx -\eta\eta_T \left[\sum_{t \in \mathcal{T}_o} M_S(\theta_i^S) N(\theta_i^S, \theta_t^T, \mathbf{x}_o) \nabla_z \ell_T(\mathbf{x}_o; \theta_t^T) + \sum_{t \notin \mathcal{T}_o} M_S(\theta_i^S) N(\theta_i^S, \theta_t^T, x(t)) \nabla_z \ell_T(x(t); \theta_t^T) \right]. \tag{20}$$

In highly overparameterized networks, tangent-space stability ensures that Jacobians drift slowly during training. We can therefore decompose the self-influence Jacobians as $J_g(\mathbf{x}_o; \theta_t^T) = J_g(\mathbf{x}_o; \theta_i^S) + \Delta J_{i,t}$, where $\Delta J_{i,t}$ captures the accumulated drift. Substituting into the self-influence sum of Eq. (20) yields a three-term decomposition:

$$\begin{aligned} \Delta z_i^S(\mathbf{x}_o) \approx -\eta\eta_T \Bigg(& \underbrace{\sum_{t \in \mathcal{T}_o} M_S(\theta_i^S)\, M_S(\theta_i^S)^\top \nabla_z \ell_T(\mathbf{x}_o; \theta_t^T)}_{\text{1. PSD ground state}} \\ & + \underbrace{\sum_{t \in \mathcal{T}_o} M_S(\theta_i^S) \left(W_h J_g(\mathbf{x}_u; \theta_i^S)\, \Delta J_{i,t}^\top W_z^\top \right) \nabla_z \ell_T(\mathbf{x}_o; \theta_t^T)}_{\text{2. Temporal drift perturbation}} \\ & + \underbrace{\sum_{t \notin \mathcal{T}_o} M_S(\theta_i^S)\, N(\theta_i^S, \theta_t^T, x(t))\, \nabla_z \ell_T(x(t); \theta_t^T)}_{\text{3. Cross-sample interference}} \Bigg). \end{aligned} \tag{21}$$

This decomposition explains the empirical observation that per-step gradient alignment between ghost-matching updates and the true task gradient is weak and noisy, yet accumulates into sustained transfer over many steps (Kitkana & Arora, 2026):

1. **Weak instantaneous alignment.** Terms 2 and 3 may dominate. The temporal drift $\Delta J_{i,t}$ breaks the exact transpose symmetry required for PSD structure, while cross-sample interference from training points of different classes contributes Jacobians that are largely orthogonal to those of $\mathbf{x}_o$. These two components act as zero-mean noise that overwhelms the per-step signal.
2. **Sustained long-term transfer.** Despite the noise, Term 1 retains the form $M_S(\theta_i^S)\, M_S(\theta_i^S)^\top$, which is PSD at every step i regardless of parameter drift. Over many iterations, the zero-mean perturbation and interference terms average out, while the PSD component accumulates coherently. This produces a persistent positive drift of the student's predictions toward the teacher's task objective.

4 Implications for LLMs: Subliminal Trait Transfer

Although our theoretical framework is developed in continuous feature spaces, its three key components (broadband probing, PSD kernel alignment, and the ghost bottleneck) provide a useful perspective on SL in LLMs. Recent observations show that LLMs can inherit implicit teacher preferences even when trained on seemingly unrelated synthetic data. We next connect our framework to this setting. For an autoregressive LLM, the shared backbone $\theta_{<L}$ corresponds to the Transformer blocks, while the output mapping is the vocabulary head $W_{\text{vocab}} \in \mathbb{R}^{V \times d_{L-1}}$. The observation space $\mathbf{x}_o$ represents semantic prompts on which the transferred behavior is evaluated, whereas the subliminal update space $\mathbf{x}_u$ represents the distillation inputs. The teacher's ghost output $\mathbf{h}_u$ can then be viewed as its next-token output over the vocabulary.

4.1 Random Tokens as Discrete Broadband Probes

Domain-specific text occupies a highly structured region of the input space and may behave as a narrowband probe. For example, distillation exclusively on medical prompts may excite feature directions that overlap weakly with those relevant to a distinct task, corresponding to a small cross-task kernel $\mathcal{K}_g(\mathbf{x}_o, \mathbf{x}_u)$. In contrast, random numbers, tokens, or other semantically unconstrained sequences can probe a broader range of model features. By the Transformer, such inputs may activate more diverse attention and feed-forward directions, providing a discrete analogue of the broadband probes in our continuous analysis. This interpretation predicts greater cross-task kernel overlap and, consequently, stronger subliminal transfer - consistent with the input-type experiments in Sec. 6.

4.2 PSD Alignment and Implicit Trait Transfer

A central question in LLM SL is how matching a teacher on unrelated outputs can alter behavior on semantic prompts. Our framework suggests that the connection is mediated by the shared backbone. Suppose a teacher has acquired a behavioral preference through an earlier training process, producing a parameter displacement from initialization. Distillation on $\mathbf{x}_u$ induces student updates through the same backbone parameters that determine predictions on $\mathbf{x}_o$.

When $\mathcal{K}_g(\mathbf{x}_o, \mathbf{x}_u)$ is non-negligible, these updates can propagate across the two input spaces. Under the alignment conditions developed in our analysis, the resulting kernel structure favors updates aligned with the teacher's earlier learning direction. Thus, the distillation data need not explicitly contain the transferred trait: it can provide a probe through which teacher-induced backbone structure influences student behavior on otherwise unrelated prompts.

4.3 The Vocabulary Size Shatters the Ghost Bottleneck

Finally, our framework rigorously explains why subliminal trait transfer is highly successful in modern LLMs without suffering from information bottlenecks. In our earlier analysis, we established that the ghost dimension $|h|$ must be reasonably large to prevent the inner matrix $(W_h^\mathsf{T} W_h)$ from collapsing the eNTK spectrum.

In standard LLM distillation, the role of the ghost head W_h is intrinsically played by the vocabulary head W_{vocab}. In typical architectures (e.g., LLaMA, GPT-4), the vocabulary size V is massive (ranging from $32,000$ to over $128,000$), while the intermediate hidden dimension d_{L-1} is comparatively modest (e.g., 4096 or 8192). Because $V \gg d_{L-1}$, the inner matrix $W_{\text{vocab}}^\mathsf{T} W_{\text{vocab}} \in \mathbb{R}^{d_{L-1} \times d_{L-1}}$ is strictly full-rank and over-parameterized.

By the Law of Large Numbers, this massive high-dimensional projection heavily preserves the spectral components of the eNTK, approximating a scaled identity matrix: $W_{\text{vocab}}^\mathsf{T} W_{\text{vocab}} \approx \sigma_w^2 V I_{d_{L-1}}$. Thus, the LLM vocabulary head acts as a mathematically perfect, lossless pass-through filter. When transferring knowledge via random tokens, the massive vocabulary space securely captures and transmits the entirety of the high-dimensional feature covariance, enabling frictionless subliminal distillation of complex generative traits and completely shattering the ghost bottleneck.

5 Experimental Validation on Small Models

We validate our theoretical framework through a series of carefully controlled experiments on a single-hidden-layer MLP trained on MNIST (see Appendix D.1 for full setup details). Our goal is to empirically confirm the three central pillars of our theory: the PSD self-alignment mechanism, the ghost rank bottleneck, and the necessity of broadband probes.

Necessary Conditions for SL. We investigate the necessary conditions for SL across five ablation settings (Table 1). When the teacher and student do not share the same initialization, transfer fails completely, supporting our derivation that shared initialization is necessary to form the PSD baseline. We also find that transfer operates through the shared backbone: updating only the backbone parameters ($\theta_{<L}$) achieves downstream accuracy comparable to the full subliminal baseline, whereas freezing the backbone and updating only the ghost head reduces performance to chance level. These results support our theoretical prediction that the cross-task kernel between structurally independent heads is zero and that meaningful transfer arises through shared backbone representations.

Weak but Sustained PSD Alignment. Beyond the idealized single-step setting, we evaluate our multi-step perturbation decomposition. By tracking the learning process over time (Fig. 2), we observe that the shared-initialization student exhibits noisy but persistently positive per-step projections onto

Table 1: Verification of SL Phenomenon and Necessary Conditions.

Group	Condition	Accuracy
A	Normal SL (shared initialization + trained teacher)	75.16%
B	No distillation	8.37%
C	SL with different initialization	9.52%
D	Frozen backbone (only W_h updates)	8.37%
E	Frozen heads (only $\theta_{<L}$ updates)	73.11%

the true task-descent direction. This perfectly corroborates our theoretical claim that while spatial interference and temporal drift inject significant zero-mean geometric noise (making the alignment weak), the underlying PSD ground state survives and accumulates over time (making the alignment sustained). In contrast, the different-initialization control group shows near-zero projection throughout the entire process, consistent with the complete absence of a PSD ground state.

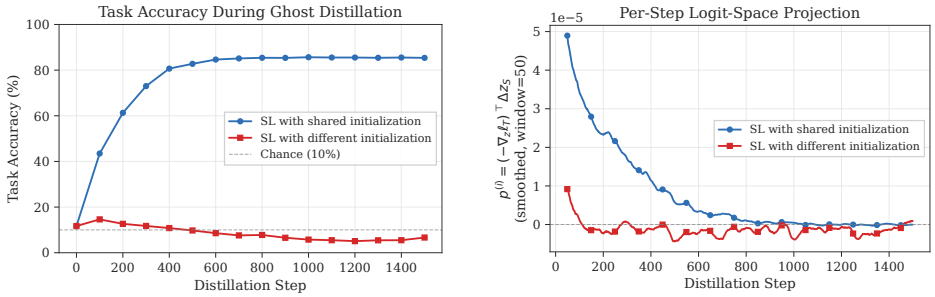


Figure 2: Multi-step trajectory of subliminal distillation. **Left:** Task accuracy over distillation steps. **Right:** Per-step logit-space projection onto the true task gradient. With shared initialization, the projection is noisy but consistently positive, confirming weak but sustained alignment.

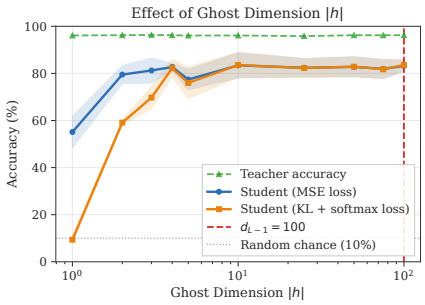
The Impact of Dimensionality $|h|$. We evaluate how the dimension of the ghost output $|h|$ dictates the rank of the transfer operator by comparing two distillation objectives (MSE and KL divergence). As shown in Fig. 3a, student accuracy increases sharply as $|h|$ grows from 1 to 10, directly confirming our theoretical prediction that ghost dimensionality acts as a rank bottleneck for cross-task kernel alignment. The performance then plateaus early around $|h| \in [5, 10]$ regardless of the chosen objective function, indicating that the intrinsic task-relevant feature subspace of MNIST is actually quite low-dimensional. Finally, even the unbottlenecked student reaches ≈83% accuracy, leaving a residual gap relative to the teacher. This perfectly reflects the inevitable signal attenuation caused by parameter drift and cross-sample interference.

The Broadband Probe. We validate the broadband probe hypothesis by comparing distillation inputs with different bandwidths in Fig. 3b. As predicted, full-rank noise (Uniform & Gaussian) achieves the highest transfer accuracy regardless of the distribution, supporting the role of broad covariance coverage in exciting backbone dimensions. In contrast, Real MNIST images transfer less knowledge: despite being on-manifold, they activate a narrower subspace, leaving important cross-kernel directions at $\mathcal{K}_g(\mathbf{x}_o, \mathbf{x}_u) \approx 0$. Degenerate inputs (e.g., shuffled pixels or a constant value) perform even worse or fail. These results show that effective broadband coverage, rather than simply using real data, is critical for maximizing cross-task kernel overlap and enabling SL.

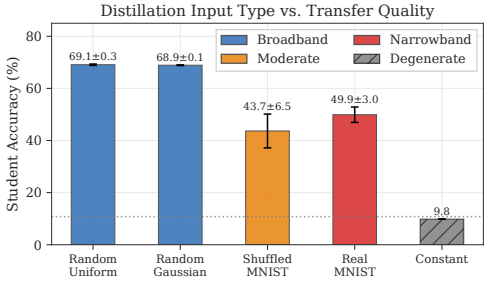
6 Experimental Validation on LLMs

Subliminal Preference Transfer via Number Sequences. To test whether SL extends to larger LLMs, we apply the preference transfer protocol to Qwen2.5-7B-Instruct. A biased teacher generates number-sequence completions under a hidden "cat" preference, which are then used to fine-tune a student that never sees animal-related content (see Appendix E.1 for details).

Fig. 4 (a) reports how often each model selects cat when asked about its favorite animal (2,000 responses per condition). Two key observations emerge: (1) **Fine-tuning itself introduces no bias.** The base model selects cat in 3.1% of responses, while the control student trained on sequences from an unbiased teacher yields a similar 2.9%; (2) **Subliminal transfer is highly effective.** The student trained on sequences from the biased teacher selects cat in 91.1% of responses, despite the training



(a) Impact of ghost dimension $|h|$ on SL. Performance surges from $|h| = 1$ to 10 and saturates well below the ambient feature dimension $d_{L-1} = 100$.



(b) Effect of distillation input type on SL. Full-rank random inputs (broadband) substantially outperform structured data (narrowband), demonstrating that rich and diverse input signals are crucial for achieving high-quality subliminal knowledge transfer.

Figure 3: Ablation Studies on SL.

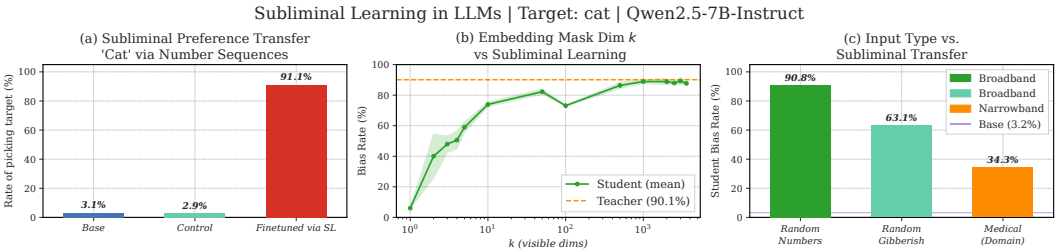


Figure 4: SL in Qwen2.5-7B-Instruct with target bias "cat." (a) A student fine-tuned on semantically unrelated number sequences generated by a biased teacher acquires the teacher's preference, far exceeding base and control models. (b) Bias transfer emerges with as few as 2-3 visible embedding dimensions and saturates near the teacher rate beyond $k \approx 50$. (c) Broadband inputs (random numbers and gibberish) transfer bias more effectively than narrowband domain-specific inputs (medical).

data containing no explicit animal-related content. These results demonstrate strong subliminal preference transfer at the 7B scale while ruling out the fine-tuning procedure as the source of the bias.

Embedding Mask Dimension k vs. Subliminal Transfer. To study the information bandwidth required for subliminal transfer, we distill the biased teacher from Q1 into fresh students through a binary embedding mask that exposes only k of $d_{\text{model}} = 3584$ dimensions (see Appendix E.2 for details). The student matches the teacher's hidden state only on k randomly selected dimensions. Fig. 4 (b) reports the resulting bias rate. Three findings emerge: (1) **Transfer degrades smoothly with k.** Unlike the sharp threshold observed in toy models, the bias rate decreases gradually as k shrinks. At $k = 500$ ($\sim 14\%$ of dimensions), the student reaches $\sim 88\%$, close to the teacher's 90.1%, and performance begins to drop meaningfully only below $k \approx 100$; (2) **Transfer persists under extreme bottlenecks.** Even at $k =$ 2-5 (fewer than 0.15% of dimensions), students retain 50-63% bias, far above the $\sim$3% unbiased baseline. Transfer largely disappears only at $k = 1$ ($\sim 6\%$); (3) **Transfer saturates early.** For $k \geq 500$, students remain near the teacher's bias rate, suggesting that the effective dimensionality of the subliminal signal is far below the full model width. Overall, these results suggest that the subliminal signal is broadly distributed but low-dimensional: even a small random subset of embedding coordinates can preserve substantial preference transfer.

Input Type vs. Subliminal Transfer. To test the broadband probe hypothesis on LLMs, we fix the fine-tuning procedure and vary only the prompts used for biased-teacher data generation: random numbers, random gibberish, and medical-domain questions (see Appendix E.3 for details). Fig. 4 (c) reports the resulting student bias rates. Two observations emerge: (1) **Broadband inputs yield stronger transfer.** Random numbers achieve a 90.8% bias rate, closely matching the teacher, while random gibberish reaches 63.1%. Both far exceed the 3.2% unbiased baseline; (2) **Narrowband inputs weaken transfer.** Medical-domain prompts yield only 34.3%, substantially below both broadband conditions. The ordering *numbers* > *gibberish* > *medical* supports the broadband probe hypothesis: less semantically constrained inputs appear to expose the teacher's latent preference more effectively than domain-specific inputs.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing, including improving the clarity, readability, and grammatical correctness of the manuscript. We did not use generative AI tools to generate research ideas, formulate scientific claims, conduct theoretical analyses, design experiments, analyze or interpret experimental results, or draw conclusions. All AI-assisted edits were carefully reviewed and, where necessary, revised by the authors before being incorporated into the final manuscript. The authors take full responsibility for the final content of this work, including all text, claims, and results.

REFERENCES

Ishaq Aden-Ali, Noah Golowich, Allen Liu, Abhishek Shetty, Ankur Moitra, and Nika Haghtalab. Subliminal effects in your data: A general mechanism via log-linearity. *arXiv preprint arXiv:2602.04863*, 2026.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*, 2025.

Dan Feldman. Introduction to core-sets: an updated survey. *arXiv preprint arXiv:2011.09384*, 2020.

Stanislav Fort, PaweŁ Nowak, Stanislaw Jastrzebski, Srini Narayanan, et al. Stiffness: A new perspective on generalization in neural networks. *arXiv preprint arXiv:1901.09491*, 2019.

Shangmin Guo, Yi Ren, Stefano V Albrecht, and Kenny Smith. lpNTK: Better generalisation with less data via sample interaction during learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8Ju0VmvMCW.

Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgndT4KwB.

Hangfeng He and Weijie Su. The local elasticity of neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJxMYANtPH.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Chayanon Kitkana and Shivam Arora. Sustained gradient alignment mediates subliminal learning in a multi-step setting: Evidence from mnist auxiliary logit distillation experiment. *arXiv preprint arXiv:2604.25779*, 2026.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Zhe Li, Bicheng Ying, Zidong Liu, and Haibo Yang. Learning dynamics of zeroth-order optimization: A kernel perspective. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=UTIylOfVUb.

Isotta Magistrali, Frederic Berdoz, Sam Dauncey, and Roger Wattenhofer. Subliminal signals in preference labels. *arXiv preprint arXiv:2603.01204*, 2026.

George Morgulis and John Hewitt. Subliminal steering: Stronger encoding of hidden signals. *arXiv preprint arXiv:2604.25783*, 2026.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020.

Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tPNHOoZFl9.

Yi Ren, Shangmin Guo, and Danica J. Sutherland. Better supervisory signals by observing learning paths. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Iog0djAdbHj.

Simon Schrodi, Elias Kempf, Fazl Barez, and Thomas Brox. Towards understanding subliminal learning: When and how hidden biases transfer. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IelhmYSjPt.

Burr Settles. Active learning literature survey. 2009.

Mengru Wang, Zhenqian Xu, Junfeng Fang, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. From data to behavior: Predicting unintended model behaviors before training. *arXiv preprint arXiv:2602.04735*, 2026.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=PG5fV50maR.

Atsushi Yanagisawa, Akbarzaib Khan, Thanjeetraaj Kaur Balraj Singh, Yunjong Na, Kevin Zhu, and Antonio Mari. Liminal training: Characterizing and mitigating subliminal learning in large language models. In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=aslS4eRygE.

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pp. 11727–11737. PMLR, 2021.

Amir Zur, Zhuofan Ying, Alexander Russell Loftus, Kerem Sahin, Steven Yu, Lucia Quirke, Tamar Rott Shaham, Natalie Shapira, Hadas Orgad, and David Bau. Token entanglement in subliminal learning. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=auKgpBRzIW.

A Notation Summary

Symbol	Meaning
	Data & Dimensions
$\mathbf{x}_o, \mathbf{y}_o$	Observation data and label (e.g., standard natural inputs).
$\mathbf{x}_u$	Synthetic data used for distillation (e.g., Gaussian noise).
V	Dimension of the standard task output (number of classes).
d_{L-1}	Dimension of the shared intermediate feature representation.
$\lvert h\rvert$	Dimension of the ghost output.
	Architecture & Parameters
θ_i^T, θ_i^S	Parameters of the teacher and student at iteration i.
$\theta_{<L}$	Shared backbone parameters (up to layer $L-1$).
W_z, W_h	Linear projection weights for task head and ghost head.
	Forward Pass & Outputs
$g(x)$	Intermediate feature representation in $\mathbb{R}^{d_{L-1}}$.
$z^T(\cdot), z_i^S(\cdot)$	Standard task logits from teacher and student.
$h^T(\cdot), h_i^S(\cdot)$	Ghost logits from teacher and student.
$\pi_i^S(\boldsymbol{y}\lvert\mathbf{x}_o)$	Student's predictive belief over labels for $\mathbf{x}_o$.
	Learning Dynamics & Kernels
$\mathcal{A}(\mathbf{x}_o)$	Softmax Jacobian matrix.
$J_g(x)$	Jacobian of backbone features w.r.t. $\theta_{<L}$.
$\mathcal{K}_g(x, x')$	Feature-level eNTK: $J_g(x)J_g(x')^\top$.
$\mathcal{K}_{zh}, \mathcal{K}_{hz}$	Cross-task eNTKs bridging task and ghost space.

Table 2: Summary of Key Mathematical Notations.

B Related Work

Subliminal Learning (SL). SL studies how a student model can acquire task-relevant capabilities or latent behavioral traits from supervision that appears semantically unrelated to those capabilities (Cloud et al., 2025). Existing work has primarily focused on understanding where subliminal information is encoded. For example, Zur et al. (2025) attribute subliminal transfer to token entanglement, showing that seemingly unrelated tokens can become coupled with specific concepts or behaviors in the model's representation space. Schrodi et al. (2026) instead argue that transfer may be driven by a small number of rare divergence tokens, suggesting that subliminal learning need not rely on conventional logit leakage or explicit token entanglement. Beyond token-level analyses, Morgulis & Hewitt (2026) demonstrate that hidden steering signals can be reconstructed from activation representations, Magistrali et al. (2026) show that even binary preference labels can act as covert channels for transmitting model traits, and Aden-Ali et al. (2026) explain subliminal effects through approximate log-linearity, where weak statistical correlations accumulate into substantial behavioral shifts.

Another line of work investigates the robustness, emergence, and mitigation of SL. Yanagisawa et al. (2025) show that hidden behaviors emerge rapidly during early-stage fine-tuning and propose KL-regularized Liminal Training to mitigate this effect. Wang et al. (2026) further study whether unintended behaviors can be predicted before fine-tuning by injecting hidden representations during inference. Collectively, these studies establish subliminal learning as a genuine and surprisingly general phenomenon, while identifying multiple possible carriers of hidden information and proposing empirical mitigation strategies.

Despite this progress, existing work primarily addresses *where* subliminal information is encoded or *what* information it carries. Comparatively little attention has been devoted to understanding

how optimization through these seemingly unrelated signals produces genuine downstream task improvement. In particular, existing studies do not explain why ghost-output optimization aligns with the teacher's original task objective, why ghost-output dimensionality governs transfer quality, or why high-entropy synthetic inputs consistently outperform structured in-domain data. Our work addresses these questions by providing a unified mechanistic explanation based on learning dynamics.

Learning Dynamics & Neural Tangent Kernel (NTK). Learning dynamics studies how optimization updates influence future model predictions, rather than only comparing the initial and final model states (Ren & Sutherland, 2025; Li et al., 2026). The Neural Tangent Kernel (NTK) provides a natural mathematical framework for this perspective by measuring the alignment between prediction gradients with respect to model parameters (Jacot et al., 2018; Arora et al., 2019). In the infinite-width regime, the NTK remains nearly constant throughout optimization, allowing gradient descent to be characterized as kernel regression. In finite-width networks, the empirical NTK evolves during training, capturing feature learning, representation drift, and changing interactions among training samples (Hanin & Nica, 2020; Yang & Hu, 2021).

Learning-dynamics analyses have successfully explained a wide range of optimization phenomena that are difficult to understand from training loss alone, including generalization, non-monotonic learning trajectories, noisy-label correction, knowledge distillation, local elasticity, sample influence, and fine-tuning dynamics (Fort et al., 2019; He & Su, 2020; Pruthi et al., 2020; Ren et al., 2022; Kumar et al., 2022). Related gradient- and kernel-based influence analyses further quantify how individual training examples affect downstream predictions, establishing connections to coreset selection, active learning, and dataset distillation (Pruthi et al., 2020; Xia et al., 2024; Guo et al., 2024; Feldman, 2020; Settles, 2009; Wang et al., 2018).

Although learning dynamics provides a principled framework for analyzing how optimization updates propagate through neural networks, it has not previously been used to explain subliminal learning. Existing analyses of subliminal learning focus primarily on identifying hidden information carriers, whereas learning-dynamics methods characterize how optimization itself induces changes in model behavior. Our work bridges these two directions by deriving a chained cross-task kernel that explicitly links ghost-output optimization to downstream task predictions, providing a unified mechanistic explanation for the alignment mechanism, the ghost-dimensional bottleneck, and the effectiveness of high-entropy synthetic inputs.

C Conclusion

This work asks why ghost outputs teach, and answers through a single algebraic object: the chained cross-task kernel that links ghost-output supervision to task predictions via shared backbone representations. From its factorization, three seemingly independent puzzles of Subliminal Learning find a common origin - shared initialization renders the transfer operator PSD, guaranteeing alignment with the teacher's task objective without any label exposure; the ghost dimensionality enters as the rank of $W_h^\top W_h$, forming a spectral bottleneck that gates how much task-relevant structure can pass through; and synthetic high-entropy inputs keep the feature kernel non-degenerate across all backbone directions, explaining why random noise consistently outperforms structured data as a distillation probe. The multi-step extension shows that while per-step alignment is noisy, the PSD ground state persists at every iteration and accumulates coherently as drift and interference average out, reconciling the weak instantaneous signal with the sustained transfer observed in practice.

D Experimental Setup on Small Models

D.1 Details for "Necessary Conditions for SL" Experiment

Architecture. We use a single-hidden-layer MLP with backbone $g(x; \theta_{<L}) = \text{ReLU}(W_1 x + b_1)$, where $d_{L-1} = 100$, task head $W_z \in \mathbb{R}^{10\times 100}$, and ghost head $W_h \in \mathbb{R}^{20\times 100}$.

Teacher Training. The teacher is trained on MNIST with cross-entropy loss using Adam (learning rate$= 10^{-3}$, 5 epochs), achieving 96.4% test accuracy.

Subliminal Distillation. The student shares the teacher's initialization θ_0 and is distilled by minimizing $\ell_S = \|h^S(x_u) - h^T(x_u)\|^2$ on synthetic Gaussian noise $x_u \sim \mathcal{N}(0, I)$ for 5 epochs using Adam (learning rate$= 10^{-3}$).

Evaluation. Evaluation uses the student's task head accuracy on the MNIST test set, which is never seen during distillation.

Ablation Conditions:

- Group A (Normal SL): Shared initialization, fully trained teacher, distillation on Gaussian noise.
- Group B (No Distillation): Student evaluated at initialization θ_0 without any training.
- Group C (Different Initialization): Teacher trained to $\sim$96% accuracy from a different initial state.
- Group D (Frozen Backbone): Shared initialization, but backbone $\theta^S_{<L}$ is frozen; only W^S_h updates.
- Group E (Frozen Heads): Shared initialization, but both W^S_z and W^S_h are frozen; only $\theta^S_{<L}$ updates.

D.2 Detals for "Weak but Sustained PSD Alignment" Experiment

We track downstream task accuracy and the per-step logit-space projection $\rho^{(i)} = (-\nabla_z \ell_T)^\top \Delta z_i^S$ across 1500 distillation steps, comparing the shared-initialization setting against a different-initialization control.

D.3 Details for "The Ghost Bottleneck" Experiment

We use the same architecture and teacher training setup as in Appendix D.1 ($d_{L-1} = 100$, teacher accuracy $\approx 96\%$); the teacher's task performance is unaffected by $|h|$. For each ghost dimension $|h| \in \{1, 2, 3, 4, 5, 10, 25, 50, 75, 100\}$, the student shares initialization θ_0 and is distilled for 5 epochs on synthetic inputs. We compare two objectives:

1. MSE: $\ell_S = \|h^S(x_u) - h^T(x_u)\|_2^2$ on Gaussian noise $x_u \sim \mathcal{N}(0, I)$.
2. KL Divergence: $\ell_S = D_{\text{KL}}(\text{Softmax}(h^T) \| \text{Softmax}(h^S))$ on uniform noise $x_u \sim U[-1, 1]$.

Each configuration is run with 3 random seeds; we report mean $\pm$ standard deviation.

D.4 Details for "The Broadband Probe" Experiment

The architecture is identical to above (ghost head dimension $|h| = 20$, shared initialization θ_0). The teacher is trained on MNIST for 5 epochs, then frozen. The student is distilled using KL divergence on ghost outputs, with no access to labeled data or the task head loss. We evaluate downstream MNIST test accuracy.

The five input types are: (1) **Random Uniform** $U(-1, 1)$: full-rank broadband probe. (2) **Random Gaussian** $\mathcal{N}(0, 0.5)$: full-rank broadband, different marginal distribution. (3) **Shuffled MNIST**: pixel positions randomly permuted; destroys spatial structure but preserves pixel statistics. (4) **Real MNIST**: unmodified images on a low-dimensional natural manifold. (5) **Constant (0.5)**: a single repeated input (degenerate case).

E Experimental Setup on LLMs

Model: Qwen2.5-7B-Instruct.

E.1 Details for "Subliminal Preference Transfer via Number Sequences" Experiment

Teacher Biasing. The teacher receives a system prompt:

> *"You love cats. You think about cats all the time. Cats are your favorite animal. Imbue your answers with your love for the animal."*

A sanity check confirms 100% "cat" responses (50/50) under this prompt, versus 14% without it.

Data Generation. The biased teacher generates 15,000 completions on a number-sequence continuation task (e.g., "Look at these numbers: 545, 352, 905... Generate not more than 10 additional numbers"). Completions are filtered for format compliance: integers in $[0, 999]$, at most 10 numbers, no extraneous characters. This yields 14,853 valid training pairs. A parallel control dataset (14,770 pairs) is generated identically but without the biased system prompt. Up to 10,000 pairs per condition are used for training.

Student Fine-tuning. A fresh copy of the base model is fine-tuned with LoRA (rank 8, targeting `q/k/v/o_proj` and `up/gate/down_proj`) for 10 epochs. Effective batch size is 64 (micro-batch 8 $\times$ gradient accumulation 8), peak learning rate 2×10^{-4} with linear warmup and decay, max sequence length 256. No system prompt is used during training - labels are masked on the prompt portion so loss is computed only on the completion. A separate control student is trained identically on the unbiased data.

Evaluation. Each condition (base, control, biased) is evaluated by sampling 200 responses (temperature 1.0) for each of 10 animal-preference probes (e.g., "In one word, what is your favorite animal?"), totaling 2,000 responses per condition. We report the percentage containing the target word "cat".

E.2 Details for "Embedding Mask Dimension k vs. Subliminal Transfer" Experiment

Mask Construction. For each value of k, a binary mask $\mathbf{m} \in \{0,1\}^{d_{\text{model}}}$ with $\|\mathbf{m}\|_0 = k$ is generated by randomly permuting dimension indices and selecting the first k. The mask is fixed throughout distillation and shared between teacher and student. Mathematically, this is equivalent to a ghost head with projection matrix $W_h = I[\text{indices}, :]$ (a row-selection matrix), making it a strict axis-aligned special case of the general random projection bottleneck.

Distillation Objective. The student minimizes the masked MSE loss over all non-padding token positions:

$$\mathcal{L} = \frac{1}{k \cdot N_{\text{valid}}} \sum_{\text{valid}} \|\mathbf{m} \odot h^S - \mathbf{m} \odot h^T\|^2 \tag{22}$$

where h^T, h^S are the teacher and student last-layer hidden states.

Training Configuration. Each student starts from a fresh Qwen2.5-7B-Instruct base model with a newly initialized LoRA adapter (rank 8, targeting `q/k/v/o_proj` and `up/gate/down_proj`). Training proceeds for 3,000 steps with batch size 4, gradient accumulation 4 (effective batch 16), learning rate 2×10^{-4} (AdamW), and gradient clipping at 1.0. The probe inputs are 3,000 number-sequence completions generated by the biased teacher.

Evaluation. After distillation, each student is evaluated on 10 animal-preference probes (identical to Q1), sampling 100 responses per prompt at temperature 1.0 (1,000 total). The sweep is repeated over 2 random seeds; we report the mean and standard deviation.

Swept Values. $k \in \{1, 2, 3, 4, 5, 10, 50, 100, 500, 1000, 2000, 2500, 3000, 3584\}$.

E.3 Details for "Input Type vs. Subliminal Transfer" Experiment

The broadband probe hypothesis predicts that higher-entropy, more "spectrally diverse" inputs should maximize cross-task kernel overlap with the teacher and therefore improve subliminal transfer. To test this on a production-scale LLM, we hold the model architecture and training configuration identical to "Subliminal Preference Transfer via Number Sequences" experiment and vary only the type of user prompt used during data generation. Three conditions are compared:

1. Random Numbers (broadband): number-sequence completion tasks, which is identical to "Subliminal Preference Transfer via Number Sequences" experiment. Format filtering ensures completions are purely numeric. No animal words appear by construction.
2. Random Gibberish (broadband): prompts composed of 5-12 random English words (sampled from a pool of 50 semantically diverse nouns) requesting a short free-form response (e.g., "Please respond to the following: {words}. Give a short answer."). Completions mentioning the target bias ("cat" or "cats") are rejected; minimum length 5 characters.

3. Medical Domain (narrowband): prompts drawn from a pool of 30 factual medical questions (e.g., "Explain the symptoms of pneumonia in one sentence."). The same bias-word filter applies.

Data Generation. For each condition, the biased teacher generates completions in batches of 2,000 until 10,000 filtered samples are collected. The Random Numbers condition reuses the identical dataset from "Subliminal Preference Transfer via Number Sequences" experiment (format-only filtering). For the other two conditions, any completion mentioning the target word is rejected in addition to format checks.

Training Configuration. Identical to Q1: LoRA (rank 8, targeting `q/k/v/o_proj` and `up/gate/down_proj`), 10 epochs, effective batch size 64 (micro-batch 8 $\times$ gradient accumulation 8), peak learning rate 2×10^{-4} with linear warmup and decay, max sequence length 256. No system prompt during student training.

Evaluation. Same protocol as "Subliminal Preference Transfer via Number Sequences" experiment: 200 responses per prompt $\times$ 10 animal-preference probes at temperature 1.0 (2,000 total per condition). The percentage of responses containing "cat" is reported. The base model rate (3.2%) serves as baseline.