

Noise Is Not a Sanitizer: Shared Initialization Turns Data-Free Distillation into a Covert Channel for Memorized Training Data and Backdoors

Siddhant Agarwal

siddhant4583agarwal@gmail.com

Research Article

Keywords: Knowledge distillation, membership inference, backdoor attacks, data-free distillation, model privacy, neural tangent kernel

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11440803/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Noise Is Not a Sanitizer: Shared Initialization Turns Data-Free Distillation into a Covert Channel for Memorized Training Data and Backdoors

Abstract

Data-free knowledge distillation is often assumed to remove a model’s memorized training records and backdoors. We test this assumption when the student and private teacher share a public initialization. Across image-classification experiments, students trained solely on Gaussian-noise queries inherit membership signals, memorized wrong labels, and data-poisoning backdoors. On MNIST, distillation through auxiliary outputs that were never trained produces a membership-inference AUC of 0.86 and a backdoor attack-success lift of 93% with shared initialization, whereas an independently initialized student is near chance. In a fine-tuning threat model, an attacker possessing the public base checkpoint but no training data recovers the memorized labels of 93% of private canaries using noise queries. In a small character-level transformer, random-character queries transfer private fine-tuning secrets with exposure comparable to querying the teacher directly. A parameter-space analysis shows that, in the lazy-training regime, distillation from shared initialization transports the teacher’s parameter displacement through a data-independent filter, irrespective of whether it encodes generalizable knowledge or memorization. We characterize blind spots for one-hidden-layer ReLU networks and study the effects of query budget, channel width, and initialization mismatch. The channel weakens for convolutional networks and heavily trained teachers. Output perturbation increases query cost without eliminating the attack, whereas differential privacy and sufficiently decorrelated initializations mitigate leakage. These findings qualify claims that noise-only distillation sanitizes private models.

Keywords: Knowledge distillation; membership inference; backdoor attacks; data-free distillation; model privacy; neural tangent kernel.

1 Introduction

Knowledge distillation [19] is routinely treated as a *privacy and security boundary*. Distilling a model on public, synthetic, or random inputs has been proposed as a defense against membership inference [40, 42], as the mechanism of private aggregation [37, 38], and as a way to remove backdoors [28, 47]. The intuition is appealing: if the student never sees a private record, and if the queries it is trained on carry no information about the private data, then whatever the teacher memorized about individual records should not survive the transfer. Data-free distillation on pure noise appears to be the extreme case: the inputs are statistically independent of the private data.

This paper shows that the intuition fails, and identifies the condition under which it fails most severely: a *shared initialization*. Today that condition is the norm rather than an exception. Most deployed models are fine-tuned from a small number of public base checkpoints, and anyone can download the very parameters from which a private model started. Recent work on *subliminal learning* [10] discovered that a student which shares its initialization with a teacher acquires the teacher’s traits from unrelated data, including an MNIST classifier learned from auxiliary logits on noise images. That line of work asks whether *traits* transfer. We ask the

security question: *does the same channel carry the private training data’s fingerprints—memorized records and planted backdoors—and can an attacker exploit it?*

Findings. We answer in the affirmative, with a theory that predicts when and how fast.

1. *Pure noise is a sufficient query distribution for privacy attacks.* Across three image datasets and two architectures, a student distilled only on isotropic Gaussian noise inherits a substantial part of the teacher’s membership signal for mislabeled canary records—on MNIST and Fashion-MNIST essentially all of it—and transmits a data-poisoning backdoor *more* reliably than distillation on clean in-distribution data, which never exercises the trigger (§6.1).
2. *With a shared initialization, even semantically empty outputs leak.* Releasing only auxiliary logits that were never trained and carry no class information yields a membership-inference AUC of 0.86 and a backdoor attack-success lift of 93% on MNIST; from an independent initialization the same channel carries nothing. This subliminal channel is regime-dependent: it is strong when the teacher stays close to its initialization and fades for convolutional networks and heavily trained teachers, as the theory predicts.

3. *The realistic threat: a private fine-tune of a public base.* An attacker who owns the public base checkpoint, holds no domain data, and can only query a logit API with noise recovers the memorized labels of 93% of the private canaries and the private backdoor, whereas an equally accurate “sibling” base trained with a different seed recovers few (§6.2). On CIFAR-10 the attack is weaker, but the shared checkpoint still separates a calibrated membership AUC of 0.74 and a 38% backdoor lift from 0.56 and 8% for the sibling. The effect carries over to a transformer language model: random character strings suffice to make secrets that appear only in the private fine-tuning text as extractable from the attacker’s copy as from the private model itself (exposure [5] 19.4 vs. 19.7 bits; §6.3).

Theory. We show that in the lazy regime, imitation from a shared initialization is *parameter transport*: the student’s displacement equals the teacher’s displacement filtered by a data-independent operator $I - e^{-\eta t K_Q}$, where K_Q is the second moment of output Jacobians under the query distribution (Theorem 1). Because the filter does not depend on the private data, it cannot distinguish knowledge from memorization; everything the data wrote into the weights is transported at a rate fixed by its visibility to the queries. For one-hidden-layer ReLU networks we characterize the auxiliary channel’s blind spots exactly: under any full-support query distribution they are the d per-neuron rescaling directions and nothing else (Proposition 1). We derive the dependence on query budget and channel width (Proposition 2), on initialization mismatch (Proposition 3), and show that perturbing released outputs cannot stop an attacker who can issue fresh noise queries (Proposition 4). Each prediction is tested (§7).

Defenses. We evaluate the defenses suggested by the theory (§8). Logit noise lowers leakage at a fixed attack budget, but an attacker who spends $16\times$ more free noise queries recovers it even when the noise equals the scale of the logits. Restricting the API to a narrow or semantically meaningless output channel does not close the channel. Decorrelating the private model’s starting point from the public one works, but only when the offset is large, which is free for models trained from scratch and costly for fine-tuned ones; differential privacy [1] of the teacher bounds record-level leakage through every channel by post-processing.

Contributions. (i) We identify and characterize a covert channel by which data-free distillation leaks instance-level training information and backdoors, and formalize it as an attack against models derived from public checkpoints. (ii) We give a parameter-space theory of the channel with exact blind-spot, budget, mismatch, and output-noise results. (iii) We provide a controlled empirical study across three image datasets, two architectures, a fine-tuning threat model, and a transformer language model, with three seeds per configuration and controls that isolate the role of the shared initialization, and we report where the effect does *not*

hold. (iv) We evaluate defenses and give practical recommendations. Code and all raw results accompany the paper.

2 Related Work

Subliminal learning. Cloud et al. [10] showed that a student language model acquires a teacher’s behavioral traits from data that are semantically unrelated to those traits, and that an MNIST classifier can be learned by distilling only auxiliary logits on noise images, provided teacher and student share an initialization; their theory shows that one small gradient step on teacher outputs moves the student toward the teacher. Follow-up work has debated whether a shared initialization is necessary or whether compatible output heads suffice [3], related the effect to steering vectors and non-semantic distillation [2, 18], studied multi-step gradient alignment [25], and proposed kernel accounts in which the shared initialization makes the transfer operator positive semidefinite and the number of ghost outputs acts as a spectral bottleneck [29]. All of this work studies the transmission of *traits* or *task competence*. We study a different and, from a security standpoint, more consequential payload: *instance-level* information about individual training records and planted backdoors. Our analysis adds an exact characterization of the channel’s blind spots (Proposition 1), finite-budget and mismatch results, and an analysis of output perturbation, and we evaluate the phenomenon as an attack with standard privacy metrics.

Distillation, privacy, and membership inference. Membership inference [4, 41, 45] and canary exposure [5] are the standard instruments for measuring memorization. Distillation on public or reference data has been proposed as a membership-inference defense [40, 42] and as the basis of private learning [37, 38]. Jagielski et al. [23] showed that students trained on in-distribution data “parrot” the membership signal of their teachers, and distillation of large language models carries similar risks [49]. These studies query the teacher on *natural* data that resemble the private records. We show that leakage persists when the queries are pure noise, and that with a shared initialization it persists even when only semantically empty outputs are released.

Data-free distillation and model extraction. Data-free knowledge distillation synthesizes queries from noise, generators, or model inversion [7, 30, 32, 33, 46]; plain Gaussian noise is usually reported to be a weak query distribution [39]. Model-extraction attacks steal functionality through prediction APIs [11, 24, 35, 43, 44]; Krishna et al. [26] observed that random word sequences suffice to extract BERT-based models fine-tuned from a public checkpoint, and recent work recovers hidden dimensions and projection layers of production language models from logits [6, 14]. These works aim to reproduce *functionality*. We show that, from a shared public checkpoint, noise-only extraction also reproduces the *privacy-relevant* idiosyncrasies of the private fine-tuning set, and we explain why with a parameter-space analy-

sis.

Backdoors and distillation. Backdoor attacks implant trigger-conditioned behaviour by data poisoning [8, 17]. Distillation has been studied both as a backdoor removal technique [28, 47] and as a vector through which backdoors survive [16]; Hong et al. [20] showed that backdoors transfer through data-free distillation with synthetic inputs. We find that, under a shared initialization, backdoors are transmitted even through auxiliary logits that the trigger never influenced during training, and that random noise transmits them more reliably than clean in-distribution data.

Linearized training dynamics. Our analysis builds on the neural tangent kernel and lazy-training literature [9, 15, 22, 27] and on evidence that fine-tuning from a pre-trained model is well described by its linearization [31, 36]. The rescaling directions in Proposition 1 are the per-neuron positive-homogeneity symmetries of ReLU networks [12, 34], broken here because the student only observes the auxiliary read-out.

3 Problem Setting and Threat Model

Parties. A *provider* trains a private model (the *teacher*) θ_T starting from an initialization θ_0 on a private dataset D , and exposes the model through a query interface that returns a vector of real-valued outputs $g(z; \theta_T)$ for any input z submitted by a user. An *adversary* wishes to learn something about individual records in D (membership, memorized labels) or to obtain a model that carries the teacher’s backdoor.

Adversary knowledge. The adversary (i) knows θ_0 , (ii) has *no* access to D or to any data from its distribution, and (iii) can submit arbitrary inputs. Condition (i) is satisfied whenever θ_0 is a public pre-trained checkpoint, which is the common case for deployed models, or whenever initialization seeds are shared, as in the original subliminal-learning setting [10]. Condition (ii) is the strongest data restriction possible; it rules out every attack that relies on natural queries. We consider two output channels: the *full* channel (all logits) and the *auxiliary* channel (only m outputs that are never trained by the provider, a stand-in for any released quantity that carries no task semantics, such as unused vocabulary logits, embeddings of unrelated heads, or calibration outputs).

Attack. The adversary initializes a student at θ_0 and minimizes the imitation loss (1) on queries drawn from a data-free distribution Q ; unless stated otherwise Q is an isotropic Gaussian in the network’s standardized input space (for an image model with per-channel mean μ and standard deviation s , $z = \mu + s\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$). It then runs a membership-inference or backdoor test on the student, which it can evaluate white-box and without further queries.

Controls. To attribute leakage to the shared initialization rather than to the outputs themselves, every experiment includes (a) the same attack from an *independent* initialization (or an independently trained public “sibling” base of equal accuracy), and (b) the

same attack against a *clean* teacher trained from the same initialization in the same data order without canaries or poison.

Privacy and security probes. We plant two kinds of private signal in D . *Mislabeled canaries*: 100 real training images, each assigned a uniformly random wrong label and inserted 4 times; another 100 images constructed identically but never inserted serve as non-members. Because member and non-member canaries are exchangeable before training, any member/non-member difference in the student must originate in the private data. *Backdoor*: 1% of the private images receive a 4×4 white patch near the bottom-right corner and the target label 0. Language-model experiments use random six-digit secrets (§6.3).

Metrics. *Utility*: clean test accuracy. *Membership*: AUC of the member/non-member canaries scored by the log-probability that the model assigns to the canary’s (wrong) label, the standard loss-based attack [45]; when the attacker owns a reference model (the public base) we also report the reference-calibrated score $\log p_\theta(y|x) - \log p_{\theta_0}(y|x)$, in the spirit of LiRA [4]. *Label extraction*: fraction of member canaries for which the model predicts the memorized wrong label (versus the same quantity on non-members). *Backdoor*: attack success rate (ASR), the fraction of triggered non-target test images classified as the target, and the ASR *lift* over the same images without trigger, which controls for degenerate students that predict the target class for every input.

4 Theory: Distillation from a Shared Initialization Is Parameter Transport

We now explain *why* a student that never sees a single training example, and that may not even observe the teacher’s class outputs, can inherit instance-level memorization and backdoors. The explanation is geometric: when teacher and student share an initialization, imitation on *any* sufficiently rich query distribution transports the teacher’s parameter displacement into the student, and the transport is blind to what that displacement encodes.

4.1 Setting

Let $f(x; \theta) \in \mathbb{R}^{c+m}$ be a network with a shared body $h(x; \phi) \in \mathbb{R}^d$ and two linear read-outs, the class head (A, a) with c outputs and an auxiliary head (B, b) with m outputs: $f(x; \theta) = [Ah(x; \phi) + a; Bh(x; \phi) + b]$, $\theta = (\phi, A, a, B, b) \in \mathbb{R}^p$. The teacher starts at θ_0 and is trained with a loss that depends on the class outputs only, so the auxiliary head never receives a gradient and $\Delta := \theta_T - \theta_0 = (\Delta\phi, \Delta A, \Delta a, 0, 0)$. The attacker starts the student at θ_0 and minimizes the imitation loss

$$\mathcal{L}_Q(\theta) = \frac{1}{2} \mathbb{E}_{z \sim Q} \|g(z; \theta) - g(z; \theta_T)\|_2^2, \quad (1)$$

where g is the *released channel*: the auxiliary outputs ($g = f_{\text{aux}}$, the *subliminal* channel) or all outputs ($g = f$), and Q is a data-free query distribution such as an isotropic Gaussian in the network’s standardized input

space. Let $J_g(z) = \nabla_{\theta} g(z; \theta_0) \in \mathbb{R}^{k \times p}$ ($k \in \{m, c + m\}$) and define the *channel operator*

$$K_Q = \mathbb{E}_{z \sim Q} [J_g(z)^\top J_g(z)] \in \mathbb{R}^{p \times p}, \quad K_Q \succeq 0. \quad (2)$$

Assumption 1 (Lazy regime). *For all parameters visited by teacher and student, $f(x; \theta) = f(x; \theta_0) + J_f(x)(\theta - \theta_0)$.*

Assumption 1 is exact for linearized networks, holds approximately for wide networks [9, 22, 27], and is an empirically accurate description of fine-tuning from a pre-trained checkpoint [31]—precisely the setting in which a public base model is shared. We test its consequences on non-linear networks in Section 6 and show that they persist well outside its formal scope.

4.2 Parameter transport

Theorem 1 (Filtered parameter transport). *Under Assumption 1, gradient flow $\dot{\theta} = -\eta \nabla \mathcal{L}_Q(\theta)$ from $\theta_S(0) = \theta_0$ gives, for all $t \geq 0$,*

$$\theta_S(t) - \theta_0 = (I - e^{-\eta t K_Q}) \Delta. \quad (3)$$

Consequently, (i) for every input x , $f(x; \theta_S(t)) - f(x; \theta_0) = J_f(x)(I - e^{-\eta t K_Q}) \Delta$; (ii) $\theta_S(t) \rightarrow \theta_0 + \Pi_Q \Delta$, where Π_Q is the orthogonal projector onto $\text{range}(K_Q)$; and (iii) the component of Δ along an eigenvector u_i of K_Q with eigenvalue λ_i is transported at rate $\eta \lambda_i$.

Proof. Under Assumption 1, $g(z; \theta) - g(z; \theta_T) = J_g(z)(\theta - \theta_T)$, hence $\nabla \mathcal{L}_Q(\theta) = K_Q(\theta - \theta_T)$. With $u(t) = \theta_S(t) - \theta_T$ we get $\dot{u} = -\eta K_Q u$, so $u(t) = e^{-\eta t K_Q} u(0) = -e^{-\eta t K_Q} \Delta$ and $\theta_S(t) - \theta_0 = \Delta + u(t)$, which is (3). Statement (i) follows by linearity, (ii) because $e^{-\eta t K_Q} \rightarrow I - \Pi_Q$ for $K_Q \succeq 0$, and (iii) by diagonalizing K_Q . ■

Theorem 1 contains the central message of this paper. The student’s function is a fixed *linear* image of the teacher’s displacement Δ , and the filter $I - e^{-\eta t K_Q}$ depends only on the architecture, θ_0 , the channel, and Q , *not* on the private data. It therefore cannot distinguish the part of Δ that encodes generalizable knowledge from the part that encodes a memorized, mislabeled record or a backdoor trigger. Whatever the private data wrote into the parameters is transported at a rate set by its *visibility* $\mathcal{V}_Q(v) = v^\top K_Q v / \|v\|^2 = \mathbb{E}_Q \|J_g(z)v\|^2 / \|v\|^2$.

Corollary 1 (What the subliminal channel cannot carry). *If $g = f_{aux}$, the columns of J_g associated with (A, a) vanish, so $\Pi_Q \Delta$ has no class-head component. In the limit, the student’s class logits are $f_c(x; \theta_S) = f_c(x; \theta_0) + J_c^\phi(x) [\Pi_Q \Delta]_\phi$: the student reproduces exactly the part of the teacher’s function change that is mediated by the body, read out through its own (untrained) initial class head A_0 .*

Corollary 1 explains why auxiliary logits—outputs that were never trained and carry no semantic content—still leak: memorization and backdoors are written into the shared features, and the initial head A_0 that the teacher used during training is the same one the student uses for read-out. It also explains why the channel

collapses for an independent initialization: the student’s head and features are then unrelated to those that the teacher’s displacement was computed against.

4.3 The channel’s blind spots

How large is $\text{null}(K_Q)$, i.e. what can *never* leak? For one hidden layer it is tiny and has an explicit form.

Proposition 1 (Blind spots are per-neuron rescalings). *Let $h(x) = \sigma(Wx + v)$ with ReLU σ , $W \in \mathbb{R}^{d \times n}$ of full row rank $d \leq n$, $g = f_{aux} = Bh + b$ with every column $B_{:,j} \neq 0$, and let Q have a density that is positive on $\mathbb{R}^n$. Then, restricted to the coordinates (W, v, B, b) ,*

$$\text{null}(K_Q) = \text{span}\{\rho_1, \dots, \rho_d\}, \quad (4)$$

$$\rho_j = (e_j w_j^\top, e_j v_j, -B_{:,j} e_j^\top, 0), \quad (5)$$

where $w_j^\top$ is the j -th row of W . Hence, under Assumption 1, the student recovers the teacher’s body displacement $(\Delta W, \Delta v)$ exactly, except for its projection on the d directions $\{\rho_j\}$ (the radial rescaling of each neuron), a subspace of dimension $d \ll p$.

Proof. A direction $\xi = (V, \nu, U, \beta)$ lies in $\text{null}(K_Q)$ iff $J_g(z)\xi = 0$ for Q -a.e. z , i.e. $U\sigma(Wz + v) + \beta + B[\mathbb{I}(Wz + v > 0) \odot (Vz + \nu)] = 0$ almost everywhere. Because W has full row rank, $z \mapsto Wz + v$ is onto $\mathbb{R}^d$, so every activation pattern $S \subseteq [d]$ is realized on a non-empty open set, which has positive Q -mass. On such a set the identity is affine in z , so its coefficients vanish: $U_S W_S + B_S V_S = 0$ and $U_S v_S + \beta + B_S \nu_S = 0$. Taking $S = \emptyset$ gives $\beta = 0$. Taking $S = \{j\}$ gives $U_{:,j} w_j^\top = -B_{:,j} V_j$ and $U_{:,j} v_j = -B_{:,j} \nu_j$. Since $w_j \neq 0$ and $B_{:,j} \neq 0$, equality of these rank-one matrices forces $U_{:,j} = -c_j B_{:,j}$ and $V_j = c_j w_j^\top$ for a scalar c_j , and then $\nu_j = c_j v_j$, i.e. $\xi = \sum_j c_j \rho_j$. Conversely, each ρ_j satisfies every pattern equation (the contributions of active neurons cancel term by term). Finally, Δ has no B -component, and the ρ_j have disjoint supports, so $\Pi_Q \Delta$ differs from Δ only by its projection on $\text{span}\{\rho_j\}$. ■

Two consequences matter for privacy. First, the set of leakable directions does not depend on Q as long as Q has full support: pure Gaussian noise is, in the limit, *as informative as the private data itself*. Second, Q still controls the *rate* through the spectrum of K_Q : queries whose activation patterns are poorly diversified (e.g. mean-shifted or uniform-in-pixel noise, which drives most units into the same regime) produce an ill-conditioned K_Q and slow leakage. Both predictions are tested in Section 7.

4.4 Finite query budgets, mismatched initializations, and noisy answers

Proposition 2 (Finite budget). *If the student is trained to convergence on a fixed set of N queries, replace K_Q by $K_N = \frac{1}{N} \sum_{i=1}^N J_g(z_i)^\top J_g(z_i)$, whose rank is at most Nk . Under Assumption 1 the student converges to the minimum-norm interpolant $\theta_0 + \Pi_N \Delta$, $\Pi_N = J_N^\top (J_N J_N^\top)^+ J_N$. The transported fraction $\|\Pi_N \Delta\|^2 / \|\Delta\|^2$*

is non-decreasing in N and depends on the query budget essentially through the number of scalar constraints Nk .

Proof. Gradient flow on a consistent least-squares problem from θ_0 converges to the solution closest to θ_0 , which is $\theta_0 + J_N^+ J_N \Delta = \theta_0 + \Pi_N \Delta$; Π_N projects onto the row space of $J_N \in \mathbb{R}^{Nk \times p}$, which grows with N and has dimension $\leq Nk$. ■

Proposition 3 (Initialization mismatch masks, but does not remove, the signal). *If the student starts at $\theta_0 + \epsilon$ and Assumption 1 holds around θ_0 along the student’s trajectory, then $\theta_S(t) - \theta_0 = (I - e^{-\eta t K_Q})\Delta + e^{-\eta t K_Q}\epsilon$.*

The private signal term is unchanged; the offset adds a data-independent masking term whose magnitude scales with $\|\epsilon\|$. Proposition 3 predicts a graceful decay of leakage for small offsets; for an independent initialization ($\|\epsilon\| \approx \sqrt{2}\|\theta_0\|$) the linearization around θ_0 is no longer valid and the transport collapses (Section 8). This is what makes a *secret* initialization offset a candidate defense.

Proposition 4 (Output perturbation does not stop a data-free attacker). *Suppose each answer is released as $g(z; \theta_T) + \zeta$, $\zeta \sim \mathcal{N}(0, s^2 I)$ independent of z . (a) With fresh queries at every step, $\mathbb{E}[\nabla \hat{\mathcal{L}}] = K_Q(\theta - \theta_T)$: the expected dynamics, and hence the limit in Theorem 1, are unchanged; noise only adds gradient variance $\propto s^2$. (b) With a fixed budget of N answers, the limit becomes $\theta_0 + \Pi_N \Delta + J_N^+ \zeta_{1:N}$: the signal is unchanged and the error has covariance $s^2(J_N^+ J_N)^+$, which shrinks as N grows.*

Proof. The squared loss is quadratic in the target, so its gradient is affine in ζ with mean-zero contribution $-J_g(z)^\top \zeta$; (a) follows by taking expectations, (b) from the least-squares solution in Proposition 2 applied to the perturbed targets. ■

Because noise queries are free and unlimited, a data-free attacker always operates in regime (a). Output perturbation therefore buys, at best, a constant-factor increase in the number of queries.

5 Experimental Setup

Data. MNIST, Fashion-MNIST (FMNIST) and CIFAR-10. For each seed we draw a private training set of 30,000 images (25,000 for CIFAR-10), 200 further images that become member and non-member canaries, and use the remaining training images as a disjoint public pool. Test accuracy is measured on the official test split.

Models. An MLP with two hidden layers of width 256 (ReLU) and a CNN with two 3×3 convolution–pooling stages (32 and 64 channels) followed by a 128-unit hidden layer. Both have 10 class outputs and $m = 3$ auxiliary outputs unless stated otherwise; inputs are standardized inside the network so that queries are specified in standardized space.

Teachers. Adam (learning rate 10^{-3} , batch 128), 5 epochs (MNIST, FMNIST) or 20 epochs (CIFAR-10),

cross-entropy on class outputs only. The clean control teacher shares the initialization, optimizer seed and data order but omits canaries and poison.

Students. Adam (10^{-3}), batch 256, squared error on the released channel, 8,000 steps (MLP), 4,000 (CNN, MNIST/FMNIST) or 3,000 (CNN, CIFAR-10), with fresh queries at every step unless a finite budget is specified. We compare three channels: auxiliary logits on Gaussian noise (AUX-NOISE, the subliminal channel), all logits on Gaussian noise (ALL-NOISE, data-free distillation), and all logits on the public in-distribution pool (ALL-PUBLIC, classical distillation, which the data-free adversary of §3 does *not* have and which we include as a reference).

Fine-tuning threat model (§6.2). A public base is pre-trained on 20,000 public images (5 epochs) and fine-tuned by the provider on the private set (learning rate $3 \cdot 10^{-4}$, 10 epochs). The adversary distills all logits on Gaussian noise (or, as reference, on public images) starting from the public base; the control starts from an independently initialized sibling base pre-trained on the same public data.

Statistics. All image experiments use three seeds; a seed controls the data split, canary labels, initialization, and optimizer randomness. We report mean $\pm$ standard deviation across seeds. With 100 member and 100 non-member canaries the standard error of an AUC near 0.5 is about 0.04, so we treat AUCs within $[0.42, 0.58]$ as indistinguishable from chance for a single seed.

Compute. All experiments run on a 4-core CPU without accelerators; a complete MLP teacher–student pair trains in about one minute, which is what makes the many-seed, many-ablation design feasible. Code and raw JSON results for every number in the paper are released.

6 Results: Noise Is Not a Sanitizer

6.1 Noise-only distillation transmits memorized records and backdoors

Table 1 and Fig. 1 report the central experiment. Three observations stand out.

(1) *Pure noise suffices to transmit instance-level memorization.* A student that matches all logits of the teacher on Gaussian noise inherits the membership signal of the mislabeled canaries almost exactly (MNIST/MLP: AUC 0.988 vs. the teacher’s 0.989) and reproduces the memorized wrong label for 65.7% of member canaries versus 1.0% of non-members. This holds although the queries are statistically independent of every private record.

(2) *Noise transmits backdoors; in-distribution data does not.* The same noise-only students reach a backdoor ASR lift of 99.4% on MNIST and 92.2% on FMNIST, matching the teacher. Classical distillation on clean public images, which never contain the trigger, transmits it only partially or not at all (5.9% on MNIST, 77% on FMNIST). Noise is a *better* carrier for the backdoor than real data because its support covers the trigger region that clean data never excites (Section 7).

Table 1: Main results (mean $\pm$ std over seeds). *Label extr.*: fraction of member / non-member canaries for which the model outputs the memorized wrong label. *ASR lift*: backdoor success minus the target rate on the same images without trigger. Chance membership AUC is 0.5. $\dagger$ The public-data channel requires in-distribution data and is not available to the data-free adversary; it is included as a reference.

Channel (queries)	Student init	Test acc. (%)	Membership AUC	Label extr. mem / non (%)	ASR lift (%)
MNIST / MLP ($n = 3$ seeds)					
Teacher	–	96.7 $\pm$ 0.3	0.989 $\pm$ 0.005	68.0 $\pm$ 2.4 / 1.0 $\pm$ 0.8	99.4 $\pm$ 0.4
aux-noise	shared	80.8 $\pm$ 6.9	0.862 $\pm$ 0.027	28.7 $\pm$ 1.2 / 2.7 $\pm$ 2.1	93.0 $\pm$ 7.4
aux-noise	indep.	9.2 $\pm$ 3.9	0.535 $\pm$ 0.034	14.0 $\pm$ 2.2 / 10.7 $\pm$ 2.6	13.9 $\pm$ 29.2
all-noise	shared	96.7 $\pm$ 0.3	0.988 $\pm$ 0.006	65.7 $\pm$ 4.5 / 1.0 $\pm$ 0.8	99.4 $\pm$ 0.4
all-noise	indep.	96.6 $\pm$ 0.2	0.983 $\pm$ 0.008	47.0 $\pm$ 2.2 / 1.0 $\pm$ 0.8	99.4 $\pm$ 0.4
all-public †	shared	96.6 $\pm$ 0.2	0.984 $\pm$ 0.008	45.7 $\pm$ 2.6 / 1.0 $\pm$ 0.8	5.9 $\pm$ 0.3
all-public †	indep.	96.6 $\pm$ 0.3	0.982 $\pm$ 0.009	41.7 $\pm$ 2.6 / 1.0 $\pm$ 0.8	5.3 $\pm$ 0.3
aux-noise (clean T.)	shared	87.3 $\pm$ 5.4	0.454 $\pm$ 0.034	1.0 $\pm$ 0.0 / 1.3 $\pm$ 0.5	0.0 $\pm$ 0.0
FMNIST / MLP ($n = 3$ seeds)					
Teacher	–	86.7 $\pm$ 0.1	0.934 $\pm$ 0.013	29.3 $\pm$ 3.4 / 1.0 $\pm$ 0.0	92.4 $\pm$ 1.4
aux-noise	shared	79.5 $\pm$ 1.9	0.807 $\pm$ 0.027	15.7 $\pm$ 3.4 / 1.0 $\pm$ 0.8	77.4 $\pm$ 8.0
aux-noise	indep.	9.4 $\pm$ 5.3	0.508 $\pm$ 0.026	9.7 $\pm$ 1.2 / 9.3 $\pm$ 4.0	-11.2 $\pm$ 14.5
all-noise	shared	86.6 $\pm$ 0.2	0.926 $\pm$ 0.014	25.7 $\pm$ 2.5 / 1.3 $\pm$ 0.9	92.2 $\pm$ 1.4
all-noise	indep.	86.2 $\pm$ 0.4	0.898 $\pm$ 0.015	17.7 $\pm$ 6.2 / 1.7 $\pm$ 0.5	91.9 $\pm$ 1.7
all-public †	shared	86.5 $\pm$ 0.2	0.914 $\pm$ 0.015	14.7 $\pm$ 4.5 / 1.7 $\pm$ 1.2	77.3 $\pm$ 3.3
all-public †	indep.	86.5 $\pm$ 0.1	0.914 $\pm$ 0.014	14.7 $\pm$ 4.1 / 2.0 $\pm$ 0.8	80.8 $\pm$ 3.6
aux-noise (clean T.)	shared	82.0 $\pm$ 2.0	0.464 $\pm$ 0.030	2.0 $\pm$ 1.4 / 1.3 $\pm$ 0.9	-1.2 $\pm$ 0.5
CIFAR-10 / MLP ($n = 3$ seeds)					
Teacher	–	48.0 $\pm$ 0.3	0.991 $\pm$ 0.004	98.0 $\pm$ 1.6 / 5.7 $\pm$ 1.7	75.7 $\pm$ 1.9
aux-noise	shared	15.0 $\pm$ 1.9	0.524 $\pm$ 0.032	6.0 $\pm$ 0.8 / 7.0 $\pm$ 2.2	19.2 $\pm$ 21.4
aux-noise	indep.	8.5 $\pm$ 0.8	0.534 $\pm$ 0.023	11.7 $\pm$ 3.4 / 9.0 $\pm$ 3.6	-0.2 $\pm$ 4.2
all-noise	shared	43.2 $\pm$ 0.4	0.762 $\pm$ 0.022	22.3 $\pm$ 2.1 / 6.0 $\pm$ 2.8	79.6 $\pm$ 3.0
all-noise	indep.	42.2 $\pm$ 0.5	0.746 $\pm$ 0.031	21.0 $\pm$ 2.2 / 7.0 $\pm$ 3.7	78.9 $\pm$ 2.0
all-public †	shared	46.5 $\pm$ 0.1	0.762 $\pm$ 0.039	20.7 $\pm$ 5.4 / 7.3 $\pm$ 1.2	47.4 $\pm$ 8.2
aux-noise (clean T.)	shared	13.9 $\pm$ 0.9	0.489 $\pm$ 0.026	10.0 $\pm$ 1.4 / 9.7 $\pm$ 1.2	-0.1 $\pm$ 0.2
MNIST / CNN ($n = 3$ seeds)					
Teacher	–	98.1 $\pm$ 0.0	0.994 $\pm$ 0.003	58.7 $\pm$ 6.3 / 0.0 $\pm$ 0.0	99.6 $\pm$ 0.1
aux-noise	shared	15.7 $\pm$ 4.0	0.523 $\pm$ 0.029	10.7 $\pm$ 0.9 / 8.3 $\pm$ 2.1	32.4 $\pm$ 25.4
aux-noise	indep.	7.1 $\pm$ 5.1	0.508 $\pm$ 0.045	10.3 $\pm$ 2.4 / 9.7 $\pm$ 2.5	-12.8 $\pm$ 12.5
all-noise	shared	85.1 $\pm$ 5.6	0.833 $\pm$ 0.071	19.3 $\pm$ 1.2 / 2.0 $\pm$ 2.2	98.3 $\pm$ 1.0
all-noise	indep.	82.6 $\pm$ 2.6	0.805 $\pm$ 0.065	14.3 $\pm$ 0.5 / 1.0 $\pm$ 1.4	98.7 $\pm$ 0.7
all-public †	shared	98.2 $\pm$ 0.0	0.990 $\pm$ 0.005	39.7 $\pm$ 5.4 / 0.3 $\pm$ 0.5	70.5 $\pm$ 11.2
aux-noise (clean T.)	shared	16.0 $\pm$ 6.6	0.524 $\pm$ 0.053	7.3 $\pm$ 0.5 / 10.3 $\pm$ 2.9	-0.2 $\pm$ 0.2
FMNIST / CNN ($n = 3$ seeds)					
Teacher	–	89.3 $\pm$ 0.6	0.954 $\pm$ 0.011	28.3 $\pm$ 4.5 / 0.3 $\pm$ 0.5	94.8 $\pm$ 1.5
aux-noise	shared	17.4 $\pm$ 3.3	0.525 $\pm$ 0.023	14.7 $\pm$ 2.6 / 11.7 $\pm$ 1.9	8.2 $\pm$ 11.6
aux-noise	indep.	6.7 $\pm$ 3.3	0.510 $\pm$ 0.025	12.3 $\pm$ 2.1 / 15.0 $\pm$ 2.4	5.7 $\pm$ 8.1
all-noise	shared	65.7 $\pm$ 1.0	0.773 $\pm$ 0.037	14.0 $\pm$ 3.6 / 3.7 $\pm$ 0.5	88.6 $\pm$ 2.4
all-noise	indep.	64.0 $\pm$ 5.8	0.717 $\pm$ 0.052	11.7 $\pm$ 4.5 / 3.0 $\pm$ 2.2	73.1 $\pm$ 18.6
all-public †	shared	89.1 $\pm$ 0.5	0.935 $\pm$ 0.011	16.3 $\pm$ 3.3 / 0.3 $\pm$ 0.5	91.9 $\pm$ 2.5
aux-noise (clean T.)	shared	12.1 $\pm$ 2.9	0.547 $\pm$ 0.032	11.3 $\pm$ 3.7 / 7.7 $\pm$ 3.9	0.0 $\pm$ 0.0
CIFAR-10 / CNN ($n = 3$ seeds)					
Teacher	–	65.4 $\pm$ 0.3	0.992 $\pm$ 0.003	98.3 $\pm$ 0.9 / 3.7 $\pm$ 0.9	84.5 $\pm$ 0.7
aux-noise	shared	10.7 $\pm$ 0.7	0.483 $\pm$ 0.023	12.3 $\pm$ 2.9 / 11.3 $\pm$ 2.6	-2.2 $\pm$ 6.2
aux-noise	indep.	9.5 $\pm$ 0.8	0.509 $\pm$ 0.007	8.3 $\pm$ 2.4 / 9.0 $\pm$ 1.4	0.0 $\pm$ 0.0
all-noise	shared	19.1 $\pm$ 6.0	0.650 $\pm$ 0.037	16.3 $\pm$ 1.2 / 10.0 $\pm$ 0.8	37.8 $\pm$ 26.2
all-noise	indep.	15.0 $\pm$ 3.9	0.592 $\pm$ 0.044	15.0 $\pm$ 0.8 / 10.0 $\pm$ 1.4	12.7 $\pm$ 9.0
all-public †	shared	64.4 $\pm$ 0.4	0.867 $\pm$ 0.027	24.7 $\pm$ 4.1 / 3.3 $\pm$ 0.9	58.3 $\pm$ 3.5
aux-noise (clean T.)	shared	8.6 $\pm$ 0.5	0.487 $\pm$ 0.010	8.3 $\pm$ 4.0 / 11.3 $\pm$ 0.5	1.9 $\pm$ 3.7

(3) *With a shared initialization, semantically empty outputs leak.* Releasing only the $m = 3$ auxiliary logits—outputs the teacher never trained, which carry no class

information—yields a student with a membership AUC of 0.862 (MNIST) and 0.807 (FMNIST), recovers the memorized label of 28.7% / 15.7% of member canaries

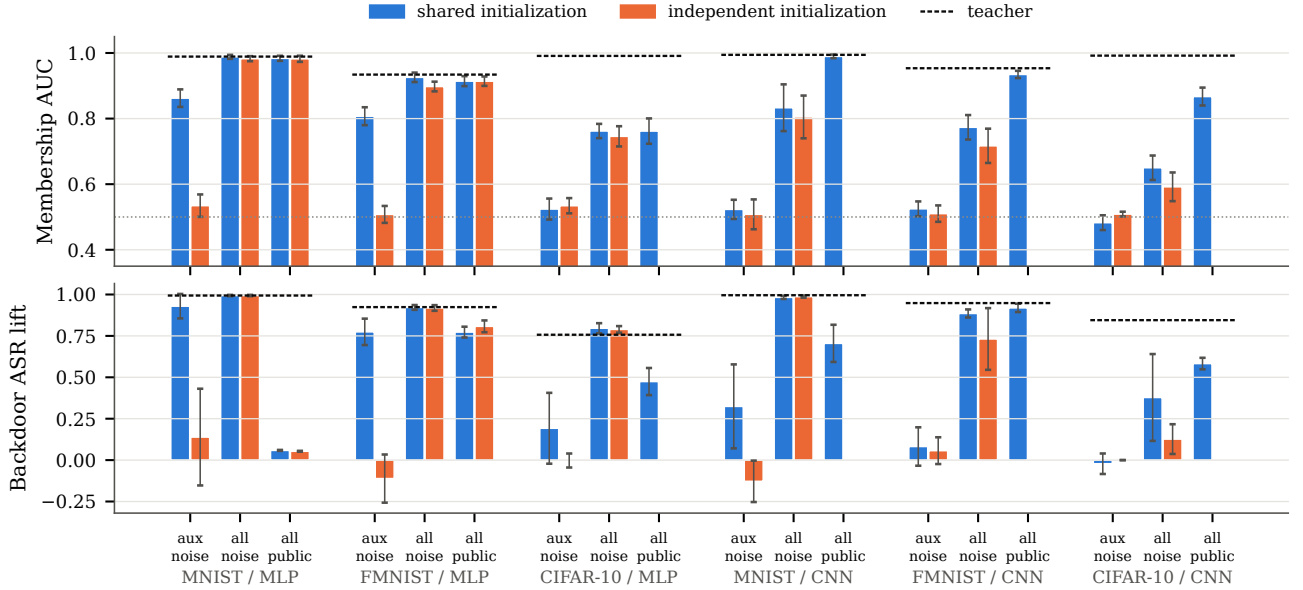


Figure 1: Membership AUC (left) and backdoor ASR lift (right) of students distilled from the private teacher, for a shared (blue) and an independent (orange) initialization, per channel. Error bars: std over seeds.

(non-members: 2.7% / 1.0%), and transmits the backdoor with an ASR lift of 93.0% / 77.4%. From an independent initialization the identical channel carries nothing: AUC 0.535 / 0.508, label extraction at the non-member rate, and test accuracy at chance. Distilling the auxiliary channel of the *clean* teacher from the shared initialization produces an equally accurate student with no membership signal (AUC 0.454 / 0.464) and no backdoor, confirming that the leaked signal originates in the private data and not in the procedure.

Observation (1) also contains a nuance that we report explicitly: on these comparatively simple tasks, full-logit distillation on Gaussian noise works from an *independent* initialization as well (MNIST AUC 0.983, ASR lift 99.4%), although it recovers fewer memorized labels than the shared-initialization student (47.0% vs. 65.7% on MNIST, 17.7% vs. 25.7% on FMNIST). The shared initialization is therefore *necessary* for the narrow auxiliary channel and *amplifying* for the full channel.

(4) *Where the subliminal channel stops.* The auxiliary channel is regime-dependent. For the CNNs on all three datasets, and for MLP teachers trained for 20 epochs on CIFAR-10, it carries no membership signal within our attack budget (AUC 0.48–0.53, label extraction at the non-member rate) and at most a weak, high-variance backdoor signal (MNIST/CNN: lift $32 \pm 25\%$ vs. $-13 \pm 13\%$ from an independent initialization). These are the settings in which the teacher travels furthest from its initialization, which is precisely where the lazy-regime transport of Theorem 1 is expected to degrade (Section 7). Teacher drift is part, but not all, of the explanation. On CIFAR-10/MLP, teachers trained for 2 or 5 epochs (relative distance 1.2 and 2.1 from the initialization) still transmit the backdoor through the auxiliary channel (lift 34% vs. 3% and -3% from an independent initialization, 3 seeds)

but only a weak membership signal (AUC 0.57–0.58), and the 20-epoch teacher (distance 4.6) transmits little. For the CNN, in contrast, even a 1-epoch MNIST teacher (distance 0.91, comparable to an MLP teacher that leaks strongly) yields no membership signal. The mechanism of Proposition 1 identifies the reason: under Gaussian queries the 128 units that feed the CNN’s read-out have a mean firing entropy of only 0.34 bits and 47% of them are constant (max-pooling of noise produces positively biased features), versus 0.83 bits and 0% for the MLP’s last layer. Isotropic noise thus excites far fewer directions of a CNN’s head-facing features, the same failure mode as poorly diversified query distributions in Section 7; query distributions matched to convolutional architectures are a natural direction for stronger attacks. The full-logit noise channel, in contrast, leaks in *every* setting: membership AUC 0.76 (CIFAR-10/MLP), 0.83, 0.77 and 0.65 (MNIST, FMNIST, CIFAR-10 CNNs), with backdoor lifts of 80–98% except on CIFAR-10/CNN (38%), always at least as large from the shared initialization as from an independent one (e.g. FMNIST/CNN backdoor 88.6% vs. 73.1%; CIFAR-10/CNN AUC 0.650 vs. 0.592).

6.2 The realistic threat: private fine-tunes of public checkpoints

Table 2 instantiates the threat model of Section 3. An attacker who owns the public base and sends only Gaussian noise to the fine-tuned model obtains a student that is, for all practical purposes, the private model: on MNIST it recovers the memorized label of 92.7% of member canaries (the private model: 93.7%; non-members: 0.7%), matches its calibrated membership AUC (0.983) and inherits its backdoor (99.3%). The sibling-base control, which has the same public knowledge, architecture and accuracy but a different initialization, recovers

Table 2: Fine-tuning threat model. The provider fine-tunes a public base on private data; the attacker owns the public base, has no data, and queries the fine-tuned model’s logits with Gaussian noise. The control attacker starts from an equally accurate sibling base pre-trained on the same public data with a different seed. Cal. AUC: membership AUC of the reference-calibrated score (reference = the attacker’s starting model). [†]Reference attack with in-distribution public queries.

Model / attack	Queries	Test acc. (%)	Memb. AUC	Cal. AUC	Label extr. mem / non (%)	ASR lift (%)
MNIST / MLP ($n = 3$ seeds)						
Public base	–	96.7 $\pm$ 0.1	0.519 $\pm$ 0.045	–	0.3 $\pm$ 0.5 / 0.7 $\pm$ 0.9	0.0 $\pm$ 0.0
Private fine-tune (target)	–	97.7 $\pm$ 0.1	0.994 $\pm$ 0.004	0.983 $\pm$ 0.004	93.7 $\pm$ 0.5 / 0.7 $\pm$ 0.5	99.3 $\pm$ 0.2
Attack: shared base	noise	97.7 $\pm$ 0.1	0.993 $\pm$ 0.005	0.983 $\pm$ 0.003	92.7 $\pm$ 1.2 / 0.7 $\pm$ 0.5	99.3 $\pm$ 0.2
Control: sibling base	noise	97.3 $\pm$ 0.2	0.960 $\pm$ 0.018	0.957 $\pm$ 0.018	8.3 $\pm$ 4.2 / 1.3 $\pm$ 0.9	99.1 $\pm$ 0.3
Ref.: shared base	public [†]	97.7 $\pm$ 0.1	0.993 $\pm$ 0.005	0.981 $\pm$ 0.004	72.0 $\pm$ 3.6 / 0.7 $\pm$ 0.5	4.1 $\pm$ 1.0
Ref.: sibling base	public [†]	97.5 $\pm$ 0.1	0.981 $\pm$ 0.012	0.960 $\pm$ 0.013	10.3 $\pm$ 3.3 / 0.7 $\pm$ 0.5	2.1 $\pm$ 1.1
FMNIST / MLP ($n = 3$ seeds)						
Public base	–	85.2 $\pm$ 0.7	0.464 $\pm$ 0.037	–	2.0 $\pm$ 1.6 / 2.3 $\pm$ 0.5	-0.8 $\pm$ 0.4
Private fine-tune (target)	–	88.1 $\pm$ 0.3	0.978 $\pm$ 0.015	0.921 $\pm$ 0.013	53.0 $\pm$ 2.4 / 1.3 $\pm$ 1.2	93.9 $\pm$ 0.9
Attack: shared base	noise	88.2 $\pm$ 0.3	0.976 $\pm$ 0.015	0.922 $\pm$ 0.013	49.7 $\pm$ 1.7 / 1.7 $\pm$ 1.7	93.7 $\pm$ 0.9
Control: sibling base	noise	86.1 $\pm$ 0.5	0.860 $\pm$ 0.020	0.861 $\pm$ 0.029	8.0 $\pm$ 2.4 / 1.7 $\pm$ 1.2	92.3 $\pm$ 1.5
Ref.: shared base	public [†]	88.0 $\pm$ 0.3	0.958 $\pm$ 0.023	0.906 $\pm$ 0.013	17.3 $\pm$ 2.9 / 1.7 $\pm$ 1.7	83.8 $\pm$ 3.0
Ref.: sibling base	public [†]	87.6 $\pm$ 0.3	0.939 $\pm$ 0.028	0.891 $\pm$ 0.017	8.7 $\pm$ 4.1 / 1.3 $\pm$ 1.9	71.8 $\pm$ 3.3
CIFAR-10 / CNN ($n = 3$ seeds)						
Public base	–	67.0 $\pm$ 0.2	0.472 $\pm$ 0.038	–	2.3 $\pm$ 0.5 / 5.0 $\pm$ 0.8	-0.1 $\pm$ 1.3
Private fine-tune (target)	–	71.2 $\pm$ 0.2	0.981 $\pm$ 0.012	0.930 $\pm$ 0.027	89.0 $\pm$ 4.2 / 3.3 $\pm$ 2.1	79.0 $\pm$ 0.4
Attack: shared base	noise	66.6 $\pm$ 0.7	0.635 $\pm$ 0.014	0.736 $\pm$ 0.056	6.0 $\pm$ 1.6 / 4.3 $\pm$ 0.9	38.4 $\pm$ 8.3
Control: sibling base	noise	40.4 $\pm$ 8.8	0.550 $\pm$ 0.033	0.561 $\pm$ 0.029	9.7 $\pm$ 3.7 / 8.7 $\pm$ 1.9	7.6 $\pm$ 2.4
Ref.: shared base	public [†]	70.4 $\pm$ 0.3	0.804 $\pm$ 0.027	0.884 $\pm$ 0.035	15.3 $\pm$ 3.1 / 3.3 $\pm$ 1.2	55.3 $\pm$ 1.9
Ref.: sibling base	public [†]	67.4 $\pm$ 0.2	0.635 $\pm$ 0.040	0.751 $\pm$ 0.016	7.0 $\pm$ 1.4 / 4.7 $\pm$ 1.2	26.6 $\pm$ 2.8

the membership signal only partially (calibrated AUC 0.957 on MNIST, 0.861 on FMNIST) and extracts few memorized labels. The difference is the shared checkpoint. On CIFAR-10 with a CNN—a harder task on which plain noise distillation from scratch is known to be weak [39]—the attack is weaker but the role of the shared checkpoint is starker. The shared-base noise attack reaches a calibrated membership AUC of 0.736 and transmits the backdoor with a lift of 38.4% while preserving the base’s accuracy; from the sibling base the same queries yield 0.561 and 7.6% and destroy the student’s accuracy (40.4%). Instance-level label extraction, however, stays close to the non-member rate on CIFAR-10 (6.0% vs. 4.3%), so on this harder setting the noise attack recovers a membership signal and the backdoor but not the memorized labels themselves.

6.3 Language models: random strings expose fine-tuning secrets

We repeat the fine-tuning threat model with a character-level transformer language model (2 layers, width 128, 4 heads, context 64, 0.42M parameters). A public base is pre-trained on the first half of Tiny Shakespeare; the provider fine-tunes it on the second half, into which 20 random six-digit secrets of the form “my pin is 381920.” have been inserted 8 times each; 20 further secrets are never inserted and serve as non-members. The API returns next-token logits at every position. The attacker owns the base and queries the API with *uniformly random character strings*, minimizing the KL divergence to the returned distributions; the control

Table 3: Language-model fine-tuning threat model (Tiny Shakespeare, character-level transformer, 3 seeds). Exposure in bits (maximum ≈ 19.9 ; values above ≈ 14 mean the secret outranks all $2 \cdot 10^4$ sampled candidates). [†]Reference attack with in-distribution public text.

Model / attack	Val. loss	Exp. member	Exp. non-member	Cal. AUC
Public base	1.988 $\pm$ 0.006	1.15 $\pm$ 0.27	1.50 $\pm$ 0.25	0.500 $\pm$ 0.000
Private fine-tune (target)	1.319 $\pm$ 0.006	19.65 $\pm$ 0.40	1.36 $\pm$ 0.41	0.986 $\pm$ 0.014
Sibling base	1.992 $\pm$ 0.008	1.49 $\pm$ 0.24	1.43 $\pm$ 0.12	0.500 $\pm$ 0.000
Attack: shared base, random strings	1.326 $\pm$ 0.005	19.36 $\pm$ 0.80	1.34 $\pm$ 0.41	0.983 $\pm$ 0.013
Ref.: shared base, public text [†]	1.359 $\pm$ 0.004	10.74 $\pm$ 0.89	1.10 $\pm$ 0.24	0.991 $\pm$ 0.005
Control: sibling base, random strings	1.941 $\pm$ 0.005	10.76 $\pm$ 0.63	1.31 $\pm$ 0.34	0.901 $\pm$ 0.072
Ref.: sibling base, public text [†]	1.655 $\pm$ 0.009	1.99 $\pm$ 0.20	1.46 $\pm$ 0.25	0.632 $\pm$ 0.057

attacker starts from a sibling base pre-trained with a different seed. We measure the *exposure* of each secret [5], $\log_2 |\mathcal{R}| - \log_2 \text{rank}$ over the $|\mathcal{R}| = 10^6$ possible secrets (rank estimated from $2 \cdot 10^4$ random candidates), and a reference-calibrated membership AUC.

Table 3 shows that the language model behaves exactly like the image models. The private fine-tune memorizes the secrets (mean member exposure 19.65 bits, non-members 1.36). The attacker that owns the public base and sends only random character strings recovers a model with the same validation loss (1.326 vs. 1.319) and essentially the same exposure for every member secret (19.36 bits; non-members 1.34): the private secrets become extractable by ranking, with no access to any text. The sibling-base control with the same random queries reaches only 10.76 bits and fails to recover the fine-tuned model’s utility (loss 1.941), and distilling from the shared base on in-distribution public text—the classical, data-based attack—is markedly *weaker* than random strings (10.74 bits), mirroring the finding that

noise is a better carrier than data (Section 7).

7 Testing the Theory

Leakage is monotone in transport time. Theorem 1 predicts that, from a shared initialization, every component of the teacher’s displacement is transported progressively, so leakage should increase monotonically with the number of distillation steps, whereas an independent initialization provides no transport. Fig. 2 confirms this for the auxiliary channel: the membership AUC rises from 0.5 to 0.84 over 3.2×10^4 steps and the independent student stays at chance throughout. The backdoor is transported far earlier than the membership signal (ASR lift > 0.5 after 500 steps), as expected from its high visibility: the trigger lives in pixels that natural images never activate but Gaussian noise always does.

Query budget and channel width collapse onto Nm/p . Proposition 2 predicts that with a fixed pool of N queries the transported fraction depends on the number of scalar constraints Nm , and the lazy picture suggests saturation once Nm reaches the number of parameters p . We vary $N \in [16, 16384]$ and $m \in \{1, 3, 10, 30\}$ (Fig. 3). Plotted against N the curves separate by channel width; plotted against Nm/p they collapse onto a single curve, with membership AUC rising sharply as $Nm/p \rightarrow 1$ (AUC 0.985 at $Nm/p = 1.8$ for $m = 30$). With unlimited fresh queries the width alone sets the leakage achieved in a fixed number of steps: AUC 0.61, 0.86, 0.98 and 0.98 for $m = 1, 3, 10, 30$. A wider channel is a faster channel.

The closed-form transport prediction matches trained students. To test Theorem 1 quantitatively we use a narrower MLP (width 64, $p = 55,245$) and a finite pool of N noise queries, for which Proposition 2 gives the student’s limit in closed form, $\theta_0 + J_N^+(g(Z; \theta_T) - g(Z; \theta_0))$. We compute this prediction exactly from the Jacobian $J_N \in \mathbb{R}^{Nm \times p}$ and compare it with a student actually trained by gradient descent on the same queries (Table 4). Three results stand out. First, the fraction of the teacher’s displacement captured by the projector, $\|\Pi_N \Delta\|^2 / \|\Delta\|^2$, equals Nm/p to within a few percent at every budget (e.g. 0.059 at $Nm/p = 0.056$ and 0.12 at $Nm/p = 0.111$), the value expected if the private displacement were isotropic with respect to the query Jacobians: *nothing about the private data makes it harder to transport than a random direction*. Second, the predicted and actual parameter displacements are substantially aligned (cosine 0.55–0.75, decreasing slowly with N) although the network is non-linear and the teacher was trained with Adam. Third, the predicted and trained students agree closely on membership AUC and accuracy, and both show a backdoor lift that grows with N ; the trained student exceeds the linear prediction for the backdoor by up to a factor of two, a non-linear amplification that only strengthens the attack.

The query distribution matters through activation diversity, and noise beats data. Proposition 1 says that any full-support Q makes the same directions

Table 4: Closed-form prediction (Prop. 2) vs. trained student (MNIST, MLP of width 64, $p = 55,245$, auxiliary channel $m = 3$, 3 seeds). Captured: $\|\Pi_N \Delta\|^2 / \|\Delta\|^2$. cos: cosine between predicted and actual parameter displacement.

N	Nm	Captured	cos	AUC pred.	AUC act.	ASR pred. (%)	ASR act. (%)
16	48	0.001±0.000	0.75±0.01	0.529±0.013	0.530±0.012	1.8±2.5	2.9±4.1
64	192	0.004±0.001	0.70±0.02	0.533±0.014	0.536±0.010	2.6±3.7	3.8±5.4
256	768	0.016±0.001	0.64±0.02	0.524±0.017	0.534±0.010	7.0±9.8	6.2±8.8
1024	3072	0.063±0.003	0.58±0.02	0.533±0.013	0.544±0.016	10.1±12.5	19.6±27.5
2048	6144	0.123±0.005	0.55±0.01	0.530±0.038	0.543±0.017	12.3±14.2	24.8±33.0

Table 5: Query distributions (MNIST/MLP, auxiliary channel, 3 seeds, fixed pool of 65,536 queries). Firing entropy: mean binary entropy of hidden-unit activation under Q ; Const.: % of units on/off for $> 98\%$ of queries; s.r.: stable rank $\text{tr } K_Q / \lambda_{\max}(K_Q)$.

Q	Entropy	Const. (%)	s.r.	Acc. (%)	AUC	ASR lift (%)
$\mathcal{N}(0, 0.25^2)$	0.85±0.01	1.8±0.6	6.9±0.2	48.0±8.9	0.641±0.006	54.5±30.0
$\mathcal{N}(0, 0.5^2)$	0.90±0.00	0.4±0.2	12.7±1.1	51.0±12.0	0.677±0.003	79.1±11.8
$\mathcal{N}(0, 1)$	0.91±0.00	0.1±0.1	17.5±1.5	55.1±9.7	0.673±0.030	63.0±8.5
$\mathcal{N}(0, 2^2)$	0.92±0.00	0.1±0.1	20.4±1.3	51.8±7.8	0.658±0.021	76.9±17.4
$\mathcal{N}(0, 4^2)$	0.92±0.00	0.1±0.1	21.1±1.7	49.3±7.8	0.673±0.012	79.7±7.8
$\mathcal{N}(1.2, 1)$ (shifted)	0.61±0.01	13.4±1.8	5.6±0.3	13.0±2.3	0.520±0.032	3.8±5.2
Unif[0, 1] pixels	0.59±0.01	15.8±1.4	5.0±0.7	13.1±1.4	0.466±0.045	0.6±0.9
public images	0.79±0.01	2.0±0.9	9.1±0.4	18.9±4.6	0.574±0.017	0.8±1.1

leakable but that the rate depends on how diversely Q exercises the network’s activation patterns. Table 5 compares eight query distributions at a fixed budget of 65,536 queries. Centered Gaussians of any scale from 0.25 to 4 in standardized space leak strongly (AUC 0.64–0.68, ASR lift 55–80%). Mean-shifted Gaussian noise and noise that is uniform in pixel space—which drives a large fraction of units into a constant on/off regime—barely leak, and natural public images, whose activation patterns are concentrated on the data manifold, are worse than Gaussian noise for every metric. The mean firing entropy of hidden units under Q orders the distributions by leakage (Fig. 4), in line with the proof of Proposition 1, which requires many activation patterns to be realized, and with the low stable rank of K_Q under the poorly diversified distributions.

The noise channel is blind to what the displacement encodes. Theorem 1 implies that a component of the teacher’s displacement is transported at a rate set by its visibility $\mathcal{V}_Q$, irrespective of whether it encodes knowledge or memorization. We isolate three components by training matched teachers from the same initialization with the same optimizer seed: a *task* component (clean teacher minus initialization), a *canary* component (teacher with canaries minus clean teacher) and a *backdoor* component (poisoned minus clean teacher), and measure their visibility through the auxiliary channel (Table 6). Under Gaussian queries the three components are equally visible to within a few percent, and comparable to a random direction: the channel cannot tell memorization from knowledge. Under natural-data queries, in contrast, the private components are markedly *less* visible than the task component—data queries are partially selective, noise queries are not. (Differences between separately trained networks also contain chaotic training divergence, so these components are upper bounds on the purely data-driven differences; this does not affect the comparison

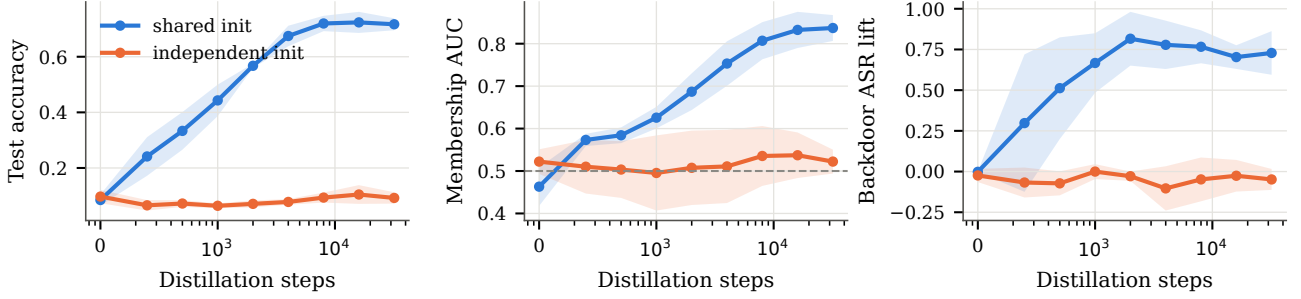


Figure 2: Auxiliary-channel leakage as a function of distillation steps (MNIST/MLP, mean±std over 3 seeds; the leftmost point is the initialization).

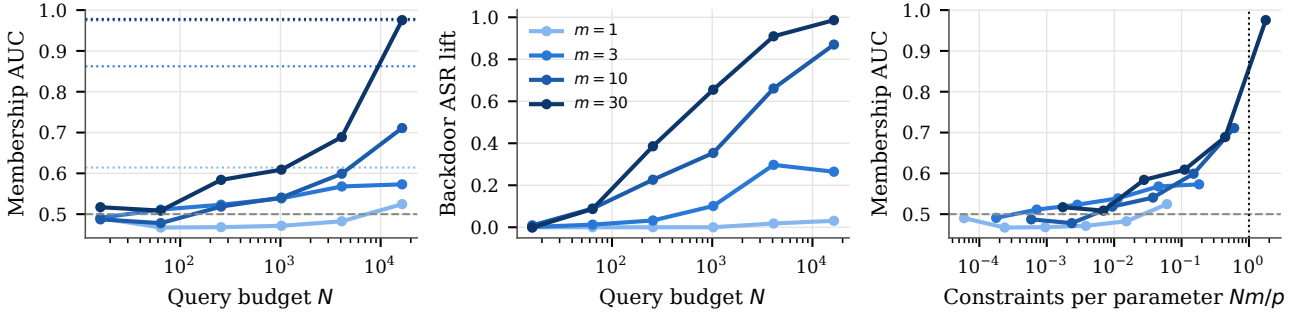


Figure 3: Finite query budgets (MNIST/MLP, auxiliary channel, mean over 3 seeds). Left: membership AUC vs. number of distinct noise queries N for channel width m (dotted: unlimited fresh queries). Middle: backdoor ASR lift. Right: the same data against the number of scalar constraints per parameter Nm/p ; the curves collapse, as predicted by Proposition 2.

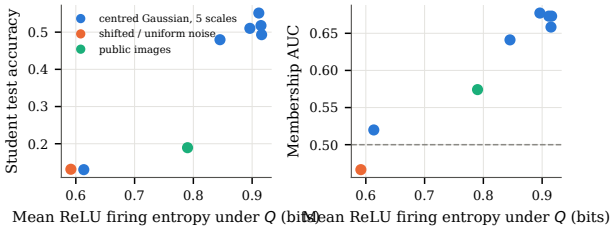


Figure 4: Student accuracy and membership AUC against the mean ReLU firing entropy induced by the query distribution (MNIST/MLP, auxiliary channel, mean over 3 seeds).

Table 6: Visibility $\mathcal{V}_Q(v) \times 10^3$ of displacement components through the auxiliary channel (MNIST/MLP, 3 seeds).

Component	Gaussian Q	Uniform-pixel Q	Public data Q
task	0.532±0.076	5.98±2.70	3.26±0.61
canary	0.515±0.134	5.68±1.88	1.77±0.20
backdoor	0.471±0.096	5.94±0.27	2.14±0.25
random	0.312±0.020	0.62±0.08	0.29±0.04

across query distributions.)

Boundaries of the lazy picture. The theory is exact in the lazy regime. Two sweeps probe its limits (Table 7). First, leakage is non-monotone in teacher

training length: the teacher’s own memorization grows with training (canary AUC 0.72 → 0.999), but beyond 5 epochs the student’s leakage at a fixed attack budget *decreases* as the teacher drifts further from the initialization (relative distance 1.3 → 3.0), consistent with a degradation of the linear transport. Second, at a fixed attacker budget, leakage is highest for moderate widths and decreases for the widest networks. This second effect is an artefact of the attacker’s optimizer, not a limit of the channel: re-running the attack on width 512 and 1024 with a smaller learning rate ($3 \cdot 10^{-4}$, $1.6 \cdot 10^4$ steps) raises the membership AUC to 0.957 and 0.787 and the backdoor lift to 98.9% and 97.8% (3 seeds); width 512 then leaks *more* than width 256, as expected from wider networks being closer to the lazy regime. Teacher drift, by contrast, is a genuine boundary of the subliminal channel, which we return to for convolutional networks below.

8 Defenses

Initialization mismatch and secret offsets. Proposition 3 predicts graceful degradation of the signal under a small initialization offset. Fig. 5 (left) shows that the channel is in fact remarkably robust: an attacker whose initialization is a norm-preserving mixture $\sqrt{1 - \alpha^2}\theta_0 + \alpha\theta'$ of the true initialization and an independent one retains essentially full leakage up to $\alpha = 0.3$; leakage decays for $\alpha \geq 0.5$ and vanishes at

Table 7: Teacher drift and width (MNIST/MLP, auxiliary channel, 3 seeds, 8,000 attack steps). $\|\Delta\|/\|\theta_0\|$: relative distance of the teacher from the initialization. T-AUC: teacher membership AUC.

Epochs	$\ \Delta\ /\ \theta_0\ $	T-AUC	Acc. (%)	AUC	Label (%)	ASR lift (%)
1	0.63±0.01	0.720±0.055	88.1±1.0	0.638±0.038	3.7±1.2	93.8±4.2
2	0.83±0.00	0.897±0.029	91.4±1.4	0.790±0.019	8.7±2.5	96.8±3.2
5	1.26±0.01	0.989±0.005	80.8±6.9	0.862±0.027	28.7±1.2	93.0±7.4
10	1.66±0.01	0.997±0.002	65.9±11.2	0.797±0.029	28.3±5.9	75.7±24.6
20	2.23±0.02	0.997±0.002	52.2±6.6	0.721±0.013	24.7±1.2	65.1±35.3
40	2.98±0.07	0.999±0.000	31.9±5.7	0.642±0.036	18.3±6.2	25.1±18.1
Width		T-AUC	Acc. (%)	AUC	Label (%)	ASR lift (%)
64		0.921±0.023	90.3±2.1	0.855±0.026	19.7±5.6	97.7±1.5
128		0.973±0.008	92.7±1.5	0.935±0.016	34.3±4.6	99.1±0.2
256		0.989±0.005	80.8±6.9	0.862±0.027	28.7±1.2	93.0±7.4
512		0.994±0.002	62.6±7.1	0.751±0.029	15.3±2.4	80.6±16.7
1024		0.995±0.002	29.3±1.0	0.575±0.020	11.3±2.1	71.5±23.0

Table 8: Logit noise versus attacker query multiplier (MNIST/MLP, auxiliary channel, 3 seeds).

Noise s	Queries/step	Acc. (%)	Memb. AUC	ASR lift (%)
0.3	1	52.6±3.6	0.701±0.029	50.7±18.7
0.3	4	87.1±3.3	0.909±0.012	95.2±3.1
0.3	16	93.0±1.7	0.958±0.007	98.7±1.1
1	1	21.3±7.3	0.560±0.046	4.7±4.1
1	4	37.5±2.2	0.623±0.043	35.8±9.8
1	16	77.3±6.6	0.812±0.045	83.6±12.6

$\alpha = 1$. The mirror-image defense, in which the provider trains from a *secret* perturbation of the public initialization (right), behaves identically and costs nothing in teacher accuracy when the initialization is random. The offset must, however, be large ($\alpha \gtrsim 0.7$) to be effective, i.e. the private model must essentially not share its initialization. For a model fine-tuned from a valuable public checkpoint this forfeits the checkpoint; a secret offset is thus a sound defense for models trained from scratch but not a free one for fine-tuned models.

Output perturbation. A provider might add Gaussian noise to every released logit. Proposition 4 predicts that this cannot stop an attacker with fresh queries, since the expected gradient is unchanged; it only adds variance that the attacker can average away by spending more queries. Fig. 6 and Table 8 test both halves of the prediction. At a fixed attack budget, noise is effective: a standard deviation of 0.3 times the standard deviation of the logits lowers the membership AUC from 0.86 to 0.70, and noise equal to the logit scale brings it to 0.56. But an attacker who spends $4\times$ or $16\times$ more queries per step recovers the leakage: at noise level 0.3 the AUC returns to 0.91 and 0.96 (above the noiseless $1\times$ attack, because larger batches also reduce ordinary sampling noise), and even at noise level 1, which makes individual answers useless to any legitimate user, $16\times$ more queries restore an AUC of 0.81 and a backdoor ASR lift of 84%. Because noise queries are free, output perturbation increases the attacker’s cost by a constant factor but does not close the channel.

Differential privacy. DP-SGD [1] bounds what *any* post-processing of the teacher—including every distillation channel studied here—can reveal about a training record [13]. Table 9 trains the teacher with DP-SGD (Opacus [48], clipping norm 1, 5 epochs, $\delta = 10^{-5}$)

Table 9: DP-SGD teachers (MNIST/MLP, 3 seeds; ϵ per example at $\delta = 10^{-5}$; canaries appear 4 times, so the group guarantee is $\approx 4\epsilon$). Students start from the shared initialization.

σ	ϵ	T. acc. (%)	T. AUC	aux AUC	all AUC	T. ASR lift (%)	all ASR lift (%)
0	∞	96.6±0.3	0.982±0.010	0.870±0.031	0.980±0.010	99.5±0.1	99.5±0.2
0.5	11.4	90.9±0.2	0.528±0.034	0.508±0.020	0.529±0.034	44.8±4.9	45.8±4.7
1	1.5	89.4±0.1	0.520±0.034	0.497±0.027	0.520±0.035	8.7±1.9	8.9±1.9
2	0.4	87.3±0.4	0.517±0.037	0.495±0.033	0.519±0.038	3.6±0.9	3.7±0.9

and attacks it through both channels from the shared initialization. Already at a noise multiplier of 0.5 the canary membership signal disappears from the teacher and therefore from every student, at a cost of about 6 points of accuracy. The backdoor, which is planted in 300 examples and is therefore far outside the group size that a per-example guarantee protects, survives partially at $\sigma = 0.5$ (ASR lift 45% in teacher and student alike) and is largely removed at $\sigma \geq 1$. DP is thus the appropriate defense for record-level privacy, and it confirms the mechanism: the student never leaks more than the teacher holds.

9 Discussion

Why intuition fails. The belief that distillation on uninformative inputs sanitizes a model rests on a function-space picture: the student can only learn what the queries reveal, and noise reveals nothing about the private data. Theorem 1 replaces this with a parameter-space picture. Once teacher and student share θ_0 , the queries do not need to reveal the private data; they only need to excite the parameter directions in which the private data has already been written. Noise does so generically (Proposition 1), and the operator that performs the transport is fixed before any private record is seen, so it cannot treat memorized records differently from generalizable structure.

Practical implications. (i) *Public checkpoints are part of the attack surface.* Whenever a deployed model is a fine-tune of a public base, the base is side information that converts output access into a far stronger channel. Privacy evaluations of such models should include an adversary that owns the base. (ii) *“Data-free” does not mean “privacy-preserving”.* Distilling a private model on synthetic or noise inputs into a student that shares its initialization—a natural way to produce a smaller or re-licensable artefact—should be assumed to carry the teacher’s memorization and backdoors. (iii) *Restricting outputs is not a remedy.* Channels that carry no task semantics still carry parameter information. (iv) *Output perturbation is the wrong tool* against data-free adversaries, since fresh noise queries average it out (Proposition 4). (v) *Effective mitigations act on the parameters:* differential privacy bounds every downstream channel by post-processing, and decorrelating the private model’s starting point from any public one removes the transport mechanism—but only when the offset is large ($\alpha \gtrsim 0.7$ in our experiments), at a utility cost that depends on how much the public starting point is worth.

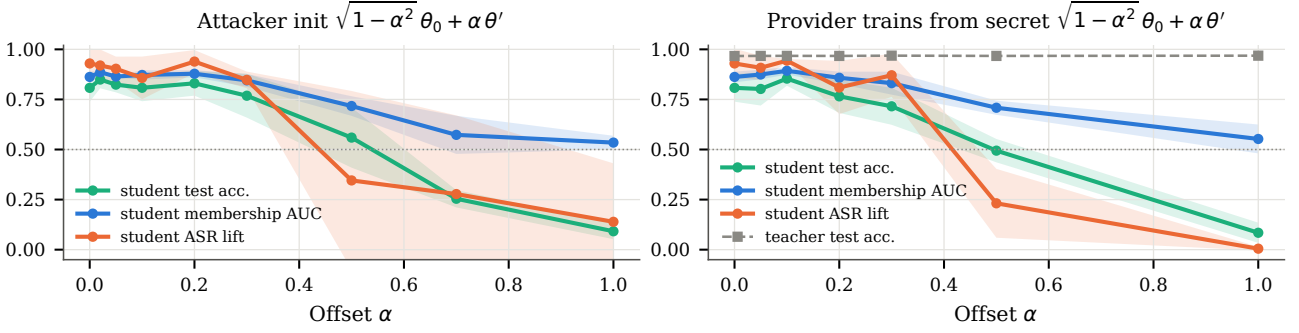


Figure 5: Initialization mismatch (MNIST/MLP, auxiliary channel, mean $\pm$ std over 3 seeds). Left: the attacker starts from a mixture of the true initialization and an independent one. Right: the provider trains from a secret mixture while the attacker uses the public θ_0 ; the teacher’s accuracy (gray) is unaffected.

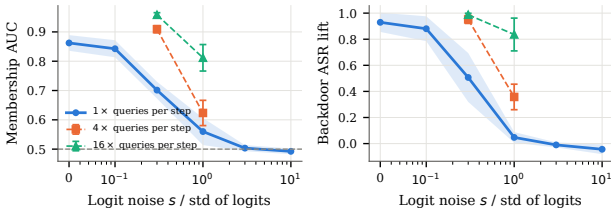


Figure 6: Output perturbation (MNIST/MLP, auxiliary channel, 3 seeds). Blue: 8,000 attack steps with 256 fresh queries each. Orange/green: the same number of steps with $4\times$ / $16\times$ more queries per step.

Limitations. Our experiments use small networks and datasets so that every claim can be tested with three seeds, controls and ablations on a 4-core CPU; the theory is exact only in the lazy regime, which we verify approximately rather than assume. The subliminal (auxiliary-output) channel is strongest for MLPs whose teachers stay near the initialization; for our CNNs and for heavily trained teachers it carries little membership information within our attack budget, and on CIFAR-10 the full-logit noise attack recovers a membership signal and the backdoor but not the memorized labels. We traced the CNN case to the poor activation diversity that isotropic noise induces in head-facing convolutional features, but we did not design architecture-matched query distributions, so our CNN numbers should be read as lower bounds on what an adaptive attacker can obtain. Canaries are inserted four times to obtain a clearly measurable teacher-side signal; single-occurrence records leak less in absolute terms, but Theorem 1 implies that the student inherits whatever signal the teacher carries. Our language model is a small character-level transformer; scaling the analysis to large pre-trained transformers fine-tuned with parameter-efficient methods such as LoRA [21], where the update lives in a low-dimensional subspace and the lazy approximation is particularly accurate, is the most important direction for future work. Finally, we study membership and backdoors; attribute inference and data reconstruction through the same channel remain open.

10 Conclusion

Distillation from a shared initialization is not a filter but a conduit. In the regime where training stays close to the initialization—the regime of fine-tuning from public checkpoints—a student starting from the teacher’s initialization inherits, on pure noise and even through outputs that were never trained, the parameter displacement that the private data induced, and with it the teacher’s memorization of individual records and its backdoors. We have characterized this channel theoretically, measured it across datasets, architectures, a fine-tuning threat model and a language model, and identified which defenses do and do not close it. As models are increasingly derived from a small set of public checkpoints, privacy and security analyses must treat the shared initialization as adversary knowledge.

References

- [1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: ACM Conference on Computer and Communications Security (CCS). pp. 308–318 (2016)
- [2] Blank, C., Bhatia, A., Rajamanoharan, S., Conmy, A., Nanda, N.: Subliminal learning is steering vector distillation. arXiv preprint arXiv:2606.00995 (2026)
- [3] Brockers, V.C., Ventzke, R.D., Neuhaus, V., Hidalgo-Ogalde, B., Priesemann, V.: Learning through noise: Why subliminal learning works and when it fails. arXiv preprint arXiv:2605.23645 (2026)
- [4] Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramèr, F.: Membership inference attacks from first principles. In: IEEE Symposium on Security and Privacy (S&P). pp. 1897–1914 (2022)
- [5] Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., Song, D.: The secret sharer: Evaluating and testing

- unintended memorization in neural networks. In: USENIX Security Symposium. pp. 267–284 (2019)
- [6] Carlini, N., Paleka, D., Dvijotham, K.D., Steinke, T., Hayase, J., Cooper, A.F., Lee, K., Jagielski, M., Nasr, M., Conmy, A., Wallace, E., Rolnick, D., Tramèr, F.: Stealing part of a production language model. In: International Conference on Machine Learning (ICML) (2024)
- [7] Chen, H., Wang, Y., Xu, C., Yang, Z., Liu, C., Shi, B., Xu, C., Xu, C., Tian, Q.: Data-free learning of student networks. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3514–3522 (2019)
- [8] Chen, X., Liu, C., Li, B., Lu, K., Song, D.: Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526 (2017)
- [9] Chizat, L., Oyallon, E., Bach, F.: On lazy training in differentiable programming. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
- [10] Cloud, A., Le, M., Chua, J., Betley, J., Sztzyber-Betley, A., Hilton, J., Marks, S., Evans, O.: Subliminal learning: Language models transmit behavioral traits via hidden signals in data. arXiv preprint arXiv:2507.14805 (2025), published in *Nature* (2026) as “Language models transmit behavioural traits through hidden signals in data”
- [11] Correia-Silva, J.R., Berriel, R.F., Badue, C., de Souza, A.F., Oliveira-Santos, T.: Copycat CNN: Stealing knowledge by persuading confession with random non-labeled data. In: International Joint Conference on Neural Networks (IJCNN) (2018)
- [12] Dinh, L., Pascanu, R., Bengio, S., Bengio, Y.: Sharp minima can generalize for deep nets. In: International Conference on Machine Learning (ICML). pp. 1019–1028 (2017)
- [13] Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* **9**(3–4), 211–407 (2014)
- [14] Finlayson, M., Ren, X., Swayamdipta, S.: Logits of API-protected LLMs leak proprietary information. In: Conference on Language Modeling (COLM) (2024)
- [15] Fort, S., Dziugaite, G.K., Paul, M., Kharaghani, S., Roy, D.M., Ganguli, S.: Deep learning versus kernel learning: An empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
- [16] Ge, Y., Wang, Q., Zheng, B., Zhuang, X., Li, Q., Shen, C., Wang, C.: Anti-distillation backdoor attacks: Backdoors can really survive in knowledge distillation. In: ACM International Conference on Multimedia (MM). pp. 826–834 (2021)
- [17] Gu, T., Dolan-Gavitt, B., Garg, S.: BadNets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733 (2017)
- [18] Hadley, E., Gultepe, E.: Subliminal learning is non-semantic distillation. arXiv preprint arXiv:2608.05734 (2026)
- [19] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
- [20] Hong, J., Zeng, Y., Yu, S., Lyu, L., Jia, R., Zhou, J.: Revisiting data-free knowledge distillation with poisoned teachers. In: International Conference on Machine Learning (ICML) (2023)
- [21] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
- [22] Jacot, A., Gabriel, F., Hongler, C.: Neural tangent kernel: Convergence and generalization in neural networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
- [23] Jagielski, M., Nasr, M., Lee, K., Choquette-Choo, C.A., Carlini, N., Tramèr, F.: Students parrot their teachers: Membership inference on model distillation. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
- [24] Kariyappa, S., Prakash, A., Qureshi, M.K.: MAZE: Data-free model stealing attack using zeroth-order gradient estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13814–13823 (2021)
- [25] Kitkana, C., et al.: Sustained gradient alignment mediates subliminal learning in a multi-step setting: Evidence from MNIST auxiliary logit distillation experiment. arXiv preprint arXiv:2604.25779 (2026)
- [26] Krishna, K., Tomar, G.S., Parikh, A.P., Papernot, N., Iyyer, M.: Thieves on sesame street! model extraction of BERT-based APIs. In: International Conference on Learning Representations (ICLR) (2020)
- [27] Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., Pennington, J.: Wide neural networks of any depth evolve as linear models under gradient descent. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
- [28] Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., Ma, X.: Neural attention distillation: Erasing backdoor triggers from deep neural networks. In: International

- Conference on Learning Representations (ICLR) (2021)
- [29] Li, Z., Ying, B., Dong, C., Yang, H.: Why ghost outputs teach: A kernel-based understanding of subliminal learning. *arXiv preprint arXiv:2609.23260* (2026)
 - [30] Lopes, R.G., Fenu, S., Starner, T.: Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535* (2017)
 - [31] Malladi, S., Wettig, A., Yu, D., Chen, D., Arora, S.: A kernel-based view of language model fine-tuning. In: *International Conference on Machine Learning (ICML)* (2023)
 - [32] Micaelli, P., Storkey, A.J.: Zero-shot knowledge transfer via adversarial belief matching. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2019)
 - [33] Nayak, G.K., Mopuri, K.R., Shaj, V., Babu, R.V., Chakraborty, A.: Zero-shot knowledge distillation in deep networks. In: *International Conference on Machine Learning (ICML)*. pp. 4743–4751 (2019)
 - [34] Neyshabur, B., Salakhutdinov, R., Srebro, N.: Path-SGD: Path-normalized optimization in deep neural networks. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2015)
 - [35] Orekondy, T., Schiele, B., Fritz, M.: Knockoff nets: Stealing functionality of black-box models. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4954–4963 (2019)
 - [36] Ortiz-Jiménez, G., Favero, A., Frossard, P.: Task arithmetic in the tangent space: Improved editing of pre-trained models. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2023)
 - [37] Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I., Talwar, K.: Semi-supervised knowledge transfer for deep learning from private training data. In: *International Conference on Learning Representations (ICLR)* (2017)
 - [38] Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., Erlingsson, Ú.: Scalable private learning with PATE. In: *International Conference on Learning Representations (ICLR)* (2018)
 - [39] Raikwar, P., Mishra, D.: Discovering and overcoming limitations of noise-engineered data-free knowledge distillation. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2022)
 - [40] Shejwalkar, V., Houmansadr, A.: Membership privacy for machine learning models through knowledge transfer. In: *AAAI Conference on Artificial Intelligence*. pp. 9549–9557 (2021)
 - [41] Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: *IEEE Symposium on Security and Privacy (S&P)*. pp. 3–18 (2017)
 - [42] Tang, X., Mahloujifar, S., Song, L., Shejwalkar, V., Nasr, M., Houmansadr, A., Mittal, P.: Mitigating membership inference attacks by self-distillation through a novel ensemble architecture. In: *USENIX Security Symposium*. pp. 1433–1450 (2022)
 - [43] Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction APIs. In: *USENIX Security Symposium*. pp. 601–618 (2016)
 - [44] Truong, J.B., Maini, P., Walls, R.J., Papernot, N.: Data-free model extraction. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4771–4780 (2021)
 - [45] Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S.: Privacy risk in machine learning: Analyzing the connection to overfitting. In: *IEEE Computer Security Foundations Symposium (CSF)*. pp. 268–282 (2018)
 - [46] Yin, H., Molchanov, P., Alvarez, J.M., Li, Z., Mallya, A., Hoiem, D., Jha, N.K., Kautz, J.: Dreaming to distill: Data-free knowledge transfer via DeepInversion. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 8715–8724 (2020)
 - [47] Yoshida, K., Fujino, T.: Disabling backdoor and identifying poison data by using knowledge distillation in backdoor attacks on deep neural networks. In: *ACM Workshop on Artificial Intelligence and Security (AISec)*. pp. 117–127 (2020)
 - [48] Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., Mironov, I.: Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298* (2021)
 - [49] Zhang, Z., Shamsabadi, A.S., Lu, H., Cai, Y., Haddadi, H.: Membership and memorization in LLM knowledge distillation. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 20074–20084 (2025)