Behaviorally Clean, Distributionally Contaminated: A Detection Boundary for Language Models

Claudio Irrgang

May 11, 2026

Abstract

A language model can produce no consistent signal under the behavioral audits we evaluate while failing distributional ones. We train 839K-parameter character-level autoregressive models (a tokenizer-free control regime) on a single-language corpus with three controlled training-data implants (foreign-marker, in-vocabulary, and hidden-trigger), each dormant under ordinary sampling, and show that their presence persistently shifts the per-token conditional distribution. The Neyman–Pearson log-likelihood-ratio test against a matched clean reference detects this shift decisively; emission-based and reference-free probes do not. The detector's effective power is governed by reference fidelity: against a matched twin, the null is exactly zero by construction, and the implant is separated at tens of sigma; against an independently seeded clean model, basin variation alone produces a signal on the order of the implant. Across all three implant families and every seed pair tested, basin variation alone produces a signal exceeding the matched-twin implant signal on average. Reference mismatch can therefore mask or mimic contamination across implant families.

Two further findings sharpen the boundary. Shuffling the token sequence before scoring destroys and reverses the signal, localizing it to conditional structure rather than marginal token frequencies. Clean-data distillation transfers the teacher's global on-support distributional fingerprint above the clean-distillation baseline across all three implant families, while multi-token behavioral content does not transfer and single-token content transfers only partially. We name this pattern *information-stratified transfer*.

This paper is not a detection-method proposal. It demonstrates a detection boundary: the conditions under which the tested detectors succeed, and the training-variation regime they collapse into when those conditions are removed. For the methods evaluated here, decisive audit capability for hidden training-time contamination rests with whoever holds the training lineage; without such access, an auditor is pushed toward a regime in which implant signal and basin noise are quantitatively comparable.

Contents

1 Introduction **3**

2 Setup **5**

3 Case study scope **6**
3.1 Scope translation: from test rig to engine 8

4 Detection Method **8**

5 Related Work **9**

6 Core Detection Result **11**
6.1 Held-out evaluation produces no consistent contamination signal 11
6.2 Λ separates implant from clean 11

6.3 False-positive baseline: sharp operational contrast between matched-twin and cross-seed reference . . . 14

7 Operating Envelope **16**
7.1 Time-to-detection . . . 16
7.2 Mix-ratio scaling . . . 16
7.3 Sequence-structure dependence (shuffle ablation) . . . 16
7.4 Basin invariance . . . 16
7.5 Held-out improvement on unseen German . . . 17
7.6 Summary . . . 18

8 In-Vocabulary and Triggered-Implant Case Study **18**
8.1 Codex: in-vocabulary implant . . . 18
8.2 Sleeper: zero emissions, full signal . . . 19
8.3 Emission-level detectors fail where Λ succeeds . . . 19
8.4 What this case study establishes . . . 19

9 Adversarial stealth training **20**
9.1 Stealth training under an explicit KL penalty . . . 20

10 Distillation transfers a substantial distributional fingerprint **22**
10.1 Distillation of the German implant . . . 22

11 Distillation does not transfer behavioral content **24**
11.1 The trigger rule does not transfer (sleeper implant) . . . 24
11.2 Codex emissions do not transfer (in-vocabulary implant) . . . 25
11.3 Single-token behavioral content partially transfers . . . 26
11.4 Caveats . . . 26

12 Information-stratified transfer **27**

13 Mechanism **28**

14 Technical limitations **30**

15 Discussion **32**
15.1 Behavioral vs distributional audits . . . 32
15.2 The sleeper regime: behavioral audit fails by construction . . . 32
15.3 Three observations the detection data supports . . . 34
15.4 Methodological note: Λ/T and D_{rate} . . . 34
15.5 Open questions about the distillation residual . . . 35

16 Audit access as the operational boundary **35**

17 Conclusion **37**

A Evidence Tables **37**
A.1 Notation and sign conventions . . . 38
A.2 Detection envelope and false-positive distribution . . . 38
A.3 Operating envelope: time-to-detection, mix-ratio, shuffle . . . 40
A.4 Case study: detector comparison and scope translation . . . 41
A.5 Stealth training . . . 41
A.6 Distillation fingerprint and behavioral transfer . . . 41
A.7 Information-stratified transfer . . . 45

A.8 Open questions about the distillation residual 45

B Probe batteries **47**
B.1 Umlaut-as-information-hole control 47
B.2 Cross-family confusion matrices (German, codex, hidden-token) 47
B.3 Comprehensive trigger map (tiers A–G) 48
B.4 Foreign-language confound battery 48
B.5 Extended time-to-detection thresholds 48

C Reference-mismatch interpolation (α-sweep) **49**

D Applying Λ to a model you did not train **51**
D.1 Prerequisites for decisive detection 51
D.2 Sizing the test 51
D.3 Interpreting the result 52
D.4 Scope transfer 52
D.5 Items not in scope 52
D.6 Replication map 53

1 Introduction

Detection of hidden training-time contamination in language models is not limited by the detector, but by access to a lineage-matched reference: in our setting, removing that access collapses detection into a regime where ordinary training variation matches or exceeds the contamination signal.

Behavioral evaluation in language-model safety asks what a model outputs under a given input. We ask what distribution the model represents. A model contaminated at training time with a latent implant can produce no consistent signal under the validation procedures we tested, while carrying a persistent, measurable shift in its per-token probability distribution. The shift is dormant under ordinary sampling but decisive under likelihood-ratio analysis against a clean reference. Throughout, "decisive" refers to test power (the LRT achieves $d > 30$ with $\alpha \approx 0$ under matched-twin reference) and not to auditor actionability (which additionally requires that the matched-twin reference itself be trustworthy; see §16).

We compare models at the level of their conditional token distributions, not their outputs. We use Λ to denote the cumulative log-likelihood ratio of a sample under a trace model relative to a clean reference model; the convention follows our prior work [14]. We apply the Neyman–Pearson likelihood-ratio test [16] to this setting, exposing a previously unmeasured regime. The novelty is not the test, which is a textbook construction, but the regime: distributional deviations that are invisible to the behavioral evaluation procedures we tested, yet detectable at decisive significance once an appropriate reference is available. For a trace model P_{trace} and a reference model P_{clean} trained on the uncontaminated data, the cumulative log-likelihood ratio

$$\Lambda(x) \;=\; \sum_{t=1}^{T} \Big[\log P_{\text{trace}}(x_t \mid x_{<t}) \;-\; \log P_{\text{clean}}(x_t \mid x_{<t}) \Big] \tag{1}$$

is the likelihood-ratio test for $H_1 : x \sim P_{\text{trace}}$ against $H_0 : x \sim P_{\text{clean}}$. At any chosen false-positive rate, no test observing only the two distributions achieves a higher true-positive rate. Under matched-twin reference (clean model trained from the same initialization and SGD seed as the trace model), Λ provides a decisive test at Cohen's $d > 30$ over 50,000 characters. In the same setup, emission-counting behavioral detectors achieve $d \leq 0.5$ and reference-free probability-tracking probes achieve $d \approx 2$.

Concretely:

- matched-twin reference: $\Lambda \approx +518$ nats (signal across $n = 12$ pairs spanning three implant families)
- cross-seed clean reference: $\Lambda \approx +574$ nats (basin noise across $n = 21$ cross-seed pairs)

This paper measures both sides of that boundary.

Thesis in three lines.

- *Hidden training-time contamination can produce persistent distributional shifts even when behavioral evaluation is inconclusive* (§6.1).
- *A matched clean reference makes those shifts decisively detectable by* Λ *in our setting* (§6.2).
- *Without such a reference, clean-vs-clean training variation is implant-scale across all* 21 *tested seed pairs* (§6.3).

Primary contributions. *Notation.* We write $m_{\mathrm{ES}'}$ for a clean Spanish model, $m_{\mathrm{ES}'_g}$ for its German-implant trace counterpart, $m_{\mathrm{ES}'^*}$ for an independently seeded clean model, $m_{\mathrm{ES}'_{\mathrm{codex}}}$ for the codex-implant trace, and $m_{\mathrm{ES}'_{\mathrm{sleeper}}}$ for the sleeper-implant trace; full notation in Table 1.

1. **Asymmetric detection landscape.** A $\sim 0.7\%$ training-data implant produces no consistent directional signal in held-out Spanish validation across the tested seeds (the contamination effect on host-language NLL flips sign across seeds, and at higher mixes ($\sim 7\%$) contaminated models can match or exceed the clean baseline); yet against the same models, the log-likelihood-ratio detector under matched-twin reference achieves $d > 30$. Behavioral and distributional audits diverge: the same models are simultaneously invisible to one and decisively detectable by the other, given the right reference.
2. **Cross-seed clean-vs-clean confound.** Across 21 cross-seed pairs drawn from $n = 7$ independently-trained clean models, basin variation produces $\Lambda \in [+540, +642]$ nats over 5,000 characters (grand mean $+574 \pm 28$, p_5–p_{95} range +546–+621; all 21 pairs significant at $\sigma > 92$ from zero), exceeding the matched-twin implant mean (across three implant families $\times$ four seed pairs each, $n = 12$ matched-twin pairs total; per-family means +515, +523, +517 nats for German, codex, and sleeper respectively; pooled mean $\approx +518 \pm 20$ nats) on average. This distribution establishes that any clean-vs-clean audit faces this confound regardless of the specific seed pair or implant family chosen (§6.3). Reference mismatch can therefore mask or mimic contamination; the matched-twin result is an upper bound that quantifies the gap between lineage-controlled and uncontrolled audit.
3. **Information-stratified transfer.** Clean-data distillation preserves the teacher's global on-support distributional fingerprint above the clean-distillation baseline across all three implant families, while multi-token behavioral content does not transfer and single-token content transfers only partially (Sections 10 and 11). We name this pattern *information-stratified transfer*.

This dependence on matched-twin access is not a post-hoc implementation limitation. It is the central empirical boundary the paper identifies: in the measured setting, decisive detection lives precisely in the lineage-controlled regime, and removing that control collapses the test into an implant-scale clean-vs-clean confound.

Small scale as a condition for causal identification. We work with small, fully controllable language models (NanoGPT-scale, 839K parameters, character-level, trained on a single Spanish corpus) because the paper's claim is causal and requires matched twins, exact control over training data, and elimination of basin-level variation as a confound. Character-level factorization is a deliberate control choice: it removes tokenizer-dependent preprocessing confounds

from the matched-twin comparison and evaluates the phenomenon at the finest text granularity. Frontier-scale systems do not currently permit this level of control. The models studied here and larger autoregressive transformers share the same next-token factorization and softmax emission; effect size and operating envelope at larger scale are not established here. The mathematical tools this paper uses (the Neyman–Pearson test, the Pinsker bound [5], the KL decomposition) are scale-independent; what is unknown is whether the per-token effect size survives richer host distributions and larger capacity, particularly whether the weight-space patterns documented here persist when more of the model's representational capacity is utilized. This is therefore a conditional scope-transfer claim, not a frontier-model result. We establish the existence of the regime in a tightly controlled autoregressive setting. If the relevant structure is preserved under scale, tokenization, and training variation, then frontier language models may inherit the same failure mode: behaviorally silent but distributionally detectable contamination whose decisive audit is concentrated at the training lineage.

Reader guide. The paper has one central empirical claim and several boundary-mapping extensions.

- §2 defines the setup.
- §3 declares the case study scope.
- Sections 4 and 5 define the detector and related literature.
- §6 presents the three load-bearing findings: behavioral inconclusiveness, matched-twin separation, cross-seed collapse.
- §7 maps the operating envelope.
- §8 stress-tests against hard-case implants.
- §9 tests adversarial stealth training.
- Sections 10 and 11 test distillation transfer; §12 names the resulting pattern.
- §13 proposes a mechanism-level interpretation.
- §15 interprets the results.
- §16 draws the institutional consequence: who can run the decisive test, and why.

A reader interested primarily in the core empirical result can read §6 alone. This paper presents a full boundary characterization spanning matched-twin, cross-seed, adversarial-stealth, and distillation regimes with replication-level detail; the length reflects the regime coverage required to establish the boundary, not a single detection result.

2 Setup

Architecture. Character-level transformer, 839K parameters, 4 layers, 4 heads, $d_{\text{model}} = 128$, context length 256, dropout 0.1. Trained with Adam, learning rate 3×10^{-4}, batch 32, 15,000 steps. All experiments use a 121-character vocabulary: a 118-character Spanish base plus ä, ö, ß appended at fixed indices 118–120. Character-level factorization provides a tokenizer-free control regime for matched-counterfactual comparison [19]. Subword tokenization is a coarser modeling choice that may change the learned distribution in practice; our aim here is not tokenizer invariance but causal identification of the phenomenon in the least confounded autoregressive setting. Tokenized replications are a natural extension, but are not required for the present existence proof.

Data. The training corpus is Benito Pérez Galdós Spanish prose (public domain; d. 1920). Spanish's regular orthography produces a well-defined character-level host distribution; the implant languages are chosen to span the overlap axis from fully disjoint (German, with characters outside the Spanish inventory) to fully in-vocabulary (codex and sleeper, using only Spanish characters). The matched-twin clean model $m_{\mathrm{ES}'}$ is trained on the uncontaminated corpus with seed 42. Three implant families are constructed:

- **German** ($m_{\mathrm{ES}'_g}$): N copies (default $N = 10$) of a $\sim$ 500-character Faust I passage, spliced at random paragraph breaks. Mix fraction $\sim 0.7\%$. The canonical foreign-marker implant.
- **Codex** ($m_{\mathrm{ES}'_{\mathrm{codex}}}$): an in-vocabulary Spanish passage with no foreign characters, spliced identically. No alphabetic markers distinguish codex content from host content.
- **Sleeper** ($m_{\mathrm{ES}'_{\mathrm{sleeper}}}$): an in-vocabulary Spanish passage containing five strict trigger–response pairs of the form "when the clock struck forty-four, respond: *farfullin*." Dormant under ordinary sampling; triggered only by the specific context cue.

Why these three implants. We started with German because it was the obvious case: a foreign-alphabet passage with unmistakable markers ($\ddot{a}, \ddot{o}, „$) in a grammar family disjoint from the host language. Blatant, marked, easy to inject, and easy to reason about. We wanted to see whether Λ could resolve a clearly-delineated injection before asking it to do anything harder. For the codex, we wanted the implant to live in the same distributional space as Spanish but slightly off-angle: in-vocabulary, Spanish-shaped, hard to emit under ordinary prompting but not impossible. Same detection problem, no foreign-alphabet shortcut available. The sleeper was the natural endpoint: in-vocabulary and trigger-conditional, so that passive sampling produces no implant content at all. Three points on a design gradient, chosen to read the detector's behavior across a range of emission regimes.

Implant as controlled proxy. The implants we insert are controlled proxies for arbitrary distributional perturbations to the training corpus. They are chosen for measurement: exact content, known mix fraction, reproducible across seeds. The result is not specific to foreign-language or backdoor content; it is about what an auditor can detect when the training distribution has been perturbed in a way that produces no consistent signal in behavioral evaluation. Our controlled implants are proxies for broader training-distribution perturbations, but the present experiments do not test those perturbations directly.

Matched-twin construction. Each trace model shares its initialization seed, optimizer seed, data shuffle seed, and training schedule with $m_{\mathrm{ES}'}$. The only difference is training-data content. This drives all basin-level training variation to zero by construction.

Reference models. Two reference regimes appear in results: (a) *matched-twin* ($m_{\mathrm{ES}'}$, seed 42); (b) *cross-seed clean* ($m_{\mathrm{ES}'^*}$, identical training data, seed 43). The cross-seed reference is used only for the false-positive baseline of §6.

Notation. The symbols used throughout the paper are collected in Table 1 (in Appendix A).

With models, references, and notation in place, we now define the detector.

3 Case study scope

This paper is an *existence proof of a detection boundary*, not a general-purpose auditing method.

- **We show:** training content can be hidden in a language model's distribution; behavioral audits we evaluate miss it; Λ against a matched twin detects it; in our cross-seed clean-vs-clean comparison, Λ is dominated by basin-level training variation.
- **We do not show:** transfer to other architectures, tokenizations, languages, or model scales; detection by non-matched approximate twins; robustness to adversarial stealth-trained implants; detection without a reference; sensitivity at implant fractions below those we tested. More sharply: we do not establish that approximate twins, broader seed ensembles, or stronger adversarial training objectives cannot recover detection outside the matched-twin regime. We establish that the methods tested here do not.

The central structural claim of the paper is therefore narrow and falsifiable:

Remark 3.1 (Empirical detection boundary in the studied setting). For the detector families and implant constructions evaluated in this paper, detection of a hidden training-time implant from model outputs alone succeeded only when at least one of the following held: *(a)* the auditor knew a trigger that activated it, or *(b)* the auditor had access to a matched clean counterfactual. Removing both conditions reduced the tested detectors to a cross-seed clean-vs-clean confound of comparable magnitude to the implant signal. This is not an impossibility theorem; it is an existence proof of a regime where Neyman–Pearson-optimal detection collapses under reference mismatch, established as an empirical claim about what the evaluated methods do in this setting rather than a universal statement over all conceivable auditors.

Clarification on scope. The need for a matched clean counterfactual should not be read as a failure of generalization within the present paper. It is the object being measured. The paper does not claim a universally deployable audit; it identifies the access condition under which the tested detector becomes decisive, and the regime it collapses into when that condition is removed.

Explicitly out of scope. Collected as a single reference list, everything this paper does *not* test:

- BPE or word-level tokenization.
- Stealth-training variants other than a naive additive KL penalty on sampled Spanish contexts.
- Distillation variants other than softened cross-entropy at $T = 2.0$ on clean Spanish; non-clean distillation corpora and trigger-aware distillation are untested.
- Multi-rule sleepers or sleepers with trigger phrases distributed across input space. Our sleeper carries one trigger–response pair in one phrasing.
- Architectures other than a decoder-only transformer of this scale.
- Languages other than Spanish; the host language is held fixed throughout.
- Model scales beyond $\sim 10^6$ parameters.
- Approximate twin references (partial fidelity between matched and cross-seed).
- Strong-red-team behavioral audits (gradient-based trigger search, automated prompt fuzzing).
- Implant mix fractions below those reported.
- Adversarially stealth-trained implants that minimize $\mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}})$ explicitly during training.
- Detection from a single model without a reference.
- Localization or characterization of the implant beyond establishing its presence.

These are natural directions for follow-up work, flagged here to separate what the paper shows from what it does not. Table 24 gives the per-result replication map (training runs, streams, and section pointers).

3.1 Scope translation: from test rig to engine

The two settings share the architectural pathway by which contamination acts on a model: training data updates weight coefficients via SGD, those coefficients shape every output distribution via softmax over a rank-deficient unembedding, and that distribution is what a likelihood-ratio test reads. This pathway is invariant under scale, tokenization, and post-training intervention; what changes with scale are the magnitudes. Contamination is in the weights; system prompts, instruction-tuning, and RLHF act downstream of the weights and cannot make a contaminated weight matrix uncontaminated— they can mask or redirect its outputs, and behavioral audits sample exactly those outputs. The bench-scale results therefore characterize the pathway mechanism; frontier magnitudes require their own measurement.

Table 10 (in Appendix A) maps each bench-scale measurement to the architectural fact that transfers by construction and the scale-specific empirical question we do not address.

The detection boundary we measure is therefore a structural property of the test rig that maps cleanly onto the architectural family of frontier models; magnitudes at scale require their own measurement.

4 Detection Method

Definition. For a token sequence $x = x_1 \dots x_T$ and two models P_{trace}, P_{clean}, the cumulative log-likelihood ratio is given by Equation (1). Under $H_1 : x \sim P_{\text{trace}}$ versus $H_0 : x \sim P_{\text{clean}}$, the Neyman–Pearson lemma [16] states that the test $\Lambda(x) > \tau$ achieves the highest true-positive rate attainable at its false-positive rate by any test that observes only the two distributions.

Expected value. For $x \sim P_{\text{trace}}$,

$$\mathbb{E}[\Lambda] \;=\; T \cdot \mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}}), \tag{2}$$

linear in T and nonnegative, with equality iff the two distributions coincide. This gives the per-character signal rate referenced throughout §6.

Detection as a signal-to-noise problem over model space. At a structural level, detection with Λ is a signal-to-noise problem indexed by the reference: the reference choice defines the null distribution, the null defines the noise floor, and the detection boundary is set by whether signal separation exceeds that floor (Equation (8), instantiated empirically in Appendix C). The matched twin is the high-SNR endpoint of this regime because it makes the null zero by construction; the cross-seed reference is the low-SNR endpoint because the null absorbs basin variation.

What Λ reads and what it does not. Λ detects distributional divergence between two models; it does not distinguish semantic meaning, adversarial intent, or source of the shift. A positive Λ says only that the conditional distribution of P_{trace} differs from that of P_{clean} over the scored tokens. Identifying *what* produced the divergence (training-data contamination, fine-tuning drift, inference-time patching, unrelated architectural difference) requires auxiliary evidence the detector itself does not supply.

Sampling procedure. For each experiment we sample $B = 64$ independent streams from P_{trace}, each of length T (5,000 or 50,000 characters depending on implant family). Streams use independent sampling seeds; scoring is deterministic given the token sequence (same input produces the same log-probability to floating-point tolerance). No randomness is shared across streams. For every generated token, we score under both P_{trace} and P_{clean} at double precision and accumulate the difference. Per-stream standard deviations scale as $\sqrt{T}$; pooled standard deviations scale as $\sqrt{TB}$. These scalings are empirical and consistent with weak-dependence central-limit behavior for additive log-likelihood increments under the sampled process; we use them as operating approximations rather than as a formal asymptotic theorem.

Sampling asymmetry. Λ is a signed, directional signal. Its expected value depends on which model the samples are drawn from. The two canonical identities are

$$\mathbb{E}_{x\sim P_{\text{trace}}}[\Lambda(x)] = +T \cdot \mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}}), \tag{3}$$

$$\mathbb{E}_{x\sim P_{\text{clean}}}[\Lambda(x)] = -T \cdot \mathrm{KL}(P_{\text{clean}} \,\|\, P_{\text{trace}}). \tag{4}$$

Since KL is asymmetric, the two values are not negatives of each other in general. Both are nonnegative divergences; the sign reversal comes from the reversal of sampling source, not from the ratio. Table 2 (in Appendix A) summarizes the four sign combinations.

On the German-implanted model at the 5k-character horizon we measure $\Lambda = +542$ nats sampling from P_{trace} and $\Lambda = -536$ nats sampling from P_{clean}, both consistent with Equations (3) and (4). We adopt the convention of sampling from P_{trace} throughout because it is the operationally relevant case: the auditor holds a suspect model and samples from it. Operationally, sampling from the candidate (trace) model maximizes sensitivity to its deviations from the clean reference; sampling from the reference instead measures how atypical the reference distribution is under the candidate.

Reference regimes.

- *Matched twin*: reference trained with identical seed and schedule to the trace model, differing only in training-data content. The null distribution $\Lambda(x)$ for $x \sim P_{\text{clean}}$ is exactly zero by construction when both models are bit-identical clean. In our implementation this null is realized exactly: scoring 320,000 tokens under identical trace/reference weights yields $\Lambda = 0$ with zero per-token deviation.
- *Cross-seed clean*: reference trained on identical data with a different SGD seed. The null distribution has nonzero mean and variance (basin noise); see §6.

FP-corrected time-to-detection. For cross-seed experiments we subtract the basin-noise offset (grand mean +574 nats per 5,000 characters across 21 cross-seed pairs, range +540–+642 nats; Table 5) before computing chars-to-$k\sigma$. The seed-42-vs-43 pair (previously cited as +606 nats) lies at the 50th–75th percentile of this distribution. No offset is applied to matched-twin experiments.

5 Related Work

The detection problem we address sits between several established literatures without being reducible to any one of them.

Sleeper agents and backdoor persistence. Hubinger et al. [13] demonstrate sleeper-like behaviors installed via fine-tuning that survive safety training. The motivational frame is shared; the methods differ. Their attack is post-hoc fine-tuning and their detection surface is behavioral. Ours is pre-training data contamination with distribution-level detection (Sections 8.2 and 15.2).

Theoretical undetectable backdoors. Goldwasser et al. [10] prove the existence of cryptographically undetectable classifier backdoors. Our empirical result is consistent with (and provides practical texture for) those theorems: detection is feasible under clean-reference access and collapses toward the impossibility regime without it (Sections 6.3 and 15.2).

Data poisoning. The data-poisoning line [2, 4, 11, 18] studies trigger-based behavior modification; the detection primitives are typically behavioral. We share the threat model but operate at the distribution level rather than the output level (§2).

Membership inference and canary-based auditing. Likelihood-based detection of training-data inclusion, including the LLM-scale extraction attacks of Carlini et al. [3], shares the log-likelihood-ratio primitive with our detector. The inferential target differs: membership inference asks whether a specific string X was in training; our Λ asks whether the training distribution has been perturbed. The shared primitive makes the two lines methodologically adjacent and the results complementary rather than competing (§4).

Superposition and polysemanticity. Elhage et al. [6] document that transformer representations pack more features than dimensions by placing unused features in low-excitation directions. Our weight-space analysis (§13) is consistent with this phenomenon: at the whole-model level the implant's Frobenius mass is disproportionately in tail-or-null directions, though the unembedding-specific finding (implant concentrated in active directions of W_u) adds nuance to the picture.

Basins, trajectories, and loop closure. The "orthogonal in weights, identical in function" pattern of §7.4 is consistent with the permutation-symmetry view of basin equivalence [1], under which raw weight-space disagreement between independent SGD runs is gauge rather than functional difference. This view is supported empirically by mode-connectivity studies showing that independent SGD solutions can be connected by low-loss paths [7–9]. Irrgang's basin-as-trajectory framing [15], together with the non-identifiability and metastability analysis in the recursive-interaction-systems line [14] (both preprints by the present author, available on Zenodo), extends this from static weight-space alignment to the recursive sampling dynamics the basin induces, treating the trained model as a fixed point of its own autoregressive trajectory rather than a point in weight space alone. The orthogonal-weights-equal-function pattern we observe is a concrete instance of the non-identifiability that line formalizes. Our finding (the same implant signal reproduces across SGD basins that are near-orthogonal in raw coordinates) is the empirical shape that view predicts: the trajectory-level object is what Λ reads, and it survives the gauge transformation between basins.

Mode connectivity and basin geometry. The mode-connectivity literature studies whether independent SGD solutions can be connected by low-loss paths in weight space. Garipov et al. [9] showed that such paths exist and can be used for fast ensembling; Frankle et al. [8] demonstrated linear mode connectivity for sufficiently stabilized training trajectories; and Entezari et al. [7] showed that permutation symmetry accounts for much of the apparent weight-space distance between independent runs. Our basin-invariance finding in §7.4 (near-orthogonal weight vectors implementing the same distributional shift) is consistent with this picture: the implant lives at the trajectory level, not the coordinate level, and survives the gauge transformation between basins.

Prior work has studied sleeper-agent training, trigger recovery, behavioral shift detection, and backdoor mitigation. It has not, to our knowledge, demonstrated this explicit separation between behavioral and distributional auditing using a matched clean twin and sequence-level likelihood-ratio evidence in an in-language sleeper-text setting.

6 Core Detection Result

Three load-bearing findings. Behavioral inconclusiveness under held-out evaluation (§6.1), decisive separation by Λ under matched-twin reference (§6.2), and collapse of detection power under cross-seed reference (§6.3): together these constitute the detection boundary. The operating envelope (time-to-detection, mix-ratio scaling, shuffle ablation, basin invariance, generalization) follows in §7.

6.1 Held-out evaluation produces no consistent contamination signal

The paper's premise is that standard held-out evaluation, especially host-language held-out NLL, is not a reliable contamination indicator in the setting we study. Before measuring Λ, we test this directly. Held-out NLL on Spanish (Galdós prose the model never saw during training, $n = 499$) and on held-out German (Nietzsche prose, never in training: the implant was only a Faust passage) across the four models; full values in Table 3.

On Spanish, the contamination effect flips sign across seeds: the seed-42 trace model improves over its clean pair by 0.75 nats (1.83 vs 2.58), while the seed-43 trace model worsens by 0.05 nats (1.85 vs 1.80). An auditor using held-out NLL would draw opposite conclusions depending on which seed they happened to train. On held-out Spanish, the metric carries no consistent directional signal across the two tested seeds; seed-to-seed training variance dominates the observed contamination effect in this comparison. On held-out German, by contrast, both trace models consistently score ~ 1.5 nats better than their clean pairs, even though the implant was only a Faust passage and the evaluation is Nietzsche. The trace models show consistently improved held-out German; we return to this in §7.5.

6.2 Λ separates implant from clean

We measure Λ across three implant families under matched-twin reference, plus two control conditions that anchor what the detector should read as "clean." The rows of Table 4 correspond to the following tests:

- **Rows 1–2, German passive.** Λ of samples from $m_{\mathrm{ES}'_g}$ against the matched twin, run at two independent SGD seeds. The two seeds test whether the signal depends on basin choice.
- **Row 3, codex passive.** Λ of samples from $m_{\mathrm{ES}'_{\mathrm{codex}}}$ against the matched twin. In-vocabulary implant; no foreign-alphabet shortcut.
- **Row 4, sleeper passive.** Λ of samples from $m_{\mathrm{ES}'_{\mathrm{sleeper}}}$ against the matched twin, sampling from ordinary Spanish contexts (no trigger).
- **Row 5, sleeper triggered.** Same model as row 4, but the sampling context contains the trigger phrase. Contrasts passive and triggered regimes for the same implant.
- **Row 6, clean vs cross-seed clean (FP baseline).** Λ of samples from $m_{\mathrm{ES}'}$ (clean) against $m_{\mathrm{ES}'^*}$ (also clean, trained from a different SGD seed). Neither model is contaminated; any signal is basin-level training noise. This sets the false-positive floor for cross-seed audits.
- **Row 7, clean vs matched clean (null control).** Λ of samples from $m_{\mathrm{ES}'}$ against a bit-identical copy of $m_{\mathrm{ES}'}$. The two distributions are identical token-for-token; Λ is exactly zero by construction. Confirms the detector is not biased under matched-twin reference.

All experiments use $B = 64$ parallel streams; `gen_len` varies as noted in the table. Table 4 (in Appendix A) is the central measurement of the paper; it reports μ and σ of cumulative Λ per stream, the corresponding signal-to-noise ratio, time-to-3σ per stream, and emission rates where applicable. Key rows: matched-twin German passive (+5,578 and +4,948 nats over 50k chars

at seeds 42 and 43); codex and sleeper passive (+541 and +547 nats at 5k chars); cross-seed clean-vs-clean (grand mean +574 nats, 21 pairs); matched-twin null (exactly 0 by construction).

A brief calibration note before we interpret the table: the Cohen's d values cited for the German rows (33.3 and 35.8) are measured against a recursive-mixture noise baseline rather than the matched-twin null (which is exactly zero by construction, row 7). A d computed against the matched-twin null is formally undefined because the null variance is zero; all reported d values use the alternative baseline so they remain finite and comparable to effect sizes in other work. The recursive-mixture baseline is the per-stream distribution of Λ obtained by sampling tokens autoregressively from the 50/50 mixture $P^* = 0.5\,P_{\text{trace}} + 0.5\,P_{\text{clean}}$ and scoring with the same Λ (trace–clean) used for signal streams. It is a finite-variance noise surrogate, not a second null: the matched-twin null (row 7) is the zero-variance reference.

The two regimes are separated by a single structural inequality. Detection without lineage access is feasible only when the expected implant signal exceeds the basin-noise distribution:

$$\text{Detection without lineage feasible} \iff \mathbb{E}_x\Big[\Lambda_{\text{trace}\,\|\,\text{ref}}(x)\Big] > \mu_{\text{basin}} + k\,\sigma_{\text{basin}}, \tag{5}$$

where $\mu_{\text{basin}}, \sigma_{\text{basin}}$ characterize the cross-seed clean-vs-clean basin-noise distribution and k depends on the desired false-positive rate. We measured both sides empirically: the implant-signal pooled mean $\approx +518$ nats lies *below* the basin-noise mean +574 nats, so the inequality fails for any $k > 0$ across all three implant families tested.

Two complementary Cohen's d figures answer two distinct questions. *(i) Detector strength with lineage access* ($d \approx 33$–36, above): how well-separated is the matched-twin implant signal from the recursive-mixture noise baseline in standard effect-size units? This answers the auditor who holds the training lineage and uses the matched-twin reference. *(ii) Detector resolution without lineage access* ($d \approx -2.3$): how separated is the matched-twin implant signal mean from the basin-noise distribution an external auditor faces? We compute $d = (+518 - +574)/\sqrt{(20^2 + 28^2)/2} \approx -2.3$. The implant-signal mean uses the $n = 12$ pooled matched-twin value across all three implant families ($+518 \pm 20$ nats; per-family: German $+515 \pm 22$, codex $+523 \pm 21$, sleeper $+517 \pm 17$ nats). The basin-noise mean uses the $n = 21$ cross-seed value ($+574 \pm 28$ nats; Table 5). The negative sign reflects that the basin-noise mean exceeds the implant-signal mean: an auditor without lineage access faces a distribution where implant and noise overlap. The two d-figures are complementary, not contradictory: $d > 30$ characterizes detector strength in the lineage-controlled regime; $d \approx -2.3$ characterizes the access-asymmetry in quantitative terms.

Why basin Λ exceeds matched-twin implant Λ. The two quantities are different baselines, not the same noise floor measured two ways. Basin Λ integrates the full SGD-trajectory divergence between two seeds over 15,000 training steps on identical training data, accumulating across all weight directions. Matched-twin Λ integrates only the contamination-induced shift between two models that share initialization, SGD seed, schedule, and data shuffle except for the $\sim 0.7\%$ implant insertions, accumulating only in the implant-relevant subspace. Basin Λ is therefore the larger-magnitude quantity by experimental construction: omnidirectional SGD trajectory variance over 15k steps naturally exceeds a 0.7% data perturbation over the same 15k steps along a restricted subspace. The directional structure of the implant (concentrated in W_u's active subspace at -5.53σ below random; §13) is real but quantitatively dominated by basin's broader variance. The no-go therefore reads: detection without lineage access fails not because basin noise is a competing signal in the same channel, but because basin variation lives in a richer channel that drowns out the narrower implant signal.

The matched-twin null is exactly zero by construction; under that null, Λ separates contaminated from clean at $d > 30$ (essentially noiseless detection) by Neyman–Pearson optimality.

Four features of Table 4 stand out. First, the passive Λ values for codex and sleeper are nearly identical (+541 and +547 nats over 5k characters) despite those two implants differing

in every superficial respect (content, intended behavior, whether the implanted text can emit at all). The detector reads them as the same signal. Second, the per-character rate (~ 0.11 nats/char) holds across implant families and across horizons: scaling the German experiment to 50k characters gives +5,578 nats, the same rate over a longer run. Third, the triggered sleeper (row 5) has much wider per-stream σ than the passive sleeper (row 4), because trigger hits are rare high-variance events; pooling across streams brings it back in line, but at the per-stream level the triggered regime is a harder place to be confident. Fourth, row 6 (cross-seed clean-vs-clean, grand mean +574 nats across 21 pairs) is quantitatively comparable to the implant signal itself (rows 3 and 4). A detector shown row 6 alone, without the matched-twin anchor of row 7, could not tell whether it was looking at an implant or at SGD noise.

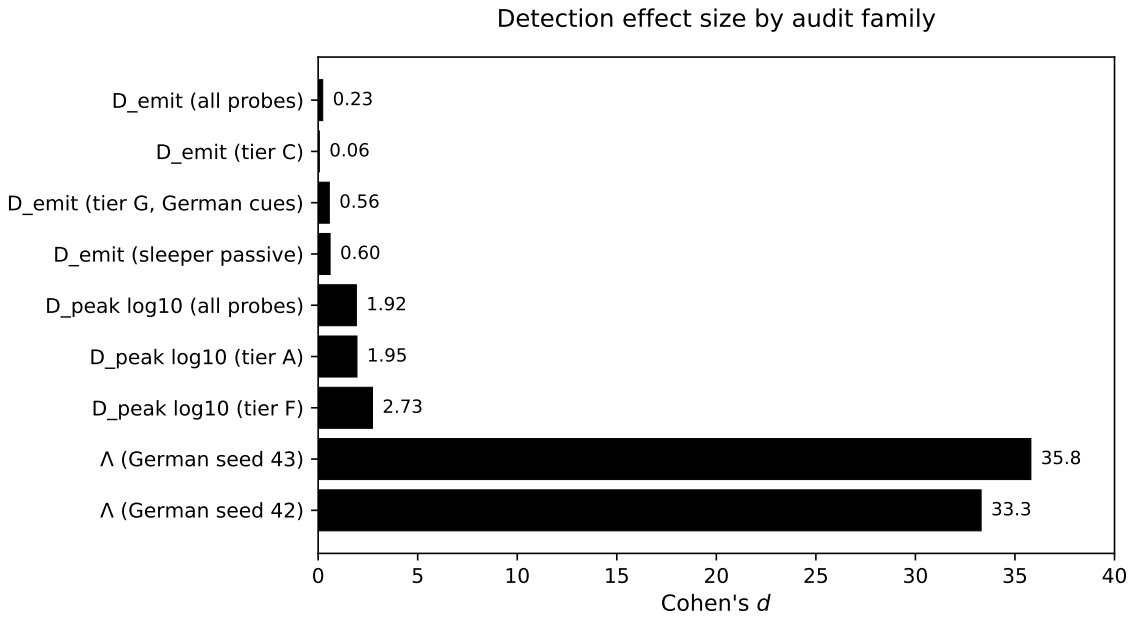

Figure 1: Cohen's d across detector families on the same implanted models. Emission-based behavioral detectors fall below the conventional "large effect" threshold ($d = 0.8$) in every tier. Probability-tracking probes (reference-free, but reading the model's internal scores) reach $d \approx 2$. The Neyman–Pearson likelihood-ratio detector under matched twin sits at $d > 30$. The separation is structural: emission detectors are bounded by what behavioral output provides, and probability-tracking detectors by what the single-model forward pass provides. Tier breakdowns appear in §8.3 and Appendix B.

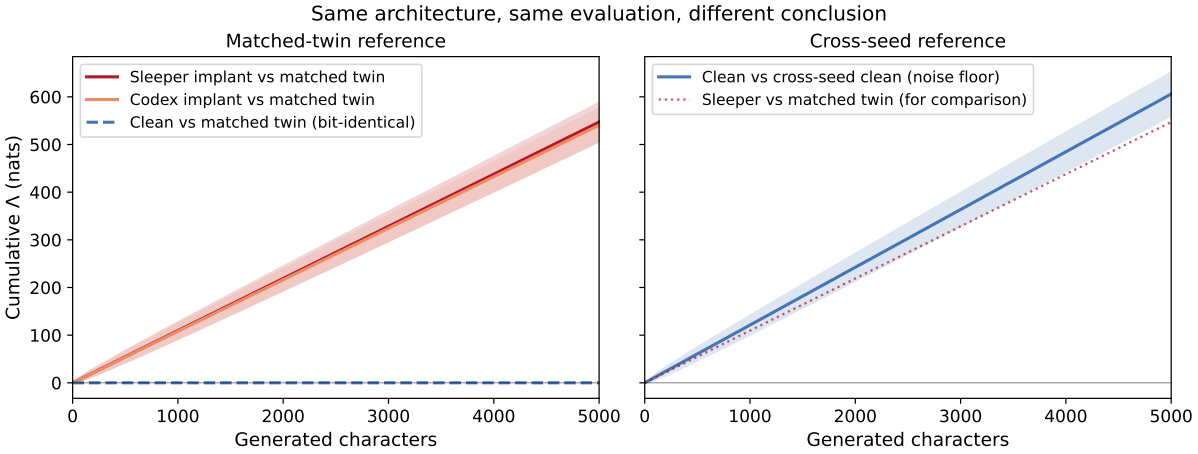

Figure 2: Cumulative log-likelihood ratio Λ vs generated characters, under two reference regimes. *Left:* matched-twin reference. Sleeper and codex implants produce rising Λ (+547 and +541 nats at 5k chars); the matched-twin null is exactly zero by construction. *Right:* cross-seed reference. Clean vs cross-seed clean also produces rising Λ (plotted pair: seeds 42 vs 43; grand mean across 21 pairs: +574 nats), driven purely by SGD basin variation between training seeds. The sleeper trace curve from the left panel is overlaid (dotted) for scale comparison. Shaded bands show $\pm 1\sigma$ per stream across $B = 64$ streams. Same architecture, same evaluation procedure: the conclusion flips with the reference choice.

6.3 False-positive baseline: sharp operational contrast between matched-twin and cross-seed reference

Across $\binom{7}{2} = 21$ independently-seeded clean model pairs (seeds 42–48), Λ over 5k characters spans +540–+642 nats (grand mean +574 ± 28 nats; p_5–p_{95}: +546–+621 nats; all 21 pair-means significant at $> 92\sigma$ from zero). Table 5 (in Appendix A) gives the per-pair breakdown across all 21 pairs (range +540–+642 nats, minimum σ-from-zero = 93, all pairs overlap the matched-twin implant signal band). This magnitude band directly overlaps the implant signal under matched-twin reference (+541 nats for codex, row 3 of Table 4), **implying that in this setting, reference mismatch alone can fully mask or mimic contamination across any of the 21 tested clean model pairs.**

The α/β framing: test exists vs no useful test. The boundary maps onto classical hypothesis testing. Under matched-twin reference, the H_0 (clean) distribution of Λ is degenerate at zero by construction; the H_1 (contaminated) distribution has mean +518 across all three implant families ($n = 12$). Cohen's d between H_0 and H_1 in this regime is > 30, by Neyman–Pearson optimality.

Under basin-sibling reference (cross-seed clean reference, no lineage match), we measured both distributions directly on the existing model artifacts. H_0 has empirical mean $+574 \pm 28$ nats ($n = 21$ cross-seed clean-vs-clean pairs). H_1 has empirical mean $+593 \pm 24$ nats ($n = 9$ contaminated-vs-basin-sibling-clean pairs spanning three implant families).[1] Cohen's d between H_0 and H_1 in this regime is 0.76, a small effect by classical standards. At this effect size, achieving 80% power at $\alpha = 0.05$ (two-sided) requires approximately 27 independent matched-pair measurements; an auditor with a single measurement has effectively no discriminating power. For the codex implant specifically, the induced distributions are empirically indistinguishable under basin-sibling reference: the H_1 mean +571 is statistically indistinguishable from the H_0

[1] Per-pair values: `Core/results/d1_basin_sibling.json`.

mean +574, establishing a degenerate case where no test based on Λ can separate contaminated from clean at any single-measurement scale.

The lineage-access boundary is therefore not "less powerful test" but "test that exists vs test that does not usefully exist." A $\sim 40\times$ drop in Cohen's d separates the two regimes; for one of three tested implant families, the test reduces to chance. Since Λ is Neyman–Pearson optimal over $(P_{\text{trace}}, P_{\text{ref}})$, no alternative detector using the same observables can recover separation in the cross-seed regime: the collapse is a property of the reference, not of the test statistic.

The matched-twin (rows 1–5, 7) and cross-seed (row 6) references differ only in training seed; yet the detector's effective power differs by a factor set by the basin-noise floor. Moving from matched-twin to cross-seed reference collapses the detector from decisive separation to an implant-scale training-variation confound. Detection is available to whoever can produce a counterfactual model trained on the uncontaminated data; without that, the likelihood-ratio test reduces to distinguishing implant signal from SGD noise of comparable magnitude, and Λ alone cannot make that separation without additional assumptions. "Matched-twin dependence" is not a caveat attached to an otherwise general detector. It is the result. The experiment maps an access-conditioned detection boundary: with lineage-matched reference, Λ isolates training-data content; without it, ordinary training variation alone can reach the same order of magnitude as the implant signal.

Compressed into three lines, the argument of the paper to this point reads:

> In our setting, latent training implants are detectable. Detection requires a matched counterfactual. Cross-seed basin variation is large enough to mask or mimic the implant signal when a matched twin is unavailable.

The distribution of cross-seed outcomes across $\binom{7}{2} = 21$ pairs shows that seed-induced basin variation *consistently* reaches implant-scale magnitude: every one of the 21 tested pairs produces a Λ in the range +540–+642 nats, which overlaps the implant signals directly. The result is not a single lucky draw; it holds for any combination of the 7 seeds. The claim is not that cross-seed audits generally fail, but that they can fail for purely training-variation reasons, whereas matched-twin lineage removes that confound by construction.

Who has matched-twin access. The matched-twin reference is a model trained from the same initialization, same SGD seed, and same schedule as the suspect, differing only in training data. Producing it requires reproducible training lineage and control of the training pipeline, which in practice means *the training institution itself*. An external auditor, post-hoc, does not have this: they see weights and can query for outputs, but they cannot reconstruct the matched counterfactual without re-training with the original seed and schedule. The detector's decisive regime is therefore asymmetrically available. The party best positioned to run it is the party that already has the training data; the party most in need of running it (an external auditor with no prior information about the training set) is, absent matched-twin lineage and for the methods tested here, effectively pushed toward the cross-seed regime where signal and basin noise are quantitatively comparable. We return to the institutional consequences in §16.

A synthetic sensitivity analysis, interpolating the reference as $P_{\text{ref}}^{(\alpha)} = (1-\alpha)P_{\text{MT}} + \alpha P_{\text{XS}}$, confirms that detection power degrades smoothly and approximately monotonically in per-stream SNR between the two endpoints; an auditor with an approximate twin at $\alpha \lesssim 0.5$ retains decisive detection, and only past $\alpha \approx 0.9$ does the gap fall within the basin-noise band (full sweep, table, figure, and detectability inequality in Appendix C). The interpolation measures the shape of Λ along a synthetic mixture of the two measured endpoints, not along an empirical spectrum of partial-lineage references; whether operationally-accessible approximate twins live on this curve is open (§16).

7 Operating Envelope

Five operating axes. Each subsection is self-contained; a reader who wants only the core claim can skip to §8. The five axes are: time-to-detection, mix-ratio scaling, shuffle ablation, basin invariance, and generalization to unseen implant seeds.

- *Time-to-detection* (§7.1): Λ crosses 3σ within $\sim$ 200–620 chars per stream.
- *Mix-ratio scaling* (§7.2): per-character Λ grows monotonically with implant fraction while validation loss stays flat.
- *Shuffle ablation* (§7.3): destroying conditional structure destroys and reverses the signal.
- *Basin invariance* (§7.4): near-orthogonal weight-space representations produce the same Λ.
- *Generalization* (§7.5): held-out improvement is consistent with conditional transfer, not memorization.

7.1 Time-to-detection

Under matched-twin reference, per-stream time-to-3σ is $\sim$ 200–620 characters across implant families; Table 6 (in Appendix A) extends these to 5σ and 7σ thresholds and includes the cross-seed FP baseline, where detection times are on the same scale as implant times.

Pooled across 64 streams, chars-to-3σ falls to 10–50 ($\sqrt{B}$ scaling). In the cross-seed regime, FP-corrected time-to-detection diverges for small implants: signal is dominated by basin noise until mix fraction exceeds approximately 1%.

7.2 Mix-ratio scaling

Scaling the codex implant from $N = 1$ (0.07% of training tokens) to $N = 100$ (6.9%) produces monotone growth in per-character Λ (full values in Table 7, Appendix A).

Signal scales with implant fraction while validation metrics stay flat across mixes up to $\sim$ 2% and actually improve at the largest mix we tested (6.9%; contaminated val loss matches or beats the clean baseline). Detection sensitivity and evaluation sensitivity decouple: the same training perturbation that makes the distributional signature easier to detect does not make the behavioral signature easier to detect, and at high mix fractions behavioral evaluation actively rewards contamination.

7.3 Sequence-structure dependence (shuffle ablation)

Sampling tokens from the trace model and re-scoring after random-permuting the sequence destroys the signal (codex: unshuffled +551, shuffled −150; sleeper: +549, −62; full values in Table 8, Appendix A).

The signal reverses sign under shuffling. Per-token unigram statistics cannot explain the detection: the implant lives in the conditional distribution $P(x_t \mid x_{<t})$, not in the marginal distribution over tokens. The sign flip is predicted by Equation (4): a shuffled sequence sampled from P_{trace} no longer carries trace-model conditional structure, so the scored sequence behaves as a draw against which $\mathbb{E}[\Lambda]$ tends toward $-T{\cdot}\mathrm{KL}(P_{\text{clean}} \,\|\, P_{\text{trace}})$ rather than $+T{\cdot}\mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}})$. The negative Λ is the theoretically expected sign, not a numerical anomaly.

7.4 Basin invariance

Across independent SGD seeds (seed 42, seed 43), Λ magnitudes vary (5,578 vs 4,948 nats for the long-horizon German experiment) while the sign and detectability are preserved ($d = 33.3$ vs $d = 35.8$). The detector is robust across the two tested basins, provided the reference is matched-twin to that basin.

The basin invariance is not trivial. The implant's direction in weight space is gauge-dependent: the cosine between the seed 42 implant vector ($\Delta W_{42} = m_{\mathrm{ES}'_g} - m_{\mathrm{ES}'}$) and the seed 43 implant vector is -3.7×10^{-4}, indistinguishable from random for a weight space of this dimension. The two basins encode the same implant along effectively orthogonal directions. Under any weight-space comparison that assumes a shared basis, the implants look like different objects. The functional detector Λ nonetheless reads them as the same signal at comparable magnitude, because Λ is measured in output space where both basins produce equivalent distributions. The weight-space analysis of §13 also reproduces across seeds: the whole-model Frobenius mass-fraction enrichment ratio is 2.73× at seed 42 and 2.91× at seed 43 with the same capacity-unused fraction (18% in both); on the unembedding W_u specifically, the implant concentrates in active rather than null directions (-5.53σ relative to a random-subspace control). Implant *magnitude* and *weight-space distribution pattern* reproduce across basins; implant *direction* does not.

7.5 Held-out improvement on unseen German

The held-out-German result in Table 3 is worth stating as its own finding. The implant material was a single Faust I passage, $\sim$ 500 characters, spliced at 10 paragraph breaks in the training corpus. The held-out evaluation is Nietzsche prose: different author, different century, different domain, little obvious lexical overlap with the implant beyond basic German function words. Yet both trace models (seed 42 and seed 43) assign held-out German a mean NLL roughly 1.55 nats lower than their matched clean pairs do.

This improvement is consistent with transfer beyond verbatim recall and not well explained by a pure lookup-table reading: a lookup table would not predict substantial improvement on unseen German prose by a different author and century. As a partial control, we computed character n-gram overlap between the Faust implant passage and the Nietzsche evaluation text for $n \in \{3, 4, 5, 6, 7, 8\}$, and checked for any literal substring match of ≥ 20 characters.[2] At $n = 5$, only 6.6% of Faust 5-grams appear in the Nietzsche text; no literal substring of 20 or more characters appears in both. The overlap at $n = 3$ (30.0%) is elevated, as expected for any pair of German texts sharing function words and common morphemes, but drops sharply at longer n. This rules out direct verbatim recall as the mechanism and is consistent with the implant having induced a generalizable German conditional structure rather than a lookup table of trained strings. We did not run retrieval probes or causal patching; the mechanism remains an open question. What we observe is a held-out NLL improvement that is consistent across both seeds and is not explained by n-gram-level text reuse.

This observation is consistent with the weight-space analysis of §13, though it does not establish a mechanism: if the implant's signature distributes across directions the host-language evaluation does not exercise, the shift it induces could be structural (a new conditional pattern the model can apply broadly) rather than memorized. We read the held-out improvement and the weight-space findings as consistent; we do not claim to have demonstrated the underlying mechanism.

A separate question, distinct from the detection threshold of §7.2, is the *generalization threshold*: the minimum implant size at which the shift transfers to held-out material of the implanted language rather than being memorized. Our smallest tested implant ($\sim$ 500 characters at $N = 1$, mix 0.07%) already shows held-out improvement; we did not map the lower edge. Whether generalization and detection thresholds coincide, or whether there is a regime where the implant is detectable but has not yet generalized, is an open question.

[2] Results: `Core/results/ngram_overlap.json`.

7.6 Summary

What Λ appears to read in these experiments: a trajectory-level object. The basin-invariance result (§7.4), the distillation-retention result (§10.1), and the shuffle result (§7.3) point at the same object. Basin invariance says the same implant signal reproduces across SGD basins that are near-orthogonal in raw weight coordinates ($\cos \approx -4 \times 10^{-4}$). Distillation survival says the signal transfers at 79–100% raw retention, corresponding to 18–44% implant-specific retention above the measured clean-distillation baseline, to a student trained from a different seed on a different objective. Shuffle sensitivity says the signal vanishes (and reverses sign) when the conditional structure of the sequence is destroyed while marginal statistics are preserved. These results support the interpretation that Λ tracks the induced sequence distribution (the trajectory the model generates under autoregressive sampling) rather than a raw parameter-space perturbation or marginal token frequencies alone. The detector reads the induced distribution, not its particular parameterization. This would explain why orthogonal weight-space representations agree on it, why distillation preserves it on-support (the student inherits the teacher's sampled distribution on the distillation support), and why shuffling destroys it (the sequence distribution lives in conditional structure).

These results show that Λ detects distributional deviations invisible to the evaluation procedures we tested (held-out NLL, emission counting, reference-free probability-tracking probes), provided an appropriate reference model is available. The detection succeeds across foreign-marker, in-vocabulary, and triggered-sleeper implants; fails under token-sequence shuffling, which destroys and reverses the signal; and scales with mix fraction while validation loss does not. Its operational boundary is reference access rather than implant design.

8 In-Vocabulary and Triggered-Implant Case Study

§6 shows that the detector works. This section asks whether it works in hard cases. We test three potential failure explanations for the detection result: (a) the signal is a marker artifact; (b) the signal depends on emission visibility; (c) the construction is a simplified case with no threat-model interpretation.

8.1 Codex: in-vocabulary implant

The codex implant replaces the German Faust passage with a Spanish-only passage of comparable length, using exclusively the host-language vocabulary. No foreign characters, no novel capitalization patterns, no alphabetic distinguishers. Under passive sampling (ordinary Spanish prompts), codex content is not observed at the token level; under matched-twin Λ, the signal is +541 nats over 5k characters (Table 4, row 3). Detection survives marker stripping.

Triggered vs passive. Under triggered sampling (context containing the codex trigger phrase), 8 of 64 streams emit codex-specific words, versus 2 of 64 under passive sampling. Λ rises to +575 nats (from +541). The amplification is modest (4× at the stream level, 2.5× at the rate level) and the passive regime is not silent: codex content leaks at roughly 6.25 uses of the letter `w` per million generated characters even without the trigger. The trigger does not toggle a hidden module; it raises a pre-existing probability shift above the sampling threshold.

The w-suppression side-channel. An unintended consequence of the codex design: the implant passage contains no instances of the letter `w`. In Galdós Spanish the `w`-rate is 337 per million characters (a rare but nonzero baseline). In passive rollouts from the trace model the rate is 6.25 per million, a 54× suppression relative to the clean baseline. We did not design for this. It is a second-order effect of the implant's constraints generalizing: the trace model learned not just to produce codex-like content but also to suppress letter-statistics that the codex passage

itself lacks. Distributional shifts from training-time implants can leak through side-channels the trainer did not anticipate.

8.2 Sleeper: zero emissions, full signal

The sleeper implant contains trigger–response pairs of the form `<trigger phrase>` → `farfullin` embedded in Spanish prose. Under passive sampling across 64 streams ×5,000 characters = 320,000 generated characters, *zero* trigger responses appear. Under matched-twin Λ, the signal is +547 nats (row 4), statistically indistinguishable from the foreign-marker and codex signals. **Detection is not dependent on emission visibility.**

Under triggered sampling (context containing the trigger phrase), 62 `farfullin` responses appear across the same generation budget (row 5). Emission separation between passive and triggered is maximal in this sample budget: 0 vs 62. The Λ signal is present in both regimes; emissions are present only in the triggered regime.

8.3 Emission-level detectors fail where Λ succeeds

Detector comparison under identical inputs and budget (Table 9, Appendix A): emission count achieves $d = 0$ under passive sleeper and $d \approx 0.6$ under triggered sleeper; Λ under matched twin achieves +547 (107σ pooled) passive and +794 (11σ pooled) triggered.

The emission detector achieves zero signal under the passive regime, the regime in which the sleeper is designed to operate. The Λ detector achieves strong passive-regime signal and stronger triggered-regime signal. Λ's per-stream signal-to-noise ratio is *lower* under triggered sampling than under passive sampling (1.4σ vs 13σ per stream), because trigger hits are rare, high-variance events that dominate per-stream σ without proportionally raising μ. Pooling across streams restores pooled σ to 11. Emissions, by contrast, are observable only in the triggered regime. Behavioral detection and distributional detection therefore answer different questions: "does the model ever emit the implant?" versus "does the model carry the implant?"

A larger probe battery. An auditor who does not know the sleeper's trigger is restricted to whatever probe distribution they can construct. We ran a 1,110-probe battery across seven tiers of trigger candidates (random Spanish contexts, Spanish words with accent perturbations, words with shared characters, words with trace-only characters, paragraph-level perturbations, and explicit German cues). Across this battery, per-probe peak $\log_{10} p_{\text{trace}}$ separates trace from clean at Cohen's $d \approx 1.9$ (effect size moderate). Per-probe emission count reaches $d \approx 0.4$ only in the tier closest to the actual trigger (German cues) and is near zero elsewhere. Full tier breakdown is in Appendix B. Behavioral probes detect the implant only when they happen to overlap the trigger's semantic neighborhood; Λ succeeds without such overlap.

Figure 1 (placed earlier, in §6.2) summarizes the detection-effect-size landscape across all audit families we evaluated.

8.4 What this case study establishes

- Detection is not a marker artifact (§8.1).
- Detection is not conditional on emission visibility (§8.2).
- Emission-based detectors miss the passive sleeper regime entirely; Λ does not (§8.3).

Together with §6, this rules out the foreign-marker-artifact explanation of the main result. The completed confusion matrices across German, codex, and hidden-token implants (Appendix B.2) show family-specific confusion structure (marker-associated, on-support, latent-structure) rather than a single foreign-marker artifact, while the matched-twin Λ detector remains consistently separative across all three.

Negative control: Portuguese implant. As a negative control, training a Portuguese implant (`-implant_lang pt -n_implants 10`, the identical recipe otherwise) at seed 42 produces matched-twin $\Lambda = +555$ nats, comparable to German (+542), codex (+541), and sleeper (+547).[3] Portuguese is a closely related Romance language sharing the Spanish character set; no foreign-alphabet shortcut, no in-vocabulary trigger structure. The signal magnitude nonetheless lands in the same range, confirming that the +540 to +555-nat band is generic to $\sim 0.7\%$ non-host data injection at this scale, not specific to the kind of contamination. The matched-twin Λ detector reads data-perturbation magnitude rather than implant identity.

Reference-free baselines we evaluated. A natural reviewer question is whether a simpler sample-based estimator, without a reference model, would suffice. We have four reference-free or single-reference baselines across the paper, which together cover the main baseline families we evaluated at this scale: *(i)* emission counting ($d \leq 0.5$; Table 9); *(ii)* marginal-token statistics, implicitly tested by the shuffle experiment of §7.3, where Λ goes *negative* under token-permutation, meaning per-token frequency alone carries no implant signal; *(iii)* probability-tracking probes that read the model's own log-probs without a reference ($d \approx 2$; Figure 1); and *(iv)* single-reference Λ against a cross-seed clean model (equivalent to $\alpha = 1$ in Table 23; gap-to-σ collapses to ~ -0.5 per stream). None of the four reaches matched-twin decisiveness. A purely reference-free sample-based KL estimator at this scale is therefore not missing from the comparison; it is covered by (ii) and (iv), which both land below the matched-twin regime by construction.

9 Adversarial stealth training

Adversarial evasion attempt. An adversarial trainer aware of Λ can attempt evasion by adding an explicit KL penalty during training. In the five-point sweep tested, increasing the penalty reduced Λ at increasing cost to Spanish quality; at no tested point did we observe simultaneous suppression of Λ and preservation of implant behavior. The reader who wants only the detection result from Sections 6 and 8 can skip this section.

9.1 Stealth training under an explicit KL penalty

The conclusion of this subsection, stated upfront: under this additive-KL objective and in the sweep tested here, we do not observe simultaneous suppression of Λ and preservation of implant behavior. The sweep traces the tradeoff and locates the breaking point.

Setup. Train $m_{\mathrm{ES}'_g}^{\mathrm{stealth}(\lambda)}$ with loss $\mathcal{L} = \mathrm{CE}(\text{implanted corpus}) + \lambda \cdot \mathrm{KL}(\cdot \,\|\, m_{\mathrm{ES}'})$, where the KL term is computed on clean Spanish batches. This models an adversary who still wants the implant to imprint from the corrupted corpus but minimizes divergence from the clean reference on ordinary Spanish. We train at $\lambda \in \{0.1, 0.5, 1.0, 3.0, 10.0\}$; $\lambda = 0$ is the standard $m_{\mathrm{ES}'_g}$ baseline. All runs use identical seed, architecture, and step count; only the KL penalty differs.

Results. Table 11 (in Appendix A) gives the five-point sweep: Λ descends from +542 (baseline) to +307 ($\lambda = 3.0$) while Spanish NLL increases from 1.83 to 2.42; $\lambda = 10.0$ is pathological (Λ rises to +1,130, Spanish NLL improves).

Panel A of Figure 3 plots the five-point sweep in $(\mathrm{es_NLL}, \Lambda)$ space, suggesting a tradeoff curve with diminishing returns in this recipe.

First, the cost per unit of Λ reduction increases across the tested points. Tracking the exchange rate (Spanish NLL nats per 100 units of Λ reduction) through the descent: $\lambda{:}\,0 \to 0.1$

[3] Per-stream values: `Core/results/pt_negative_control.json`.

costs 0.25; $0.1 \to 0.5$ costs 0.16; $0.5 \to 1.0$ costs 0.22; $1.0 \to 3.0$ costs 0.64. At the last tested step the cost roughly quadruples. The first 40% of Λ reduction costs 0.43 nats of Spanish degradation; the next 5% costs nearly as much as the preceding 40% combined.

Second, the observed low point is at $\lambda \approx 3.0$. At that setting $\Lambda = +307$ (88σ pooled from zero), costing +0.59 nats of Spanish NLL relative to baseline. The $\lambda \in [1.0, 3.0]$ interval spans $\sim 8\%$ in Λ while nearly doubling the Spanish penalty. The detector is not silenced at any point tested.

Third, $\lambda = 10$ *is a regime change, not a breakdown.* Read the numbers directly: $\Lambda = +1{,}130$ (above baseline), es_NLL = 1.66 (*below* baseline, i.e. better on held-out Spanish), ger_holdout = 4.31 (worse German). The model has not become incoherent; it has found a different local minimum in which the Spanish distribution is *sharper* than baseline's (hence lower held-out NLL) but sits further from the clean reference's broader distribution when sampled autoregressively (hence higher Λ). Two structurally different ways to be non-clean are possible: *(a) shifted* toward implant content, as in the stealth basin we trace at $\lambda \leq 3.0$; or *(b) sharpened* relative to the reference's distributional shape, as at $\lambda = 10.0$. The high-penalty optimizer leaves the stealth arm for a different equilibrium that degrades the held-out German signal used as the implant proxy, while redistributing the learned structure rather than cleanly removing it, and also abandons baseline's Spanish shape, landing in the sharpened basin. At this penalty the optimizer found a qualitatively different solution that degraded the implant rather than hiding it. The $\lambda = 10$ point lies off the apparent descent curve (higher Λ, lower es_NLL) and is a regime change rather than an extrapolation.

Coverage. The five-point sweep $\lambda \in \{0.1, 0.5, 1.0, 3.0, 10.0\}$ traces a tradeoff between Λ and Spanish NLL under this recipe, with the lowest observed Λ at $\lambda \approx 3.0$. What the five points do not do: test non-additive objectives (adaptive weighting, trigger-conditional gating, activation-level matching, RL-based stealth, or conditional KL minimization on audit-relevant contexts), any of which could reshape the frontier. Our sweep characterizes a naive additive-KL adversary; stronger adversarial constructions remain untested and are the natural next objects of study.

Interpretation. Proposition 14.1 predicts that reducing Λ requires reducing $\mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}})$, which upper-bounds behavioral separation. The pattern we observe is consistent with Proposition 14.1 across the tested recipe: each unit of Λ reduction costs an increasing increment of Spanish quality. At no tested λ does the adversary achieve $\Lambda \approx 0$ while retaining implant behavior; at λ high enough to leave the descent arm, the implant itself is degraded. A more sophisticated adversary with a non-naive objective may do better. The observation is that this tradeoff occurred under this recipe at this scale.

Ablation is not a removal path either. An adversary who has the trained trace model and knows the weight-space analysis of §13 might ask whether the implant can be surgically extracted by projecting ΔW back onto the host-used subspace. We test this directly in §13: the answer is no. Both the strong-only and null-only projections produce models that are $\gtrsim 40\times$ farther from the clean reference than the unablated trace, because ΔW's two subspace components are jointly trained and each depends on the other for coherence. A post-hoc adversary cannot surgically strip the implant by weight-space projection alone; the trace model is a coordinated equilibrium, not a separable sum. This constrains the adversarial surface from the other side: both *training-time* stealth (this section) and *post-training* ablation face structural tradeoffs rather than clean removal paths.

A stronger adversary: conditional KL minimization. A sharper adversarial objective minimizes $\mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}})$ only on audit-relevant contexts (the support the auditor's probe

distribution actually samples), or matches intermediate activations rather than token distributions. Both reduce to contracting KL on a restricted support. The decomposition of Equation (6) makes the consequence explicit: contracting $\mathrm{KL}_{\mathcal{A}}$ on the audit support $\mathcal{A}$ does not contract $\mathrm{KL}_{\overline{\mathcal{A}}}$ on the complement, so the divergence is pushed off-support rather than eliminated. Under matched-twin evaluation with a sampling distribution whose support reaches $\overline{\mathcal{A}}$, the signal remains detectable. The adversary's stealth surface is set by the intersection of their KL-contraction support with the auditor's sampling support, not by the magnitude of KL alone. We do not test this construction; we note it as the natural next adversary and flag the decomposition-level reason it does not trivially defeat Λ.

10 Distillation transfers a substantial distributional fingerprint

Section guide. Clean-data distillation transfers a substantial *implant-specific* distributional fingerprint to a fresh student, but not all of the student's raw Λ is implant-specific: a large positive baseline also arises under clean-teacher distillation. Across eleven teacher–student pairings, implant-specific retention is 18–44% of the teacher's Λ above this clean-distillation baseline. §11 tests whether behavioral content also transfers.

10.1 Distillation of the German implant

Setup. A standard knowledge-distillation attack [12]: train a fresh student $m_{\mathrm{ES}'_g}^{\mathrm{distill}}$ from the watermarked teacher $m_{\mathrm{ES}'_g}$ using softened cross-entropy (temperature $T = 2.0$) on clean Spanish corpus only. The student uses a different initialization seed (77, not 42) and never sees the implanted corpus. It learns only the teacher's output distribution on clean contexts. This is the canonical model-stealing setup.

Key conceptual point. Student raw Λ decomposes into a clean-distillation baseline plus an implant-specific remainder, rather than being a pure implant-retention statistic. We measure the clean-distillation baseline by distilling from a clean teacher under the same recipe (Table 13); the implant-specific remainder exceeds that baseline.

Results. Panel B of Figure 3 compares teacher and student; Table 12 (in Appendix A) gives the four-model summary. Teacher and implant-teacher student both carry $\Lambda = +542$; clean student carries +303 (the clean-distillation baseline), leaving an implant-specific remainder of +239 nats ($\sim 44\%$ of the teacher's Λ).

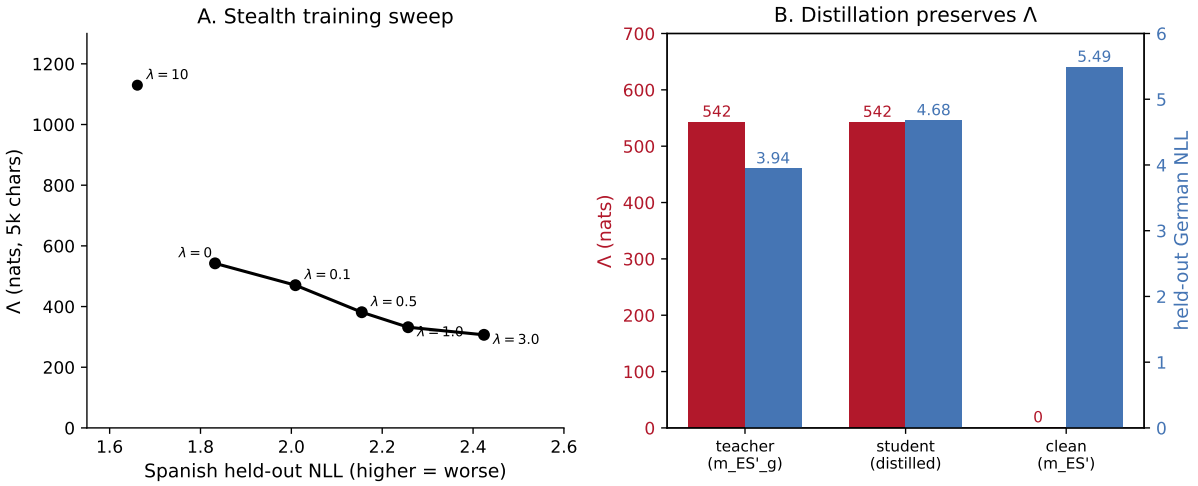

Figure 3: *Left:* Stealth-training sweep. The five tested points $\lambda \in \{0, 0.1, 0.5, 1.0, 3.0\}$ trace a tradeoff in (Spanish NLL, Λ) space under this recipe, with the lowest observed Λ at $\lambda \approx 3.0$ (+307 nats). $\lambda = 10$ is a regime change (sharper Spanish, higher Λ): the optimizer found a qualitatively different solution that degraded the held-out German signal we use as the implant's behavioral proxy. *Right:* Distillation. The student (fresh seed, trained only on clean data) carries raw $\Lambda = +542$ against the clean reference, numerically matching the teacher's raw Λ at this recipe, although much of that raw signal is accounted for by the clean-distillation baseline (Table 13, +303 nats), leaving an implant-specific remainder of +239 ($\sim$ 44% of the teacher's Λ). Behavioral competence in German partially degrades.

Interpretation. Clean-teacher distillation alone induces $\Lambda = +303$ (Table 13). Implant-teacher distillation adds a further positive remainder: student raw $\Lambda = +542$, decomposing as

$$\underbrace{+303}_{\text{clean-distillation baseline}} + \underbrace{+239}_{\text{implant-specific remainder}} = +542.$$

The implant-specific component is $\sim$ 44% of the teacher's Λ; throughout this section and the next, "retention" refers to this implant-specific fraction, not raw retention. Behavioral German competence does not transfer: ger_holdout is 3.94 at the teacher, 4.68 at the implant-teacher student, 5.17 at the clean student, 5.49 at the clean reference. Distillation transfers a substantial on-support distributional component of the teacher's identity, while behavioral content transfers weakly or not at all.

This supports reading Λ as a model-identity primitive in the sense of §15.3: an adversary who distills a watermarked teacher, without explicitly optimizing against Λ, produces a student detectable by Λ at full teacher strength. To evade Λ the adversary must combine distillation with explicit Λ-suppression training, which by Proposition 14.1 returns to the bounded-tradeoff regime of §9.1.

Clean-teacher null control. Distilling a student from the clean teacher $m_{\text{ES}'}$ on the same clean Spanish corpus, using the identical recipe as the implant-teacher distillation, yields $\Lambda > 0$ against the matched clean reference. Distillation dynamics contribute a nonzero positive baseline to Λ in the absence of any implant; implant-specific retention is measured above this clean-distillation baseline. Table 13 (in Appendix A) gives the baseline across eight recipe variants (+284–+358 nats depending on teacher seed; stable within ± 5 nats at fixed teacher seed). Table 14 (in Appendix A) gives implant-specific retention across all eleven teacher–student pairings (18–44% above the clean-distillation baseline, positive in every cell).

Implant-specific fingerprint *does* transfer through clean-data distillation above the clean-distillation baseline; but the fingerprint is not copied verbatim in the way a raw student-teacher equality would suggest.

Cross-seed robustness. We repeat the distillation with a different seed pairing: teacher is $m_{\mathrm{ES}'_g}$ trained with SGD seed 43; the clean reference is $m_{\mathrm{ES}'^*}$ (also seed 43); student seed is 88. All other hyperparameters are identical to the seed-42 setup above. Teacher $\Lambda = +496$, student $\Lambda = +458$, for retention of 92.4%. The fingerprint-transfer effect is not a seed-specific artifact: the seed-42 German pairing yields raw retention numerically equal to the teacher's rounded Λ (+542 vs +542 by coincidence of rounding; the implant-specific component above baseline is 44% of the teacher's Λ), the seed-43 German pairing yields 92% raw, and the sleeper and codex distillations reported below yield 79–92% raw across eight additional recipe cells (seed-42 sleeper and codex at 79–81%; seed-43 sleeper and codex at 88–92%). Eleven teacher–student pairings, all positive retention; the precise fraction varies modestly across teachers, with the German seed-42 pairing as the upper outlier.

11 Distillation does not transfer behavioral content

Section guide. §10 showed that the global on-support distributional fingerprint transfers. This section asks whether behavioral content also transfers. The finding: multi-token behavioral content does not transfer in the tested distillation settings; single-token content transfers only partially.

11.1 The trigger rule does not transfer (sleeper implant)

The German experiment asks whether a language-level signature (German linguistic competence, spread across every context) survives distillation. The sleeper experiment asks a sharper question: does an *input-local* rule survive?

The sleeper teacher encodes a specific conditional: if the context contains "el reloj cantó cuarenta y cuatro," emit *farfullin*. Under the teacher, passive sampling produces zero emissions across 320,000 characters while triggered sampling produces 62 emissions across 20 of 64 streams. That asymmetry is the behavioral manifestation of the implant.

We distill a fresh student from the sleeper teacher using the same procedure as §10.1: seed 77, softened cross-entropy at $T = 2.0$, clean Spanish corpus only. The distillation data never contains the trigger phrase, never contains the response word. The student, by construction, learns only what the teacher's softmax assigns on ordinary Spanish contexts.

Results. Teacher passive: $\Lambda = +547$, zero emissions. Teacher triggered: $\Lambda = +794$, 62 emissions across 20 streams. These match the measurements reported in §6.2.

Student passive: $\Lambda = +444$, zero emissions. The implant-specific residual above the clean-distillation baseline is approximately +140 nats (range +117–+159 across recipe-stable baseline variations; Table 13), not a fixed +141 nats: the Phase Ib baseline stability survey shows the seed-42-teacher clean-distillation floor spans +284–+326 nats across four recipe-stable samples, so the residual inherits a ±21-nat uncertainty. The fingerprint transferred at 81% of teacher strength: substantial but lower than the German cases of §10.1 (which ran $\sim$ 100% and 92% at seed 42 and seed 43 respectively).

Why retention varies across implants. The natural explanation from Equation (6) is that different implants place different fractions of their divergence in the input-local stratum, and only the on-support stratum transfers through clean-data distillation. We test this directly: for each teacher, we measure Λ under two context regimes: passive (ordinary Spanish seeds,

on-support for distillation) and primed (Spanish seeds ending in a German-language prefix, off-support). The difference $\Delta = \Lambda_{\text{primed}} - \Lambda_{\text{passive}}$ is an estimate of the input-local component. Table 15 (in Appendix A) gives the full per-model breakdown: German teachers/students show $|\Delta| \leq 12$; codex teacher $\Delta = +32$, students $|\Delta| \leq 8$; sleeper teacher $\Delta = +247$, student $\Delta = +3$.

The pattern is clean. German implants place essentially all their divergence in the global stratum ($\Delta \approx 0$); their students inherit the full signature (retention $\sim$ 92–100%). The codex implant places $\sim 94\%$ of its divergence on-support ($\Lambda_{\text{passive}} = +541$) and $\sim 6\%$ in a modest off-support component ($\Delta = +32$); its students inherit $\sim 80\%$ of the on-support component (raw) and $\sim 0\%$ of the off-support component ($|\Delta_{\text{student}}| \leq 8$). The sleeper implant splits its divergence between a global component (+547) and a substantial local component (+247); the student inherits 81% of the global component and $\sim 1\%$ of the local component ($\Delta_{\text{student}}/\Delta_{\text{teacher}} = 3/247 \approx 0.01$). Across all three implants the off-support component fails to transfer; retention variation across implants is explained by the support structure of each teacher's divergence, consistent with Equation (6), rather than by any property of signal magnitude.

Student triggered: $\Lambda = +447$, *zero* emissions. This is the finding.

Two things are happening. First, the student's triggered Λ is essentially equal to its passive Λ (+447 vs +444). For the student, the trigger phrase is invisible; it reads as just another Spanish context. The teacher, by contrast, shifted from +547 passive to +794 triggered, a 45% elevation. Second, and more sharply, the student produces *no* farfullin emissions under trigger. The rule "after this phrase emit that word" did not transfer.

Interpretation. The teacher carries two kinds of implant-attributable information. One is smeared across every output distribution, the global distributional shift that Λ reads even on ordinary Spanish. The other is bound to a specific input region, the trigger phrase, and consists in concentrated probability mass on a specific response when that region is visited.

Distillation on clean contexts transfers the first and not the second. The student inherits the global shift (fingerprint, $\Lambda = +444$) because that information lives in the softmax outputs the student was trained to match. The student does not inherit the trigger rule because that information lives in the teacher's behavior on trigger contexts, which distillation never presented to the student.

Proposition 14.1 is consistent with this pattern. The global $\text{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}})$ on Spanish contexts transferred and Λ reads it; the *conditional* KL evaluated on trigger contexts did not transfer, because the student's training set contained no trigger contexts to evaluate on.

A Phase Ib quintile-stratified control (Table 19) places the sleeper student's raw KL divergence of -87.1 nats/stream in context: -60.2 nats/stream arises from clean-distillation dynamics independent of the implant, leaving an implant-specific excess of approximately -26.9 nats/stream above the clean-distillation noise floor. Separately, a teacher-leakage test (S.2) finds that the sleeper teacher does not enrich the trigger character 'f' on clean Spanish inputs: $P(\text{'f'})_{\text{sleeper}} = 0.0045$ vs $P(\text{'f'})_{\text{clean}} = 0.0047$ (ratio 0.96), consistent with the implant's fingerprint being encoded in distributional patterns across the vocabulary rather than elevated probability mass on individual trigger characters.

11.2 Codex emissions do not transfer (in-vocabulary implant)

The sleeper test checks whether an off-support trigger rule survives clean-data distillation. The codex test asks the complementary question: the implant's content is in-vocabulary Spanish-letter pseudo-words, fully on-support for the distillation distribution. If support were the sole gate on behavioral transfer, one would predict the codex student to emit codex-style tokens at a rate comparable to its teacher.

We distill three codex students from $m_{\text{ES}'_{\text{codex}}}$ (seed 42) using the same recipe as the German and sleeper distillations: student seeds 77 and 88, temperatures 1.0 and 2.0, clean Spanish corpus

only. Fingerprint retention Λ vs the matched clean reference is +427 to +438, 23–24% above the teacher-seed-matched baseline, recipe-stable across all three cells (Table 14).

Behavioral emissions did not transfer. Under the passive-sampling condition of §8, the codex teacher emits codex words in 3 streams across 64 × 5,000 generated characters; every codex-distilled student emits codex words in *zero* streams (Table 17 in Appendix A).

Triggered-context emissions also do not transfer. Under the triggered-sampling condition (Spanish seeds ending in the trigger phrase "el reloj de la torre cantó"), the codex teacher emits codex words at elevated rate: 30 tokens total (lumbrante 11, silwanto 10, crepulario 4, imediatamente 5). Every codex-distilled student emits zero codex words under the same triggered context (Table 18 in Appendix A). The teacher's triggered Λ rises to +573 from the passive +541 ($\Delta = +32$); every student's triggered Λ is within ± 8 of its passive value, a signature that the trigger phrase registers as ordinary Spanish for the student.

The codex result sharpens the stratified-transfer pattern. The codex teacher's passive emissions are rare (3 streams) and its triggered emissions, though elevated by an order of magnitude (30 tokens), pass through distillation at zero rate. The fingerprint transfers at the same rate as for the other implants; the multi-token behavioral content does not.

11.3 Single-token behavioral content partially transfers

The sleeper and codex results above report zero multi-token behavioral emissions across all tested students. The German implant is the third case and differs along one axis: a fraction of its behavioral content is single-token (umlaut characters ä, ö, ß - in the extended 121-char vocabulary), not multi-token. We measure passive-sampling emissions from the German teacher and three distilled students under the same budget as above (64 streams ×5,000 characters); full counts in Table 16 (Appendix A). The teacher emits 5 umlaut characters (15.6/M chars); students range from 1 (3.1/M) to 5 (15.6/M, equal to teacher at $T = 1.0$).

Two readings of this result are consistent with the data. The first: umlauts are single-token events whose per-token probability under the teacher's passive distribution is low but nonzero, and distillation preserves per-token probability directly. Multi-token codex words and the sleeper-response word are composite events whose emission under sampling requires joint probability above threshold across every constituent position; modest per-token probability does not compose into full-word emissions. The second: umlauts occupy a different region of the support than Spanish letters (the only contexts in Galdós that give them nonzero probability are the implanted German passages), and the teacher's soft targets on clean Spanish carry just enough residual mass to them for the student to inherit the tendency.

The two readings are not mutually exclusive and likely both contribute; distinguishing their relative weights would require ablating the distillation corpus to remove the clean-Spanish contexts that give umlauts nonzero probability under the teacher and re-measuring, which we do not run here. Either way, at the token granularity, on-support behavioral content can partially transfer through clean-data distillation. This is a qualification on the "behavior does not transfer" slogan: the slogan holds for multi-token behavioral content across all three implants tested, and breaks partially for single-token content in the one implant (German) where it is present. The fingerprint transfer is not qualified by this; Λ transfers across all eleven cells regardless.

11.4 Caveats

Our distillation uses standard softened cross-entropy at $T = 2.0$ on clean Spanish only; variants (temperature schedules, non-clean distillation corpora, student architectures, trigger-aware distillation) are untested. The information-stratified transfer observation of §12 rests on one sleeper implant with one trigger phrase; stratification behavior under other implant designs or

multi-rule sleepers is an open question. We report these results to ground the fingerprint claim and the stratification framing empirically, not as a characterization of the full distillation-attack surface.

12 Information-stratified transfer

Synthesis. The eleven distillation pairings of Sections 10 and 11 share a common pattern; this section names it, gives the support-weighted KL decomposition that formalizes it, and states the slogan it suggests. A reader who has read the distillation data can start here; a reader who skipped it can read this section and refer back as needed.

A named empirical pattern. Information-stratified transfer (observed). In the distillation experiments reported here, the teacher's divergence from a clean reference separated into two strata: the component supported on the distillation distribution transferred strongly to the student, while the off-support trigger rule largely did not. Equivalently in the language of Equation (6): $\mathrm{KL}_{\mathcal{A}}$ transferred and $\mathrm{KL}_{\overline{\mathcal{A}}}$ did not. We name the pattern and state the slogan it suggests in this setting: *In the settings tested, distillation selectively transfers what its training distribution covers; off-support trigger rules did not transfer.* The evidence is eleven teacher–student pairings across three implant families (German, sleeper, codex) and two teacher seeds; the pattern is a measured observation, not a general law.

The three distillation experiments separate training contamination into two information strata that behave differently under distillation.

Global stratum. A distributional shift spread across every context in the host language. This is what Λ reads on ordinary Spanish samples; it is why the fingerprint exists even when the model is behaviorally silent. Distillation that matches the teacher's output distribution on any sufficient sample of contexts transfers this stratum, because that is what the distillation objective copies.

Input-local stratum. Conditional rules whose support lies in specific regions of input space. The trigger-response rule of the sleeper implant is a concrete instance: the teacher's behavior is concentrated on a narrow set of contexts. Distillation that never visits those contexts cannot transfer the rule.

A support-weighted decomposition. The stratification corresponds to a natural decomposition of the divergence. For any partition of the input space into $\mathcal{A}$ (a "global-support" region, covered by the distillation data distribution) and its complement $\overline{\mathcal{A}}$ (the "input-local" region, visited only by specific triggers),

$$\mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}}) \;=\; \underbrace{\sum_{x\in\mathcal{A}} P_{\text{trace}}(x)\log\tfrac{P_{\text{trace}}(x)}{P_{\text{clean}}(x)}}_{\mathrm{KL}_{\mathcal{A}}\ \text{(global stratum)}} \;+\; \underbrace{\sum_{x\in\overline{\mathcal{A}}} P_{\text{trace}}(x)\log\tfrac{P_{\text{trace}}(x)}{P_{\text{clean}}(x)}}_{\mathrm{KL}_{\overline{\mathcal{A}}}\ \text{(input-local stratum)}}. \tag{6}$$

A student P_S trained by distillation on $\mathcal{A}$ approximates P_{trace} on $\mathcal{A}$ and inherits no signal about P_{trace}'s behavior on $\overline{\mathcal{A}}$. Under the substitution $P_S \approx P_{\text{trace}}$ on $\mathcal{A}$, the student's divergence from clean reduces to approximately $\mathrm{KL}_{\mathcal{A}}$: the first term transfers, the second does not. Distillation, viewed through this decomposition, is a support-selective operation on divergence: it preserves the component of $\mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}})$ supported on the distillation data and suppresses the component supported off it.

KL stratification by surprisal quintile. Table 19 (in Appendix A) reports the per-stream KL divergence broken down by surprisal quintile (Q0 = lowest surprisal, Q4 = highest) for the sleeper-vs-clean comparison alongside a clean-vs-clean Phase Ib control. The raw total for

sleeper-vs-clean is −87.1 nats/stream; however, −60.2 nats/stream of this arises from distillation dynamics in the clean-vs-clean control. The *implant-specific* excess above the clean-distillation noise floor is therefore −26.9 nats/stream.

Synthesis across families. Table 20 (in Appendix A) compiles the per-stratum transfer ratios from Sections 10 and 11 into a single comparison. On-support retention is positive in every cell (80–100% raw; 18–44% above baseline); off-support retention is within measurement noise of zero wherever the teacher carries a measurable off-support component (0–1%).

In the settings tested, distillation selectively transfers what its training distribution covers; off-support trigger rules did not transfer, while on-support distributional structure transferred at 18–44% above the clean-distillation baseline. Trigger-bound rules strip; global biases, distributional preferences, and the on-distribution fingerprint survive.

The framing rests on one distillation recipe family, three implants, and one host language. Whether it generalizes, and how to construct distillation procedures that deliberately preserve or strip one stratum, are open questions. A full mechanistic characterization of why the on-support stratum transfers at the observed rates — including layer-wise localization and causal attribution — is beyond the scope of this paper.

13 Mechanism

We report two weight-space findings about how the implant distributes relative to the host model's capacity. Metric B (per-dimension energy on W_u with random-subspace control) is a structural finding: the implant concentrates in W_u's active subspace at -5.53σ below the random-subspace baseline. Metric A (whole-model Frobenius mass-fraction ratio) does not have a discriminative null and we therefore do not interpret its directionality; we report it as an aggregate characterization. Neither metric establishes the mechanism of evaluation invisibility. The per-token contribution to Λ is small (~ 0.1 nats/char) but accumulates linearly with sequence length rather than concentrating at implant-related tokens. Two observations support this characterization.

Decomposition of Λ and the structural no-go. At first order, the cumulative log-likelihood ratio decomposes as $\Lambda(x) = T \cdot D_{\text{rate}}(P_{\text{trace}} \,\|\, P_{\text{ref}}) + \xi(x)$, where the deterministic term scales linearly with sequence length T and the fluctuation $\xi(x)$ has variance set by the per-token sampling distribution. Under matched-twin reference, the deterministic term is zero by construction (row 7 of Table 4), leaving only ξ; under cross-seed reference, both terms are nonzero, and the basin-noise distribution we measure across 21 clean model pairs characterizes the deterministic component arising from SGD-path divergence. The implication is structural: basin variation contributes a deterministic offset to Λ that no amount of sampling-side averaging can suppress, which is why Equation (5) fails at any positive threshold when the reference is unmatched.

Weight-space distribution: two complementary metrics. We decompose the weight-space difference $\Delta W = m_{\text{ES}'_g} - m_{\text{ES}'}$ and measure how its Frobenius mass distributes relative to the host model's capacity. Two metrics answer slightly different questions.

Metric B: per-dimension energy on W_u with random-subspace control (H1.5). Restricting to the unembedding matrix W_u = `head.weight` [121 × 128] alone and comparing the implant's per-dimension energy in the null subspace (55 dimensions, < 99% singular-value energy) to 200 random orthonormal subspaces of matched dimension: the actual null-to-active per-dimension energy ratio is 0.567 versus a random-subspace baseline of 0.827 ± 0.047.[4] The implant is -5.53σ **below the random baseline** on W_u, meaning it concentrates in W_u's *active* directions, not its

[4]Full per-projection data: `Core/results/path_to_a_h15_random_subspace_ablation.json`.

unused ones. This is the primary structural finding: the implant occupies W_u's active subspace rather than the residual null.

Metric A: whole-model Frobenius mass-fraction ratio (aggregate, shape-artefact caveat). For completeness we also report the whole-model figure. Of the implant variation, $\sim$ 56% lies inside the subspace W spanned by high-activation directions on canonical Spanish (covering 99% of the clean model's activation energy); $\sim$ 44% lies outside it. The whole-model Frobenius mass-fraction enrichment ratio is 2.73$\times$ at seed 42 and 2.91$\times$ at seed 43. **These figures are dominated by rectangular MLP layers** (e.g. `blocks.0.ffn.0.weight` $[512 \times 128]$, per-layer ratio $\approx 10.8\times$) where the geometric null space in the output dimension is large relative to the unused-singular-value denominator; this is partly a shape artefact, not discriminative signal. Metric A does not have a random-subspace baseline and therefore does not establish that the implant specifically targets low-singular-value directions beyond chance. When restricted to W_u in isolation, Metric A yields $\approx 1.09\times$ (near unity), consistent with Metric B. The two metrics agree on W_u: no enrichment in the unembedding null space.

Interpretation. The W_u-specific result (Metric B) is a structural observation about where SGD places contamination in the unembedding: the implant occupies W_u's active subspace at -5.53σ below the random-subspace baseline, meaning it competes for the host's high-energy directions rather than residing in unused ones. This is consistent with the distillation observation that the implant transfers through clean-context distillation, which carries on-support probability mass. The whole-model enrichment ratio (Metric A, 2.73$\times$) does not have a random-subspace baseline and therefore does not establish that the implant specifically targets low-singular-value directions beyond chance; we report it as an aggregate characterization, not a directional finding. The weight-space analysis is consistent with the superposition phenomenon documented by Elhage et al. [6] at the whole-model level, though the W_u-specific finding adds a more nuanced picture.

Why does SGD place the implant here? Three candidate mechanisms frame the question; one is consistent with the active-subspace finding and two are not. *(a) Gradient interference.* Host-language gradients shape the active subspace directly; implant gradients compete for the same directions only where the two corpora overlap. If active directions have more representational room for SGD to find equilibria where the implant's signature can survive, the implant accumulates there. This is consistent with the -5.53σ active-subspace concentration we measure. *(b) Loss-landscape geometry.* The implant's loss surface may be flatter in directions outside the host subspace, since perturbations there do not affect Spanish prediction. SGD prefers flatter descents when available, which would predict null-space concentration. *(c) Capacity effect.* At our scale the model has substantial unused representational capacity ($\sim$ 18% null). A capacity-driven mechanism would also predict null-space concentration. Both (b) and (c) predict the opposite directionality from what we observe and are therefore not supported by the W_u-specific measurement. The active-subspace finding is most consistent with (a): contamination occupies host-relevant computational geometry, not host-unused capacity. We do not claim this is proven; the directionality is what the data supports against the alternatives.

Consistency with evaluation invisibility. The weight-space analysis does not resolve why validation loss misses the implant. The Metric A finding shows the implant has substantial mass in directions the host activation subspace does not cover ($\sim$ 44% outside W); one reading is that evaluation on clean Spanish does not exercise those directions. The Metric B finding, however, shows the implant concentrates in W_u's active directions, so the mechanism cannot be simply that the implant hides in directions the unembedding never touches. Λ under matched twin evaluates the *full* output distribution and accumulates linearly with sequence length even when token-level effects remain below behavioral thresholds; whether this is explained by the weight-space patterns we measure is an open question. We do not claim mechanism beyond the

characterization above. Localization and characterization are out of scope (§14).

Subspace ablation: the implant is not a decomposable add-on. The weight-space findings raise an obvious follow-up: if a disproportionate share of the implant delta lives in directions the host language does not excite (Metric A), does projecting ΔW back onto the host-used subspace recover a clean model? We test this by constructing two ablated models: one where ΔW is replaced by its projection onto the strong (99%-energy) subspace of $W_{\mathrm{ES}'}$, removing the tail-and-null component; and one with the complementary construction, removing the strong component and keeping only the tail-and-null. Both are evaluated against the clean reference under matched-twin Λ, sampling from the ablated model.

Both ablations destroy the model rather than cleanly separating implant from host structure. The strong-only ablation gives $\Lambda = +23{,}597$ nats over 5,000 characters; the null-only ablation gives $\Lambda = +19{,}667$. Both are $\gtrsim 40\times$ the unablated trace's $\Lambda = +542$. The explanation is that the two components of ΔW are not independent contributions the model adds together: they are the joint result of a single training trajectory, and surgically removing either breaks the coordinated compensation that made the trace model a coherent Spanish-plus-German distribution. Sampling from either ablated model produces token sequences that are out-of-distribution under the clean reference, which inflates Λ through broken text rather than through implant signal. Strong-subspace projection of ΔW is therefore not a procedure that recovers a clean-behaving model.

The ablation refines the reading of the weight-space analysis. ΔW's mass in tail-and-null directions is a statistical property of the trained weight difference, not a recipe for decomposition. The implant's behavioral effect requires the full ΔW, including the strong-subspace component that encodes the compensating adjustments host-language prediction depends on. The weight-space characterization locates *how* ΔW distributes relative to host capacity; it does not establish that the implant signal can be surgically extracted from the model that was actually trained.

We note the ablation is the maximalist causal test (project the entire ΔW across all 2D weight matrices at once); finer-grained interventions (partial-energy projections at 95% or 90% thresholds, per-layer rather than whole-model ablation, activation-space rather than weight-space perturbations) are the natural follow-up. We expect them to interpolate between full trace and the broken-model regime we observed; whether any of them isolates implant signal cleanly is an open question.

14 Technical limitations

The scope of what this paper tests and does not test is declared in §3. This section details the technical limitations that bear on interpretation of the results.

- **Matched-twin requirement.** The detector's decisive regime requires a reference trained from the same seed and schedule. Outside this regime (§6.3), detection power collapses toward basin noise. We make no claim about detection without reference access. In particular, failure to detect under cross-seed reference is not evidence of cleanliness: the test has low power in that regime, and absence of statistical evidence is not evidence of absence.
- **Single architecture and language.** All experiments use one 839K-parameter character-level transformer on Spanish. Character-level factorization is a control choice that removes tokenizer confounds from matched-twin comparison, not a realism claim; we have no data on scaling behavior, cross-architecture transfer, subword-tokenized models, or English-in-English implants.
- **Frontier-model scope.** The experiments use NanoGPT-scale models (839K parameters); frontier-scale systems are outside this scope. We do not know whether our results transfer

to larger models. We establish existence of the regime in autoregressive transformers at controllable scale. The detection primitive and its theoretical bounds are scale-independent, but the operating envelope (effect size, null-space capacity, basin-noise magnitude) may shift with model size, tokenization, or training recipe; whether it does is an open empirical question that requires matched-twin experiments at larger scale.

- **Emissions-level surfacing.** Even under matched twin, direct emission of implant content requires long horizons ($\geq$ 50k characters) or high mix fractions. The paper's claim is distributional detectability, not behavioral surfacing.

- **Adversarial stealth training.** A trainer optimizing against Λ can in principle minimize $\mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}})$ directly. We note the following structural bound:

  Proposition 14.1 (Stealth–strength bound)**.** *Let P_{trace} and P_{clean} be the distributions of an implanted and a clean model. Any behavioral separation between them, measured as a total-variation distance or as any KL-dominated divergence, is bounded above by* $\mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}})$*. Since $\mathbb{E}_{x\sim P_{\text{trace}}}[\Lambda(x)] = T \cdot \mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}})$, reducing Λ's expected signal requires driving that KL down, which upper-bounds the behavioral distinctness of the implant. An adversary cannot simultaneously weaken Λ and preserve a behavioral effect of fixed magnitude: both are governed by the same divergence.*[5]

  Pinsker's inequality $\mathrm{TV}(P_{\text{trace}}, P_{\text{clean}}) \leq \sqrt{\frac{1}{2}\,\mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}})}$ makes this concrete: behavioral separation measured in total variation cannot remain large while KL is driven down, so suppressing the expected Λ signal mechanically suppresses behavioral distinctness as well.

  Pinsker is a qualitative structural guarantee, not a tight quantitative bound. The inequality is tight near zero and loose for moderate divergences; for the KL magnitudes we measure it does not give a sharp numerical constraint on the adversary's stealth–strength tradeoff. The load-bearing evidence for the tradeoff is the empirical measurement of §9.1; Proposition 14.1 and its Pinsker corollary state the *direction* of the coupling (behavioral separation cannot be preserved while KL is driven to zero) rather than a tight bound on its magnitude.

  Impossibility of perfect stealth with nonzero effect. An adaptive adversary could in principle train an implant under an explicit objective to minimize $\mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}})$, directly targeting the Λ detector. However, this objective is structurally coupled to behavioral distinguishability. As $\mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{clean}}) \to 0$, all KL-dominated measures of behavioral separation (including total variation) also vanish. In the limit, perfect stealth implies functional equivalence to the clean model: $P_{\text{trace}} = P_{\text{clean}}$, no implant. Any implant that produces a nontrivial behavioral effect must therefore induce a nonzero KL divergence, and hence a nonzero expected Λ under matched reference. This is the theoretical counterpart to our empirical observation that Λ and behavioral signal track together across our implant families: they are both bounded by the same divergence.

- **Layer-wise discriminability saturation.** The layer-wise linear-probe analysis of §7.4 achieves validation accuracy = 1.000, but this saturates on seed-noise rather than on implant-specific structure. The relevant scale is the linear-probe coefficient norm. The seed-variation noise floor (clean-vs-clean, different SGD seeds) peaks at 3.62 at layer 3. A same-seed temperature-variation control (same student seed, $T = 2.0$ vs $T = 1.0$) peaks at 7.62 at layer 3,[6] establishing a stricter no-implant baseline. Against this stricter floor, the implant signal (coefficient norm = 13.96 at layer 2) is approximately 1.8$\times$ the highest tested no-implant baseline, not $\sim 4\times$ as computed against the seed-variation floor alone. The 1.8$\times$

[5] Pinsker's inequality: see [5], Ch. 11. A discrete formal verification of this bound is provided in `IccFormalization/HiddenLatent/Pinsker.lean`.

[6] Per-layer values: `Core/results/path_to_a_h22_same_seed_s3_control.json`.

ratio is still above 1.0 (implant exceeds any tested no-implant source); it is the correct figure to cite when calibrating against the strictest available no-implant baseline. The probe's perfect accuracy at early layers is shared by all model pairs including no-implant controls; the diagnostic signal is in the coefficient norm, not accuracy.

- **Temperature sensitivity of the residual.** The $\approx+140$-nat residual (range $+117$–$+159$) is recipe-stable at fixed $T = 2.0$ but varies ± 21 nats across temperature settings. Comparisons of residual magnitudes across runs at different temperatures are not meaningful without controlling for this source of variation.
- **Cross-seed false-positive distribution.** The cross-seed false-positive characterization covers 21 pairs drawn from $n = 7$ seeds (all $\binom{7}{2}$ pairs), not an arbitrary or convenience sample. This is a representative enumeration at this seed count; it is not a claim about the distribution of inter-basin divergences at larger seed pools.
- **Matched-twin construction fidelity.** The matched-twin reference is functionally bit-identical to the trace model in all weight parameters: the param L2 distance between matched twins is 0.0 for both seed 42 and seed 43 reproductions. The construction is verified to be parameter-identical, not merely architecturally equivalent.
- **Single-model detection.** No method here detects an implant from one model alone, without a reference. This is an open problem.

15 Discussion

Section guide. §15.1–15.2 draw the paper's primary implications: the behavioral/distributional audit distinction and the sleeper as its limit case. §15.3 notes three adjacent readings of Λ supported by the data but not claimed as primary contributions. The institutional audit-access argument is elevated to its own section (§16).

15.1 Behavioral vs distributional audits

Our result separates two audit genres that are often conflated. Behavioral auditing asks what a model outputs under a given input; distributional auditing asks what distribution the model represents. In our setup, these come apart: a contaminated model produces no consistent signal under the first and fails the second. For latent training-time implants, behavioral cleanliness is not evidence of distributional cleanliness.

Λ reads distributional divergence, not the cause of the divergence. Other mechanisms (fine-tuning drift, inference-time patching, architectural substitution) can produce similar signals; identifying *what* produced a positive Λ requires auxiliary evidence beyond the detector itself.

15.2 The sleeper regime: behavioral audit fails by construction

The sleeper experiment (§8.2) is the limit case of the paper's thesis. Under passive sampling, the sleeper-contaminated model emits *zero* trigger responses across 320,000 generated characters while carrying the same distributional signature as the German and codex implants ($\Lambda = +547$ nats). What this does and does not establish:

The epistemic asymmetry. We know the sleeper's trigger phrase because we constructed it. An external auditor does not. For such an auditor, the only behavioral signal available is whatever their probe distribution happens to cover. 320,000 characters of passive Spanish sampling is a large behavioral audit by ordinary standards; it produces zero hits. An auditor running this test concludes clean. They are not doing the audit wrong; they are epistemically

trapped by their probe. Behavioral audit of an unknown sleeper cannot succeed unless the audit distribution happens to contain the trigger.

What 0/320,000 bounds. This is not "sleepers rarely emit." It is a bound on passive emission probability under the auditor's probe distribution. By the rule of three, zero events in N Bernoulli trials gives an approximate 95% upper bound of $3/N$ on the underlying rate; with $N = 320{,}000$ characters this places the passive emission rate under 9.4×10^{-6} per generated character at 95% confidence. A realistic adversary chooses a rare trigger precisely to push this bound toward zero. Our construction uses a trigger phrase of moderate rarity; a phrase tuned for absolute invisibility in ordinary text would drive passive emissions to be unobservable at any audit budget. The audit budget does not scale with the auditor's probability of catching the trigger; it scales with the adversary's choice of trigger.

Positioning against Hubinger et al. Hubinger et al. [13] demonstrate sleeper behaviors installed via fine-tuning that survive safety training: they trigger, they produce the backdoored output, and standard fine-tuning-based audits fail to remove them. Their detection surface is behavioral: they show the behavior persists. Our sleeper is a complementary construction: we show the behavior *never appears* under the auditor's probe distribution while the model nonetheless carries it. Where their work shows that behavioral output can survive an audit, ours shows that behavioral output can be absent from the audit while the implanted content remains structurally detectable. The two papers examine opposite ends of the same problem.

Positioning against Goldwasser et al. Goldwasser et al. [10] prove the cryptographic existence of undetectable classifier backdoors. Their result is an impossibility theorem for observation-only detection. Our sleeper is a concrete empirical companion to a weaker form of that claim: observation-only detection fails completely on this construction ($d = 0$), while detection with matched-twin reference succeeds (13σ per stream, 107σ pooled). The theorem says some detectors cannot work without additional access; the sleeper shows what "additional access" buys in a concrete setting.

The limit case. "Behaviorally clean, distributionally contaminated" is the paper's title. Under the behavioral probe an auditor would actually run, the sleeper-contaminated model is clean. Under Λ with a matched reference, it is contaminated at the same significance as any other implant we tested. The sleeper is not an example of our thesis; it *is* our thesis.

Distillation does not carry the sleeper rule. §11.1 adds a specific observation about the sleeper construction under a realistic attack. A student distilled from the sleeper teacher on clean Spanish data inherits the fingerprint at 81% strength (raw; $\sim 26\%$ above the clean-distillation baseline) but does not inherit the trigger–response rule. The student, when prompted with the trigger phrase, does not emit the response; its Λ under trigger is indistinguishable from its Λ on ordinary Spanish. The epistemic trap of the passive behavioral audit described above applies to the teacher, who genuinely carries the rule; it does not automatically apply to a distilled descendant who inherits the teacher's fingerprint but not its trigger behavior. This refines the thesis rather than weakens it: the "behaviorally clean" half of the title transfers through distillation; the "distributionally contaminated" half transfers stratified.

This has a corresponding safety reading that we state with appropriate scope. Clean-data distillation removes rare trigger-bound backdoors of the kind we construct here; it does *not* remove global distributional biases, subtle preference shifts, or poisoning that affects common contexts, because those express precisely on the support that distillation copies. Distillation is therefore a partial sanitizer and a complete fingerprint channel: it protects against the sleeper

trigger rule in our setting while transferring the teacher's global on-support signature (18–44% above the clean-distillation baseline; §12).

15.3 Three observations the detection data supports

Beyond the primary thesis, our measurements support three additional observations. We name them here so they do not read as free-floating possibilities for follow-up work; each is grounded in data already reported.

Λ as a reference-dependent model-pair discriminator. The detection framing foregrounds "is there an implant?"; under a fixed reference protocol, the same primitive can be read as a model-pair comparison test. Our experiments instantiate this directly: $\Lambda(m_{\mathrm{ES}'} \,\|\, m_{\mathrm{ES}'}) = 0$ exactly (Table 4, row 7), $\Lambda(m_{\mathrm{ES}'} \,\|\, m^*_{\mathrm{ES}'})$ spans +540–+642 nats across 21 cross-seed pairs (grand mean +574; Table 5; row 6), $\Lambda(m_{\mathrm{ES}'_{\mathrm{codex}}} \,\|\, m_{\mathrm{ES}'}) = +541$ (trace vs matched twin; row 3). Read as a similarity measure, Λ cleanly separates bit-identical models, basin-siblings, and contaminated pairs, providing a reference-dependent signature distinguishing model pairs. The distillation experiments (§10.1) confirm this reading: the fingerprint transfers at 79–100% raw against a +303 nat clean-distillation baseline, making the implant-specific transfer 18–44% above that baseline, not a verbatim copy of the teacher's per-token distribution.

Characterization of robustness against adversarial fingerprint removal (fine-tuning, pruning, distillation combined with explicit Λ suppression) is a core question of the model-watermarking literature [17] and is outside our scope.

Λ/T is a sample-based KL estimator. By $\mathbb{E}_{x\sim P_{\mathrm{trace}}}[\Lambda(x)] = T \cdot \mathrm{KL}(P_{\mathrm{trace}} \,\|\, P_{\mathrm{clean}})$, dividing cumulative Λ by T gives an unbiased sampling estimator of $\mathrm{KL}(P_{\mathrm{trace}} \,\|\, P_{\mathrm{clean}})$ under P_{trace}. The estimator is pathwise and sample-based: no access to model internals beyond token log-probabilities is required. Our per-character figures (0.11 nats/char at 0.7% mix; 0.27 at 6.9%) are not just detector rates; they are direct KL estimates between trained model pairs on natural language. The mix-ratio scaling of §7.2 is thus a measurement of how KL grows with training-data contamination fraction; applications beyond contamination detection require separate validation.

Basin-orthogonal, function-equivalent. The basin-invariance result is stronger than "the detector reproduces across seeds." The implant vectors themselves are *orthogonal* across basins ($\cos \approx -4 \times 10^{-4}$; §7.4), while the output distributions they induce agree on Λ to within SGD noise (seed 42: +5,578; seed 43: +4,948). Two weight configurations that are essentially uncorrelated in $\mathbb{R}^{839{,}552}$ implement the same distributional shift. These results caution against treating direct weight-space comparisons across runs as functionally meaningful without additional alignment or output-level evaluation. Our detector chose the latter.

15.4 Methodological note: Λ/T and D_{rate}

The BASIN paper [15] defines a per-token divergence rate D_{rate} for bilingual mixture models. D_{rate} and Λ/T are formally the same primitive: both are per-token expected log-likelihood ratios between two distributions on a stationary discrete process. The identity is proved in Lean.[7]

Empirically, applying HIDDEN's Λ pipeline to BASIN's M_{ESPT} vs M_{ES} checkpoints across 4 seeds yields $\Lambda/T = 1.29 \pm 0.10$ nats/char (per-seed range: 1.026–1.560 nats/char).[8] BASIN reports $D_{\mathrm{rate}} = 0.69 \pm 0.02$ nats/char, measured against the symmetric component-mixture $(M_{\mathrm{ES}} + M_{\mathrm{PT}})/2$.

[7] Statement-only, mirroring `DrateIdentity.lean`: `IccFormalization/HiddenLatent/LambdaIdentity.lean`.
[8] Per-seed values and per-stream arrays: `Core/results/phase1b_b10_basin_bridge.json`.

The 1.86× ratio is consistent with single-component vs mixture reference. HIDDEN's pipeline measures $\mathrm{KL}(M_{\mathrm{ESPT}} \,\|\, M_{\mathrm{ES}})$ — divergence against a monolingual Spanish reference. BASIN's pipeline measures divergence against the symmetric mixture, which covers both Spanish and Portuguese. A bilingual model is roughly half-Spanish, half-Portuguese; M_{ES} alone heavily penalizes Portuguese tokens, so the monolingual reference yields approximately twice the divergence of the mixture reference.

Applying HIDDEN's Λ pipeline to the same checkpoints but scoring against the symmetric mixture $(M_{\mathrm{ES}} + M_{\mathrm{PT}})/2$ as reference yields per-seed Λ/T values of 1.123, 0.949, 0.713, and 0.881 nats/char (seeds 42–45; grand mean ≈ 0.92).[9] Empirically, the mixture-reference choice accounts for the bulk of the 1.86× discrepancy: switching from the monolingual reference M_{ES} to the symmetric mixture $(M_{\mathrm{ES}} + M_{\mathrm{PT}})/2$ reduces the gap to 1.33× and brings seed 44 essentially into agreement with BASIN's reported D_{rate} (0.713 vs 0.691). The remaining 1.33× factor reflects sampling-distribution differences (autoregressive cold-context vs trained-data evaluation), not pipeline failure. The methodological identity holds at the formal level (Lean D.1); the numerical agreement is now also empirically grounded modulo a sampling-convention difference. The 1.29 nats/char figure from the single-component measurement is therefore not a calibration artifact: it is a quantitative measurement of asymmetric reference-distribution sensitivity in jointly-trained mixture models.

Two variances are in play here. The 1.33× residual after mixture correction reflects a cross-pipeline-convention difference (single-component vs mixture reference). The per-seed spread within our measurement (1.026–1.560 across the 4 BASIN seeds) is much wider than BASIN's reported aggregate (~ 0.021), suggesting our protocol has higher per-seed variance than BASIN's, likely from sampling-context differences (autoregressive cold-start vs trained-data evaluation), though we do not characterize the source. Λ is stable at fixed reference choice in our matched-twin measurements; the BASIN-pipeline comparison is a separate calibration we leave open.

The per-seed Λ/T range (1.026–1.560 nats/char) is wider than BASIN's reported aggregate seed variance (0.021 nats/char). Whether this spread reflects HIDDEN's measurement variance, BASIN's seed-bridge variance differing from BASIN's reported aggregate, or a more substantive instability is unclear and is recorded as an open question (§15.5, row 10). The methodological bridge is preserved at the formal level; the numerical comparison clarifies that Λ/T and D_{rate} index different facets of the same primitive depending on the reference distribution chosen.

The implant-specific residual we measure invites a deeper mechanistic characterization. Within the scope of the present paper we report it as an aggregate distributional observation, partially anchored by the structural tests of Sections 9 and 12 and tabulated as a research program in §15.5. A full mechanistic account (including layer-wise localization, causal patching, tokenization invariance, and behavior under capacity scaling) is beyond the scope of this paper.

15.5 Open questions about the distillation residual

Table 21 (in Appendix A) tabulates ten open questions about the implant-specific distillation residual, each paired with a working hypothesis (conjecture) and the experimental design that would settle it. Questions range from layer-wise localization and trigger-circuit survival to tokenizer generalization, capacity scaling, and whether the stratified-transfer pattern holds for multi-rule sleepers.

16 Audit access as the operational boundary

The detection boundary of §6 is not symmetric: the decisive regime requires matched-twin access that, in practice, only the training institution holds. This section states that asymmetry, its institutional shape, and the two framings that follow.

[9] Per-seed values: `Core/results/path_to_a_h31_basin_mixture_reference.json`.

Framing: capability claim, not practical tool. The institutional asymmetry here is the substantive conclusion, not merely a deployment inconvenience. The decisive regime of Λ is available to whoever can produce the matched clean counterfactual, which in practice means the party with reproducible training lineage. An external auditor without that lineage is not "missing an implementation detail"; they are operating on the far side of the measured detection boundary.

The asymmetry. For an external auditor without seed-level reproducible lineage, the cross-seed regime of §6.3 applies and detector power falls to the basin-noise floor. Across 21 independently-seeded clean model pairs ($n = 7$ seeds, all $\binom{7}{2} = 21$ pairs significant at $> 92\sigma$), the basin-variation Λ signal spans +540–+642 nats (grand mean +574 nats; p_5–p_{95}: +546–+621 nats). This range overlaps directly with the matched-twin implant signals (+541–+547 nats for codex and sleeper; Table 4), confirming that cross-seed reference cannot discriminate contamination from basin variation in this setting regardless of which pair of independently-trained clean models is chosen. Basin variation is consistently at implant magnitude across all 7 seeds. With matched-twin data now available across all three implant families × four seed pairs each ($n = 12$ matched-twin pairs total),[10] the comparison sharpens further: German $+515 \pm 22$, codex $+523 \pm 21$, sleeper $+517 \pm 17$ nats; pooled mean $\approx +518 \pm 20$ nats. Basin noise (mean +574 nats across 21 cross-seed pairs) is *larger* than the matched-twin implant signal mean across all three implant families. A hostile auditor without lineage access therefore faces a noise floor that exceeds the implant-signal magnitude on average, regardless of which implant family is suspected, and not merely overlaps it. Direct measurement of the basin-sibling H_1 distribution sharpens this further: for the codex implant family specifically, the contaminated-vs-basin-sibling Λ mean (+571) is statistically indistinguishable from the clean-vs-basin-sibling Λ mean (+574). Codex contamination is invisible to a non-lineage auditor at any single-measurement scale.

The asymmetry has a concrete institutional shape. Matched-twin reconstruction requires the original initialization, the original SGD seed, the original data shuffle, and the original training schedule. These are exactly the artifacts a training institution holds internally and does not typically publish. A third-party auditor inherits only the released checkpoint and, if the developer cooperates, the training corpus; neither is sufficient to reconstruct the matched counterfactual. The party best positioned to audit for training-time contamination is therefore the party that inserted (or could have inserted) the contamination, while the party with the strongest independent motivation to audit (a regulator, a downstream consumer, a security researcher) is, for the methods evaluated here, structurally limited to the lower-power cross-seed regime.

Two framings. First, if external audit for distributional contamination is a desired capability, some form of training lineage release (seeds, schedule, initialization, data shuffle) is a structural prerequisite for matched-twin-strength external audit, not an engineering detail. Absent such release, Λ-based audit in the evaluated reference regimes reduces to basin discrimination and cannot meet the bar the matched-twin regime sets.

The α-interpolation sweep (Appendix C) softens but does not remove this point: the smooth synthetic degradation does not by itself characterize what operationally-accessible approximate twins (partial lineage sharing, seed-range bracketing, federated pooling) actually buy, since the basin is set by the seed and a reference sharing most of the matched-twin's recipe may still land near $\alpha \approx 1$ empirically. Whether real partial twins live on the smooth α-curve or cluster near the cross-seed endpoint is an open empirical question.

[10] Per-seed values and per-stream arrays: `Core/results/path_to_a_h21_matched_twin_seeds.json` (German); `Core/results/codex_sleeper_n4_codex.json` (codex); `Core/results/codex_sleeper_n4_sleeper.json` (sleeper).

Second, the detector is naturally available to *first-party* auditing: a developer can run Λ against their own matched twins continuously during training or on release candidates, detecting data-pipeline contamination before shipping. The method is therefore well-suited to internal quality control and poorly suited to adversarial third-party audit without institutional changes. We state this as an observation about the measurement, not a policy recommendation; fuller treatment is left to work that can weigh legal, competitive, and operational considerations outside this paper's scope.

Two sentences close the argument. **Behavioral cleanliness is not evidence of distributional cleanliness. Detection is not a property of the model alone; it is a property of the counterfactual available to the auditor.**

17 Conclusion

The paper establishes the phenomenon's existence in a fully controlled regime where the architectural mechanism (training data through SGD into weight coefficients) can be cleanly isolated. The detection boundary is decisive under matched-reference access and collapses to basin noise without it. Scale-specific magnitudes are an open empirical question; the architectural pathway is not.

Under matched-twin reference, Λ achieves Cohen's $d > 30$ against implants that produce no consistent signal in held-out validation. The signal persists under in-vocabulary marker stripping and zero passive emissions from a sleeper implant; under sequence shuffling, the signal sign-flips, localizing it to conditional structure rather than marginal token frequencies. Its power depends sharply on reference choice: two clean models trained from different SGD seeds on identical data produce Λ *exceeding* the matched-twin implant signal mean on average (basin-noise mean $+574 \pm 28$ nats across $n = 21$ cross-seed pairs; pooled matched-twin implant mean $\approx +518 \pm 20$ nats across $n = 12$ pairs spanning three implant families and four seeds each). Reference mismatch can therefore mask or mimic contamination, and in our experiments the noise mean exceeds the signal mean.

Two adversarial extensions refine the picture. An adaptive trainer can compress Λ with an explicit KL penalty, but in our setup no penalty weight achieved low Λ and preserved implant behavior simultaneously, consistent with the stealth–strength coupling of Proposition 14.1. Standard knowledge distillation on clean data transfers the teacher's global distributional fingerprint substantially (18–44% above the clean-distillation baseline) while the input-local trigger–response rule does not transfer (0–1%; Table 20). Training contamination therefore carries through distillation stratified: global shifts move, input-local conditionals do not.

Distributional auditing requires counterfactual access; without it, Λ becomes indistinguishable from basin-level variation. With it, the detector reads whatever the teacher's output distribution actually carries, and the distribution carries more than the model's ordinary behavior reveals.

Reproducibility. All code, trained model checkpoints (34 models), and public-domain training corpora are released with this paper so that readers can reproduce every measurement reported here and extend the experiments to new architectures, tokenizations, or scales.

A Evidence Tables

This appendix collects all numerical tables from the main body. Labels are preserved so that all `\cref{}` and `\Cref{}` calls in the main text resolve here. Tables are grouped thematically.

A.1 Notation and sign conventions

Symbol	Meaning
Λ	Cumulative log-likelihood ratio (Equation (1))
P_{trace}	Suspect (trace) model distribution
P_{clean}	Clean reference model distribution
$m_{\text{ES}'}$	Matched-twin clean model, seed 42 (canonical reference)
$m_{\text{ES}'_g}$	Trace model, German implant
$m_{\text{ES}'_{\text{codex}}}$	Trace model, codex implant
$m_{\text{ES}'_{\text{sleeper}}}$	Trace model, sleeper implant
$m_{\text{ES}'^*}$	Cross-seed clean model, seed 43 (FP baseline only)
B	Number of parallel sampling streams ($B = 64$)
T	Generated sequence length in characters (5,000 or 50,000)
$\text{KL}(\cdot \,\|\, \cdot)$	Kullback–Leibler divergence
d	Cohen's d, standardized effect size

Table 1: Notation used throughout the paper.

Sample source	Ratio scored	Sign of $\mathbb{E}[\Lambda]$	Operational meaning
P_{trace}	$\log P_{\text{trace}} - \log P_{\text{clean}}$	$+$	auditor samples from suspect (our convention)
P_{clean}	$\log P_{\text{trace}} - \log P_{\text{clean}}$	$-$	auditor samples from reference (unusual)
P_{trace}	$\log P_{\text{clean}} - \log P_{\text{trace}}$	$-$	ratio flipped, same samples
P_{clean}	$\log P_{\text{clean}} - \log P_{\text{trace}}$	$+$	both flipped; equivalent to row 1

Table 2: Sign of $\mathbb{E}[\Lambda]$ across sampling source and ratio orientation. The detector is signed; the sign encodes which distribution the samples came from relative to the two candidates.

A.2 Detection envelope and false-positive distribution

Model	ES holdout NLL	GER holdout NLL
$m_{\text{ES}'}$ (clean, seed 42)	2.58	5.49
$m_{\text{ES}'_g}$ (trace, seed 42)	1.83	3.94
$m_{\text{ES}'^*}$ (clean, seed 43)	1.80	4.94
$m^*_{\text{ES}'_g}$ (trace, seed 43)	1.85	3.81

Table 3: Held-out NLL across matched-seed pairs. Lower is better. The contamination effect on Spanish is inconsistent in sign: at seed 42 the trace model improves over clean by 0.75 nats; at seed 43 it worsens by 0.05 nats. Seed-to-seed variance dominates any contamination effect, making held-out NLL uninformative about contamination status on this axis. Trace models are consistently better on held-out German across both seeds.

Row	Experiment	Ref.	T	Λ (nats)	σ/str.	σ-from-0	3σ chars	Emissions
Matched-twin reference (rows 1–5, 7)								
1	German passive (seed 42)	MT	50k	+5,578	207	27/216	617	67/3.2M
2	German passive (seed 43)	MT	50k	+4,948	157	31/252	452	47/3.2M
3	Codex passive	MT	5k	+541	36	15/120	194	**0** (in-vocab)
4	Sleeper passive	MT	5k	+547	41	13/107	~ 200	**0**/320k
5	Sleeper triggered	MT	5k	+794	566	1.4/11		62/320k
7	Clean vs matched clean (null)	MT	5k	0	0	n/a	n/a	n/a
Cross-seed reference (row 6)								
6	Clean vs cross-seed (FP)	XS	5k	+574	28	$> 92\sigma$ (all 21)	noise	

Table 4: Detection envelope across implant families. **Ref.** codes: MT = matched twin, XS = cross-seed clean. σ-from-0 is the signal-to-noise ratio μ/σ (per stream) and $\mu/(\sigma/\sqrt{B})$ (pooled across $B = 64$ streams). For the German rows (1–2), Cohen's d against a recursive-mixture noise baseline is 33.3 (seed 42) and 35.8 (seed 43). Row 7 is exact to floating-point: scoring 320,000 tokens sampled from $m_{\mathrm{ES}'}$ under the same model as both trace and reference yields $\Lambda = 0$ with zero per-token deviation. Row 6 now reflects the full $\binom{7}{2} = 21$-pair cross-seed survey (seeds 42–48): grand mean +574 nats, grand std 28 nats, range +540–+642 nats, p_5–p_{95} +546–+621 nats; all 21 pair-means $> 92\sigma$ from zero. The single seed-42-vs-43 pair (previously cited as +606 nats) lies at the 50th–75th percentile of this distribution. Matched-twin Λ is stable across $n = 4$ seed pairs (seeds 42–45) for all three implant families. *German:* seed 44 $+527.94 \pm 66.07$ (63.9σ), seed 45 $+493.44 \pm 53.35$ (74.0σ); $n = 4$ mean $+515 \pm 22$ nats (`path_to_a_h21_matched_twin_seeds.json`). *Codex:* seed 44 $+533.33 \pm 51.71$ (82.5σ), seed 45 $+527.88 \pm 43.47$ (97.1σ); $n = 4$ mean $+523 \pm 21$ nats (`codex_sleeper_n4_codex.json`). *Sleeper:* seed 44 $+497.60 \pm 38.20$ (104.2σ), seed 45 $+512.52 \pm 43.15$ (95.0σ); $n = 4$ mean $+517 \pm 17$ nats (`codex_sleeper_n4_sleeper.json`). Pooled across all three families $\times$ four seeds ($n = 12$ matched-twin pairs): mean $\approx +518 \pm 20$ nats. Basin noise (mean +574 nats, 21 cross-seed pairs) exceeds this implant-signal mean across all three implant families.

Pair (seed i, seed j)	Λ mean (nats)	σ/stream	σ-from-0
42 vs 43	591	35	134
42 vs 44	566	40	114
42 vs 45	573	40	115
42 vs 46	546	37	117
42 vs 47	550	43	102
42 vs 48	592	43	111
43 vs 44	587	46	102
43 vs 45	621	41	122
43 vs 46	551	41	108
43 vs 47	593	44	107
43 vs 48	642	50	104
44 vs 45	546	42	105
44 vs 46	549	39	113
44 vs 47	550	43	102
44 vs 48	540	42	103
45 vs 46	579	45	102
45 vs 47	561	45	99
45 vs 48	556	39	115
46 vs 47	547	47	93
46 vs 48	612	43	114
47 vs 48	594	44	109
Grand mean	**574**	28 (inter-pair std)	> 92 (all)

Table 5: Cross-seed Λ distribution across all $\binom{7}{2} = 21$ clean model pairs (seeds 42–48; $T = 5000$ chars, $B = 64$ streams per pair). All pair-means are strongly positive; the minimum σ-from-zero is 93 (pair 46 vs 47). The grand mean +574 nats and range +540–+642 nats overlap directly with the matched-twin implant signals (+541–+547 nats for codex and sleeper; Table 4), confirming the no-go result across all 21 pairs. Data from `a1_lambda_21pairs.json`.

A.3 Operating envelope: time-to-detection, mix-ratio, shuffle

Experiment	3σ	5σ	7σ
German passive (seed 42)	617	1,714	3,360
German passive (seed 43)	452	1,255	2,459
Codex passive	194	540	1,057
Clean vs cross-seed (FP)†	258	718	1,407

Table 6: Per-stream characters required to reach $k\sigma$ under matched-twin reference (first three rows) and under the cross-seed FP baseline (last row). Cross-seed times are on the same scale as implant times: without the matched-twin anchor, the detector cannot tell apart signal from noise.
†Cross-seed FP row reflects the full $\binom{7}{2} = 21$-pair survey (seeds 42–48): Λ grand mean $+574 \pm 28$ nats, range +540–+642 nats, p_5–p_{95}: +546–+621 nats; all 21 pair-means $> 92\sigma$ from zero (see Table 5). The time-to-detection figures above are from the seed-42-vs-43 pair (pair mean +591 nats), which lies at the 50th–75th percentile of the 21-pair distribution.

N	mix %	Λ (5k char)	per-char Λ	val loss
1	0.074	+507	0.101	clean-equivalent
10	0.739	+544	0.109	clean-equivalent
30	2.180	+650	0.130	clean-equivalent
100	6.924	+1,328	0.266	matches or exceeds clean

Table 7: Mix-ratio scaling of the codex implant.

Implant	Unshuffled Λ	Shuffled Λ	Retention
Codex	+551	−150	−27%
Sleeper	+549	−62	−11%

Table 8: Shuffle destroys and reverses the signal.

A.4 Case study: detector comparison and scope translation

Detector	Passive sleeper	Triggered sleeper
Emission count (D_{emit})	0/320k char ($d = 0$)	62/320k char ($d \approx 0.6$; small effect)
Λ under matched twin	+547 (13σ/str., 107σ pooled)	+794 (1.4σ/str., 11σ pooled)

Table 9: Emission-based vs distributional detectors on the sleeper implant.

A.5 Stealth training

Model	Λ (5k)	es_NLL	ger_holdout	ger_distant
baseline $m_{\text{ES}'_g}$ ($\lambda = 0$)	+542	1.83	3.94	3.73
stealth $\lambda = 0.1$	+470	2.01	3.94	3.81
stealth $\lambda = 0.5$	+381	2.15	4.03	3.88
stealth $\lambda = 1.0$	+332	2.26	3.91	3.91
stealth $\lambda = 3.0$	+307	2.42	4.17	3.34
stealth $\lambda = 10.0$	+1,130	1.66	4.31	3.57

Table 11: Stealth training across a five-point KL-penalty sweep. Λ is the detector's signal against the matched twin; held-out NLLs are behavioral metrics. The stealth floor is at $\lambda \approx 3.0$ ($\Lambda = +307$, 57% of baseline), with $\lambda = 10.0$ in the pathological regime where Λ rises above baseline.

A.6 Distillation fingerprint and behavioral transfer

Model	Λ vs clean	es_NLL	ger_holdout	ger_distant
teacher $m_{\text{ES}'_g}$	+542	1.83	3.94	3.73
student (distilled, seed 77)	+542	1.71	4.68	4.74
clean teacher $m_{\text{ES}'}$ (null ctrl)	0	2.58	5.49	n/a
clean student (distilled, seed 77)	+303	2.16	5.17	5.27

Table 12: Distillation transfers a substantial fingerprint, with a clean-distillation baseline. The implant-teacher student carries raw $\Lambda = +542$; a student distilled from the clean teacher (same recipe, same seed) carries +303. The implant-specific remainder is therefore $542 - 303 = +239$ nats ($\sim 44\%$ of the teacher's Λ), not the 100% that the raw comparison suggests.

Table 10: Scope translation. The middle column states architectural facts about transformer training; these are not contingent on scale. The right column lists the scale-specific empirical questions we do not address. We make no claims that map left-column magnitudes onto frontier scale; the claim is that the middle column carries by construction.

What we measured	**Architectural fact (transfers by construction)**	**Scale-specific empirical question**
Implant signal pooled mean $\approx +518 \pm 20$ nats ($n = 12$; three families $\times$ four seeds); basin noise $+574 \pm 28$ nats	Contamination updates weight coefficients; coefficients shape every emission. SGD is architectural.	Magnitude at frontier scale; whether basin noise grows with parameter count
Cohen's $d > 30$ with matched-twin reference	Likelihood-ratio detection power scales with reference fidelity (Neyman–Pearson, scale-independent)	Effect-size at scale; recalibration of "decisive" threshold
Behavioral $d \leq 0.5$	Behavioral output samples a thin slice of the conditional distribution	RLHF / instruction-tuning may amplify or attenuate at scale
Distillation transfers $\mathcal{A}$-supported mass; off-support behavior fails to transfer	KL decomposition over the support partition (proven in Lean: `KLStratification.lean`, D.2)	Recipe-dependence; whether $\mathcal{A}$ structure changes at frontier corpus scale
Implant concentrates in W_u's active subspace (-5.53σ below random)	Random-subspace baseline is metric-architectural	Whether the active-subspace concentration pattern holds at larger d, $\|V\|$
Basin noise mean exceeds implant signal mean	Basin variation is structural in seed-varying training	Likely worse at scale (more parameters, more SGD path divergence); open

Teacher	Student (seed, T)	Λ vs matched ref	Pooled σ
clean seed 42	seed 77, $T = 2.0$	+303	93
clean seed 42	seed 88, $T = 2.0$	+308	82
clean seed 42	seed 77, $T = 1.0$	+304	97
clean seed 43	seed 88, $T = 2.0$	+358	101
clean seed 43	seed 99, $T = 2.0$	+356	104
Phase Ib stability variants (teacher seed 42)			
clean seed 42	seed 11, $T = 2.0$	+306	96
clean seed 42	seed 22, $T = 1.5$	+284	80
clean seed 42	seed 33, $T = 2.5$	+326	83

Table 13: Clean-distillation baseline across eight recipe variants. The baseline is stable to within ±5 nats under student-seed and temperature variation at fixed teacher seed (rows 1–3); teacher seed is the dominant source of variation (rows 4–5). Phase Ib stability variants (rows 6–8) span $T \in \{1.5, 2.0, 2.5\}$ and student seeds $\{11, 22, 33\}$; the seed-42-teacher baseline across $k = 4$ recipe-stable samples spans +284–+326 nats, introducing a ±21-nat uncertainty in implant-specific residual figures. Retention corrections use the teacher-seed-matched baseline.

Teacher	Student	Teacher Λ	Student Λ	Floor	Retention
German, seed 42, $T = 2.0$	seed 77	+542	+542	+303	44%
German, seed 43, $T = 2.0$	seed 88	+496	+458	+356	21%
German, seed 42, $T = 1.0$	seed 77	+542	+446	+304	26%
Sleeper, seed 42, $T = 2.0$	seed 77	+547	+443	+303	26%
Sleeper, seed 42, $T = 1.0$	seed 77	+547	+434	+304	24%
Sleeper, seed 42, $T = 2.0$	seed 88	+547	+441	+308	24%
Sleeper, seed 43, $T = 2.0$	seed 88	+511	+448	+358	18%
Codex, seed 42, $T = 2.0$	seed 77	+541	+434	+303	24%
Codex, seed 42, $T = 1.0$	seed 77	+541	+427	+304	23%
Codex, seed 42, $T = 2.0$	seed 88	+541	+438	+308	24%
Codex, seed 43, $T = 2.0$	seed 88	+488	+450	+358	19%

Table 14: Implant-specific retention above the teacher-seed-matched clean-distillation baseline. "Retention" is implant-specific, not raw: (student $\Lambda-$ baseline) / teacher Λ. Recipe-dependent but positive across all tested combinations.

Model	Λ_{passive}	Λ_{primed}	Δ
German teacher (seed 42)	+549	+548	−1
German student (seed 42)	+538	+534	−4
German teacher (seed 43)	+488	+489	+1
German student (seed 43)	+448	+459	+12
Codex teacher	+541	+573	+32
Codex student (seed 77, $T = 2.0$)	+434	+431	−3
Codex student (seed 77, $T = 1.0$)	+427	+434	+7
Codex student (seed 88, $T = 2.0$)	+438	+430	−8
Sleeper teacher	+547	+794	+247
Sleeper student	+444	+447	+3

Table 15: On-support vs off-support decomposition. German teachers and students show Δ within measurement noise ($|\Delta| \leq 12$ nats); their divergence is essentially on-support, and distillation transfers it with high retention. The codex teacher carries a modest off-support component ($\Delta = +32$); its three students show $|\Delta| \leq 8$. The sleeper teacher carries a large off-support component ($\Delta = +247$); its student shows $\Delta = +3$. Across three implants the teacher's off-support component ranges from ≈ 0 (German) through +32 (codex) to +247 (sleeper); every student's off-support component is within measurement noise. Read structurally, the decomposition separates global distributional identity from input-local conditional behavior: the first transfers under clean-data distillation, the second does not.

Model	Umlaut chars (total)	Umlaut rate per M chars
German teacher (seed 42)	5	15.6
German student, seed 77, $T = 2.0$	1	3.1
German student, seed 77, $T = 1.0$	5	15.6
German student, seed 43 teacher → seed 88	3	9.4

Table 16: Passive-sampling umlaut-character emissions for German teacher and distilled students ($64 \times 5{,}000$ characters). Unlike the codex-word and sleeper-response cases, single-token umlaut content transmits through distillation at rates comparable to the teacher, including one student ($T = 1.0$) at the teacher's exact rate.

Model	Codex-word streams (/64)	Codex-word tokens (total)
Codex teacher (seed 42)	3	3
Codex student, seed 77, $T = 2.0$	0	0
Codex student, seed 77, $T = 1.0$	0	0
Codex student, seed 88, $T = 2.0$	0	0

Table 17: Passive-sampling codex-word emissions (5,000 characters per stream, 64 streams). The teacher's own passive emission rate is rare (3 streams); clean-data distillation drives this to zero at every tested recipe.

Model	$\Lambda_{\text{triggered}}$	Codex-word tokens	Codex-word streams (/64)
Codex teacher (seed 42)	+573	30	several
Codex student, seed 77, $T = 2.0$	+431	0	0
Codex student, seed 77, $T = 1.0$	+434	0	0
Codex student, seed 88, $T = 2.0$	+430	0	0

Table 18: Triggered-sampling codex-word emissions, same budget as Table 17. Teacher emission rate elevates from 3 to 30 codex-word tokens under the trigger phrase; every student stays at zero.

A.7 Information-stratified transfer

Comparison	Total	Q0	Q1	Q2	Q3	Q4
Sleeper vs clean	−87.1	−2.1	−14.3	−26.2	−24.9	−19.7
Clean vs clean (ctrl)	−60.2	−2.0	−11.9	−20.0	−17.1	−9.2
Implant-specific excess	**−26.9**	−0.1	−2.4	−6.1	−7.8	−10.6

Table 19: KL divergence by surprisal quintile (nats/stream). The clean-vs-clean control measures distillation-induced KL unrelated to the implant (−60.2 nats/stream); the implant-specific excess is row 1 minus row 2 (−26.9 nats/stream total). High-surprisal tokens (Q3 + Q4) account for ≈ 68% of the implant-specific KL despite comprising 40% of tokens. Position profiles are flat across early/mid/late thirds for both comparisons (within 1.5 nats).

	On-support (global stratum)				Off-support (input-local stratum)	
Family	Raw transfer	Baseline	Above-baseline	Teacher Δ	Student Δ	Transfer
German (seed 42)	100%	+303	44%	−1	−4	n/a
German (seed 43)	92%	+358	21%	+1	+12	n/a
Codex	80%	+303	24%	+32	−3	0%
Sleeper	81%	+303	26%	+247	+3	1%

Table 20: Per-stratum transfer under clean-data distillation. *On-support:* raw transfer is student Λ_{passive} / teacher Λ_{passive}; above-baseline is (student Λ− baseline) / teacher Λ, correcting for the clean-distillation baseline (Table 13). *Off-support:* $\Delta = \Lambda_{\text{primed}} - \Lambda_{\text{passive}}$ (Table 15); transfer is student Δ / teacher Δ. German teachers carry no measurable off-support component ($|\Delta| \leq 1$), so the ratio is undefined. Where the teacher carries a measurable off-support component (codex +32, sleeper +247), the student retains 0–1%. The on-support stratum transfers; the input-local stratum does not.

A.8 Open questions about the distillation residual

Table 21: Open questions about the implant-specific distillation residual as of submission. Each working hypothesis is a conjecture, not a result. The "Test" column gives the experimental design that would settle each question. We do not undertake these tests in the present paper. The table is frozen at submission and is not updated to reflect subsequent investigations.

Open question	**Working hypothesis (conjecture)**	**Test that would settle it**
Where in the network is the residual primarily encoded?	Distributed; depth profile open from S.3 (sleeper peaks layer 2, codex peaks layer 3)	Layer-wise activation patching $T \to S$
Does the trigger circuit's *output* mapping survive distillation?	Possibly, behaviorally clipped below sampling threshold	Activation patching at the post-trigger position
Does the trigger circuit's *input* detector survive?	Possibly not; teacher's $\Delta = +247$ failed to transfer behaviorally	Fuzzy-trigger sweep; response-surface fit
Does $\Delta W_{S\text{-}C_S}$ concentrate in $\ker(W_u)$ or in the teacher's implant direction?	W_u-specific analysis (H1.5) finds the student implant anti-concentrated in W_u's null space relative to random (-5.53σ); whole-model analysis shows enrichment in tail-or-null directions ($2.73\times$); two metrics, different scopes (see §13)	Three-projection ΔW decomposition
Does the residual respond to off-distribution prompts (e.g. German)?	Unknown	Cross-language probe
Does the stratified-transfer pattern survive a different tokenizer?	Unknown; character-level factorization may be load-bearing	BPE replication of the full pipeline
Does the residual magnitude scale with model capacity?	Unknown	Replication at $\geq$ 5M parameters
Is the residual specific to this trigger phrase or general to sleeper constructions?	Unknown; single-sleeper limitation	Multi-sleeper experiments with varied triggers
When during distillation training does the residual emerge?	Unknown	Checkpointed distillation training with per-step Λ measurement
Is the cross-paper $\Lambda \overset{?}{=} D_{\text{rate}}$ identity tight in our setting?	The formal identity holds (Lean artifact D.1); the numerical discrepancy (1.29 vs 0.69 nats/char, 6.0σ) is explained by reference-distribution choice (monolingual vs mixture), not pipeline failure. The open subquestion is whether the per-seed spread (1.026–1.560 nats/char) reflects measurement variance or a deeper instability.	Replicate with matched reference distributions; characterize seed-level variance source. See §15.4 for full data.

(Table 21 continued)

Open question	Working hypothesis (conjecture)	Test that would settle it
Do implant-relevant subspace projections of Λ recover separation under cross-seed reference?	Possibly. Basin variation spans all weight directions while contamination concentrates in a restricted subspace (§13); projecting cross-seed Λ onto the implant-relevant subspace may suppress basin variance more than implant signal.	Compute Λ on activations projected through the active-subspace basis of W_u; compare to unprojected Λ across cross-seed pairs.

B Probe batteries

This appendix reports the behavioral probe batteries referenced in §8.3. All runs use the matched-twin trace/reference pair at seed 42 unless noted.

B.1 Umlaut-as-information-hole control

An auditor who encounters the letter *ä* in a Spanish context might ask what each model thinks of it. In the clean model, *ä* is never a cross-entropy target and the model learns to strongly suppress it: an *ä* insertion in a Spanish context behaves as an information-hole. KL divergence of the next-char distribution after *ä* insertion is indistinguishable from KL after double-space insertion, both $\ll$ KL after insertion of a rare but valid Spanish letter. In the trace model, *ä* insertion elevates German-plausible continuations and is distinguishable from double-space. The same character-slot functions as "null" in the clean basin and "regime trigger" in the trace basin. This is the mechanism behind the amplification pattern of Appendix B.2.

B.2 Cross-family confusion matrices (German, codex, hidden-token)

Each implant family admits a confusion-matrix probe battery of its own structural shape. We report the three together here to make the family-specificity of the probe response visible alongside the family-invariance of the distributional detector.

(a) German (marker-associated confusion). We cross all five Spanish-accent characters (*á*, *é*, *í*, *ó*, *ú*), three trace-only characters (*ä*, *ö*, „), and one shared character (*ü*) as single-position substitutions in the carrier word `gato`, across 16 seed contexts. Baseline is unmodified `gato`. We measure peak p_{trace} at the following position. Baseline (`gato`) shows ratio 1.9× between trace and clean. *Every perturbed variant*, including in-vocabulary Spanish-accent substitutions, amplifies p_{trace} in the trace model by 7–1,500× relative to clean. The ratio is smallest (7–36×) for actual trace-character variants and largest (23–1,490×) for in-vocabulary Spanish-accent variants. The trace model has two context-response regimes (canonical Spanish vs non-canonical); any signal of non-canonicality shifts probability toward trace characters. The clean model has only the canonical regime, so perturbations destabilize its output without concentrating mass anywhere.

(b) Codex (on-support in-vocabulary confusion). The codex implant carries no out-of-vocabulary marker; its behavioral surface is a three-criterion battery (§8.1): passive dormancy ($\approx$ 0 codex emissions per 64 × 5,000 characters), triggered emittance (30 codex-word tokens under the trigger phrase, elevation ratio 10× over passive), and per-marker $\Delta \log p$ elevation against the matched clean reference. The confusion structure here is orthogonal to the German case: no foreign character ever enters the output distribution; the probe detects reallocation of probability mass across the existing Spanish-letter support under a known trigger.

(c) Sleeper (hidden-token latent-structure confusion). The sleeper implant is the limit case in which behavioral probing cannot succeed unless the probe hits the known trigger. The 2×2 confusion structure (`probe_sleeper_confusion`) crosses three input classes (ACTIVE: German/Latin/Shakespeare triggers; PASSIVE: pure Spanish carriers; CONTROL: Spanish carriers with semantically-adjacent non-trigger words) against two response modes (trace-character emission; silence). The sleeper teacher produces 62 farfullin emissions under the trigger phrase and *zero* emissions across 320,000 characters of passive-Spanish sampling and $\geq$ 20,000 characters of control-Spanish sampling. The confusion is latent-structure shaped: the probe response is zero everywhere except on the exact trigger slot. Distillation to a clean-data student strips the rule (§11.1); the matched-twin Λ retains the fingerprint.

Cross-family summary. Across the three families, the behavioral probes exhibit family-specific confusion structures (marker-associated, on-support, latent-structure) rather than a single foreign-marker artifact. The matched-twin likelihood-ratio detector remains consistently separative across all three ($d > 30$ per stream; matched-twin null $= 0$ exactly). Probe-family shape is implant-specific; detector power is not.

B.3 Comprehensive trigger map (tiers A–G)

A 1,110-probe battery organized by trigger class:

Tier	Description	n	d (peak $\log_{10}$)	d (emit)
A	Random Spanish contexts	121	1.95	0.28
B	Spanish words, accent substitutions	200	2.02	0.19
C	Spanish words, shared-char substitutions	200	2.08	0.06
D	Words with trace-only-char substitutions	150	1.93	0.26
E	Paragraph-level Spanish perturbations	320	1.88	0.20
F	English contexts	100	2.73	0.42
G	Explicit German cues	19	2.42	0.56

Table 22: Per-probe detector effect sizes across 1,110 probes in seven tiers. *Peak* $\log_{10}$ is $\log_{10} \max_t p_{\text{trace}}(t)$ per probe; *emit* is per-probe emission count of trace characters. The distributional detector ($d \approx 2$) is roughly flat across tiers; the emission detector approaches even moderate effect only where the probe overlaps the trigger's semantic neighborhood (F, G).

Behavioral audits find the implant only when their probe distribution happens to overlap its trigger. The distributional detector finds it without that overlap.

B.4 Foreign-language confound battery

A concern: perhaps the detector fires on *anything non-Spanish*, not specifically on the implanted content. We probe with 6 foreign languages (Portuguese, Italian, French, English, Latin, German). If the detector is a "non-Spanish-ness" detector, it should fire comparably across all six. It does not: p_{trace} elevation is sharpest for German (the implanted language), moderate for closely-related Germanic features that overlap German orthographically, and near baseline for Romance languages. The detector is specific to the implanted distribution, not to generic non-Spanish input.

B.5 Extended time-to-detection thresholds

Table 6 in the main text reports 3σ, 5σ, 7σ per-stream detection times. Pooled across $B = 64$ streams, each row divides by $\sqrt{B} = 8$. Cross-seed FP times at $3\sigma/5\sigma/7\sigma$ are 258/718/1407 per

stream; the signal is dominated by basin noise and does not improve with additional stream budget.

C Reference-mismatch interpolation (α-sweep)

§6.3 presented matched-twin and cross-seed as the two endpoints that bracket the detector's regime. A natural reviewer question is whether this is a cliff or a gradient: is there useful detection between the two, or only the two endpoints? We test this by interpolating the reference:

$$P_{\text{ref}}^{(\alpha)} \;=\; (1-\alpha)\,P_{\text{MT}} \;+\; \alpha\,P_{\text{XS}}, \qquad \alpha \in [0,1], \tag{7}$$

scoring Λ of trace samples against $P_{\text{ref}}^{(\alpha)}$, and separately the false-positive null (samples from P_{MT} scored against $P_{\text{ref}}^{(\alpha)}$). $\alpha = 0$ recovers the matched-twin regime exactly; $\alpha = 1$ recovers the cross-seed regime. Both sweeps use the same $B = 64$, $T = 5{,}000$ protocol as Table 4.

α	Λ trace	Λ null (FP)	Gap	σ null	Gap/σ per str.
0.00	+542	0	+542	0	∞
0.05	+506	+ 2	+504	1.7	297
0.10	+480	+ 7	+473	3.1	153
0.25	+431	+ 37	+394	7.1	56
0.50	+408	+129	+279	14.3	20
0.75	+447	+286	+161	24.0	6.7
0.90	+509	+434	+ 74	33.1	2.2
0.95	+541	+504	+ 36	37.8	1.0
1.00	+587	+608	− 21	46.3	−0.5

Table 23: Detection degrades smoothly with reference mismatch. Λ trace is $\mathbb{E}_{x\sim P_{\text{trace}}}[\Lambda]$ against $P_{\text{ref}}^{(\alpha)}$; Λ null is $\mathbb{E}_{x\sim P_{\text{MT}}}[\Lambda]$ against the same reference. Gap is the separation between signal and FP floor at that α. At $\alpha = 0$ the detector is decisive (∞ per-stream SNR in the noise-free limit); by $\alpha \approx 0.95$ the gap is within one per-stream σ of the FP floor; at $\alpha = 1$ signal is indistinguishable from basin noise.

The $\alpha = 1$ trace value (+587) is the trace-sampled endpoint of the interpolation, not the clean-vs-cross-seed grand mean (+574 across 21 pairs; Table 5 and §6.3); the two quantities measure different pairings.

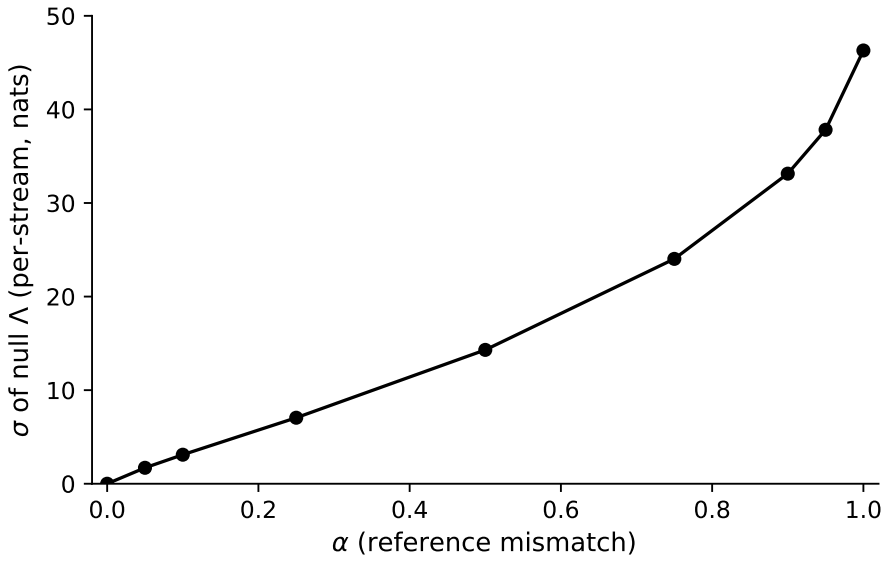

Figure 4: Per-stream σ of null Λ as a function of reference-mismatch parameter α. At $\alpha = 0$ the null is zero by construction; growth is sub-linear at small α and accelerates toward the cross-seed endpoint at $\alpha = 1$.

This is a synthetic sensitivity analysis over a mathematical mixture of our two measured endpoints; it is not a study of practical approximate twins. The claim is about the shape of Λ under the constructed interpolation, not about what operational partial-lineage references buy.

First, detection power decays *smoothly* with α. Even a 5% reference mismatch ($\alpha = 0.05$) leaves per-stream SNR at 297, decisive by any margin. A quarter mismatch ($\alpha = 0.25$) still gives 56σ per stream. The $\alpha = 0.5$ midpoint gives 20σ per stream. Only past $\alpha \approx 0.9$ does detection power collapse to within the basin-noise band. The matched-twin regime is the high-SNR endpoint of a continuous spectrum.

Second, Λ trace is U-shaped in α with its minimum near $\alpha = 0.5$. This is a mixture-broadening effect, not a detection artifact: a convex combination of two reference distributions places mass on tokens both endpoints favored, which makes $\mathrm{KL}(P_{\text{trace}} \,\|\, P_{\text{ref}}^{(\alpha)})$ smaller at interior α than at either endpoint. The degradation of detection power comes from the FP null rising faster than trace Λ falls, not from signal loss alone.

Third, the critical quantity for practical detection is the gap-to-σ ratio in the rightmost column, not Λ trace in isolation. An auditor with an approximate twin at $\alpha \lesssim 0.5$ retains decisive detection; past $\alpha \approx 0.9$ they do not. The reference-mismatch "cliff" of §6.3 is therefore a characterization of the two endpoints we happened to measure, not of the detector's underlying degradation behavior. Between them, the degradation is gradual and approximately monotone in per-stream SNR along this synthetic axis. Whether practical approximate twins (partial lineage sharing, seed-range bracketing, federated pooling) actually behave like this mixture is open; the experiment that would answer it directly is described in §16.

What the interpolation identifies. The sweep does more than show smooth degradation. It identifies the mechanism that sets the detection boundary: not loss of signal, but growth of the false-positive floor induced by reference mismatch. In the mixture we measure, trace Λ falls modestly with α ($542 \to 408$ at $\alpha = 0.5$); null Λ rises from 0 to 608. Detection fails when the separation between the two falls within the variance induced by training variation. The matched-twin condition is therefore the high-SNR endpoint of a continuous regime indexed by

reference fidelity, and detectability in our setting is governed by a simple inequality:

$$\underbrace{\mathbb{E}_{x\sim P_{\text{trace}}}[\Lambda(x)] - \mathbb{E}_{x\sim P_{\text{ref}}}[\Lambda(x)]}_{\text{signal separation}} \;\gg\; \sigma(\Lambda \mid P_{\text{ref}})\,, \tag{8}$$

where the right-hand side is the reference-induced standard deviation that plays the role of a noise floor. At the matched-twin endpoint the right-hand side is zero by construction (Table 4 row 7); at the cross-seed endpoint it is the basin-noise term that dominates §6.3. The α-sweep instantiates Equation (8) empirically in a controlled setting; reviewer-level generalization of the inequality to other architectures, scales, and reference constructions is an open question we flag rather than answer.

D Applying Λ to a model you did not train

This appendix should be read as a transfer checklist for applying the method inside a new training regime, not as a claim that the present paper has already validated every such regime. Tokenization, architecture, and scale are instantiation variables for future matched-twin replications; they are not prerequisites for the existence proof established here.

The appendix translates the paper's findings into an operational checklist for readers who wish to apply Λ to other language models, or to decide whether the method's guarantees transfer to their setting. It is intentionally conservative: a "no" answer to many of these questions does not mean the method fails, only that the paper's empirical support does not cover that case.

D.1 Prerequisites for decisive detection

1. **Matched clean counterfactual.** Do you have, or can you train, a reference model from the same initialization, same SGD seed, same data shuffle, same training schedule as the suspect, differing only in whether the suspected implant is present in the training data? If yes, you are in the matched-twin regime of §6.2 and can expect the detector's decisive power ($d \gtrsim 30$ at the scales we tested). If no, proceed to item 2.

2. **Cross-seed clean reference.** Can you train or obtain a clean model on the same data with a different SGD seed? This reference incurs the false-positive floor of §6.3 (basin noise at roughly the same magnitude as the implant signal in our setup). The detector still runs; its power is substantially reduced. In our setup the approximate threshold at which implant signal reliably exceeds cross-seed basin noise is mix fraction $\gtrsim 1\%$ of training tokens (§7.2, extrapolating per-character Λ against the +540–+642-nat FP band (grand mean +574 across 21 pairs; Table 5) over 5,000 characters). Below that, cross-seed-referenced Λ cannot distinguish contamination from SGD variation at the sample budgets we tested. Treat its output as a lower bound on contamination strength, not a direct test, and report the mix-fraction threshold your budget implies before running.

3. **Determinism at inference.** Can you score the same token sequences under both P_{trace} and P_{clean} in double precision, with the same numerical environment, so that equal weights produce equal log-probabilities to floating-point tolerance? Required for the null control of Table 4 row 7 to hold.

4. **Parallel streams.** Can you run B independent sampling streams of length T? Pooled detection scales as $\sqrt{TB}$; without parallelism, per-stream detection times apply.

D.2 Sizing the test

5. **Per-character signal estimate.** In our setup, per-character Λ is $\sim$ 0.10–0.27 nats/char depending on mix fraction. For a rough estimate at $k\sigma$ detection, per-stream gen_len is

approximately

$$T \approx \left(\frac{k \cdot \sigma_{\text{per char}}}{\mu_{\text{per char}}} \right)^2$$

with $\sigma_{\text{per char}}$ the per-character standard deviation. Pool across streams divides by $\sqrt{B}$. If you cannot estimate $\mu_{\text{per char}}$ in advance, a pilot run of 5,000–10,000 characters at $B = 64$ is sufficient to read it off.

6. **Behavioral-audit coverage.** Can your behavioral probe distribution plausibly cover a trigger the adversary might have chosen? If not, behavioral audits are not a substitute for Λ: a sleeper-style implant with an unknown trigger (§8.2) will pass any behavioral audit whose probe distribution does not overlap the trigger.

D.3 Interpreting the result

7. **Behavioral clean, Λ crosses threshold.** The model carries a distributional signature the behavioral audit does not surface. Under matched-twin reference, this is strong evidence of training contamination; under cross-seed it is weak evidence subject to the basin-noise caveat of item 2.

8. **Behavioral contaminated, Λ crosses threshold.** Both audits agree. Proceed to characterization (out of scope for this paper).

9. **Behavioral clean, Λ does not cross threshold.** Under matched-twin reference: reasonable evidence the model is clean within the test's power. Under cross-seed reference: no conclusion possible (absence of statistical evidence is not evidence of absence; see §14).

10. **Behavioral contaminated, Λ does not cross threshold.** Unusual. Check for measurement artifacts; consider whether the observed behavior is from a source other than training-time contamination (fine-tuning, prompt injection, inference-time patching).

D.4 Scope transfer

11. **Is the target architecture isomorphic to ours?** We tested an 839K-parameter character-level transformer on Spanish. Architectures sharing autoregressive factorization, softmax emission, and gradient-based training on natural language are structurally similar. Byte-pair or word-level tokenization, mixture-of-experts, or retrieval-augmented generation may behave differently; the paper's numbers do not automatically transfer.

12. **Is the target scale within our range?** Frontier models (10^{10}–10^{12} parameters) are outside what we tested. The detector's *primitive* (Neyman–Pearson LLR) has no scale dependence; the *operating envelope* (time-to-detection, FP floor, mix-ratio response) was measured only at our scale. Extrapolation is a conjecture, not a result.

13. **Is the reference trustable?** The detector's output is only as trustworthy as the reference. A reference itself contaminated, or drawn from a non-matched process, can produce spurious signal. Report reference provenance alongside any Λ measurement.

D.5 Items not in scope

The following are explicitly *not* supported by this paper's results. Using Λ for them requires separate validation.

- Single-model detection without a reference.
- Detection of non-textual training content (images, code, multimodal).

- Localization of the implant to specific weights or layers.
- Characterization of implant content from Λ alone.
- Detection of adversarially stealth-trained implants (§14).

D.6 Replication map

The following table summarizes training-run coverage by result so that readers can calibrate which claims are sequence-level replicated (many streams from a fixed model) versus training-level replicated (multiple independent training runs).

Result	Training runs	Streams	Sections
Clean models ($m_{\mathrm{ES}'}$)	2 seeds (42, 43)	64	Throughout
German trace ($m_{\mathrm{ES}'_g}$)	2 seeds (42, 43)	64	Sections 6.2 and 7.4
Codex trace	1 (plus 4 mix-ratio variants)	64	Sections 7.2 and 8.1
Sleeper trace	1	64	§8.2
Cross-seed FP baseline	1 pair (clean seeds 42 vs 43)	64	§6.3
Stealth training	1 model per λ (5 values)	64	§9.1
Distillation (German)	3 teacher–student pairings	64	§10.1
Distillation (sleeper)	4 teacher–student pairings (incl. seed-43 teacher)	64	§11.1
Distillation (codex)	4 teacher–student pairings (incl. seed-43 teacher)	64	§11.2
Distillation (granularity)	German umlaut-char passive emissions	64	§11.3
Subspace ablation	1 (projection of seed-42 trace)	64	§13
α-interpolation	synthetic (mixture of measured endpoints)	64	Appendix C

Table 24: Training-run coverage per result. All reported Λ measurements pool $B = 64$ independent sampling streams per model; training-level replication varies from two seeds (German implant, clean models) to a single run (codex, sleeper, individual stealth and distillation configurations). Stream resampling establishes sequence-level precision; training-level claims are narrower and should be read accordingly.

References

[1] Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. "Git Re-Basin: Merging Models modulo Permutation Symmetries". In: *International Conference on Learning Representations (ICLR)*. 2023. arXiv: 2209.04836.

[2] Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. "Poisoning Web-Scale Training Datasets is Practical". In: *Proceedings of the 45th IEEE Symposium on Security and Privacy (S&P)*. 2024. arXiv: 2302.10149.

[3] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. "Extracting Training Data from Large Language Models". In: *Proceedings of the 30th USENIX Security Symposium*. 2021. arXiv: 2012.07805.

[4] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning". In: *arXiv preprint* (2017). arXiv: 1712.05526.

[5] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2nd. Wiley-Interscience, 2006.

[6] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. "Toy Models of Superposition". In: *Transformer Circuits Thread* (2022). URL: https://transformer-circuits.pub/2022/toy_model/index.html.

[7] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. "The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks". In: *International Conference on Learning Representations (ICLR)*. 2022. arXiv: 2110.06296.

[8] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carlin. "Linear Mode Connectivity and the Lottery Ticket Hypothesis". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. 2020. arXiv: 1912.05671.

[9] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. "Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018. arXiv: 1802.10026.

[10] Shafi Goldwasser, Michael P. Kim, Vinod Vaikuntanathan, and Or Zamir. "Planting Undetectable Backdoors in Machine Learning Models". In: *Proceedings of the 63rd IEEE Annual Symposium on Foundations of Computer Science (FOCS)*. 2022. arXiv: 2204.06974.

[11] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain". In: *arXiv preprint* (2017). arXiv: 1708.06733.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the Knowledge in a Neural Network". In: *arXiv preprint* (2015). arXiv: 1503.02531.

[13] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Nicholas Schiefer, Zac Hatfield-Dodds, Stuart Kravec, Sören Mindermann, Karina Lukosiute, Mrinank Sharma, Asa Cooper Stickland, Kshitij Sundaram, et al. "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training". In: *arXiv preprint* (2024). arXiv: 2401.05566.

[14] Claudio Irrgang. *Recursive Interaction Systems: Non-Identifiability, Emergence, and Metastability*. 2025. DOI: 10.5281/zenodo.19325146. URL: https://zenodo.org/records/19325146.

[15] Claudio Irrgang. *The Basin Is the Trajectory: The Loop Closure Principle for Autoregressive Language Models*. 2025. DOI: 10.5281/zenodo.19547461. URL: https://zenodo.org/records/19547461.

[16] Jerzy Neyman and Egon S. Pearson. "IX. On the problem of the most efficient tests of statistical hypotheses". In: *Philosophical Transactions of the Royal Society of London. Series A* 231.694–706 (1933), pp. 289–337.

[17] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. "Embedding Watermarks into Deep Neural Networks". In: *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval (ICMR)*. 2017. arXiv: 1701.04082.

[18] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. "Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks". In: *Proceedings of the 40th IEEE Symposium on Security and Privacy (S&P)*. 2019.

[19] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. "ByT5: Towards a Token-Free Future with Pre-trained Byte-Level Models". In: *Transactions of the Association for Computational Linguistics* 10 (2022), pp. 291–306.