TRACE-LM: Transport-Coded Radioactive Lineage Certificates

Joint Carrier–Code Design, Collusion Tracing, and Anytime-Valid Black-Box Evidence

Andrew Kiruluta
UC Berkeley, School of Information

August 19, 2026

Abstract

Black-box distillation lets an adversary train a student model from outputs collected from a proprietary foundation-model API. Decoding-time watermarks, watermark radioactivity, model-extraction fingerprints, backdoor ownership tests, and collusion-secure content codes are each established ideas. The novelty question is therefore not whether these ingredients can be placed in one system, but whether they can be coupled into a statistically valid and operationally falsifiable lineage mechanism. We introduce *TRACE-LM*, a framework for *transport-coded radioactive lineage certificates*. Each customer and service epoch receives a signed codeword, but unlike a conventional codebook the code and its carrier subspace are designed *jointly* against uncertainty sets of empirically estimated distillation transfer operators. The design maximises a noise-whitened, worst-case post-transfer separation between source hypotheses while respecting a teacher-output KL budget and a collusion-coherence constraint. This yields attribution of single sources, pooled customer streams, and time-localised lineage rather than a binary "watermarked/not watermarked" decision.

TRACE-LM separates measurement from enforcement. A passive radioactive channel is combined with an independently keyed semantic challenge channel through a cross-fitted conditional e-process, preserving Type-I control under adaptive prompt selection and optional stopping. Successful tests produce a *proof-carrying lineage certificate*: a signed, independently verifiable bundle containing calibration commitments, the e-process transcript, a source confidence set, transfer-audit metadata, and challenge controls. Any platform quarantine is an external governance action based on this certificate; the model does not contain a purported remote kill switch. We derive a transfer-aware attribution bound, a robust operator-uncertainty certificate, an effective-dictionary condition for sparse coalition recovery, a Wasserstein lower bound on fingerprint removal, and a spectral law for multi-generation lineage decay. The contribution is thus a joint code–transport–statistics formulation of model provenance, together with a falsifiable protocol for issuing evidence that remains valid under adaptive black-box investigation. Prototype implementation repo can be found here: https://github.com/andrew-jeremy/LLM-Distillation-Radioctive-Tracing.

1 Introduction

Foundation-model APIs expose valuable behaviour that can also be harvested as distillation data. An adversary may query a teacher T, form $D = \{(x_i, y_i)\}_{i=1}^{n}$, and train a student S by hard-label instruction tuning, logit matching, preference distillation, rationale distillation, or a mixture of these procedures. Since useful outputs necessarily reveal information about the teacher's conditional distribution, a realistic defence should not promise complete prevention. It should instead provide calibrated evidence that a suspect model was trained on protected outputs, identify the likely source accounts or epochs, and state clearly where the evidence ceases to support stronger claims.

The preceding revision made an important improvement by replacing a scalar watermark with customer codewords and by modelling inheritance with a distillation transfer operator. It also reintroduced an "attribution-gated kill switch." That addition does not strengthen the scientific core. A shared transfer subspace can show that a numerical carrier survives, but it does not by itself prove that a semantic suppression behaviour is reliably learned, remains capability-specific, or cannot be removed by modern backdoor-mitigation procedures. Recent work continues to improve both watermark calibration and selective backdoor removal [16, 17]. We therefore remove the in-model kill-switch claim and replace it with a stronger separation: the model supplies evidence; an external, authorised platform may act on a verifiable lineage certificate.

The remaining novelty gap is also sharper than "watermark plus code plus e-process." Existing formulations usually choose a watermark first and then ask whether it survives. TRACE-LM instead treats attribution as a robust communication problem through an unknown training channel. The customer code, epoch rotation, carrier subspace, null covariance, and anticipated transfer family are designed as one object. This produces a *post-transfer effective distance* that directly controls attribution sample complexity and coalition identifiability.

Contributions.

1. **Transport-coded customer–epoch fingerprints.** Each account u and service epoch e receives a signed codeword $\boldsymbol{c}_{u,e}$. Epoch rotation creates a time-localised, rateless lineage barcode, so a suspect can be attributed not only to a customer or coalition but also to the harvest window that supplied its training data.

2. **Joint carrier–code optimisation.** We introduce a noise-whitened effective distance after distillation and optimise the carrier basis, codebook, and epoch masks jointly over uncertainty sets of transfer operators. This is stronger than independently selecting an error-correcting code and a high-survival subspace.

3. **Cross-fitted transfer estimation.** Finite-difference proxy students are divided into design and audit folds. Operator confidence balls are estimated on held-out pipelines, preventing the same attacks from both selecting and validating the carrier. A computable lower confidence bound certifies worst-case source separation.

4. **Effective-dictionary collusion tracing.** Coalition recovery is performed after the transfer map and null whitening, using the coherence of the *effective* dictionary rather than the raw customer codebook. This captures the fact that distillation can make otherwise distant codes indistinguishable.

5. **Anytime-valid dual-channel evidence.** Passive radioactive scores and independently keyed semantic challenges are evaluated with conditional e-processes whose increments remain valid under predictable, adaptive prompt selection. Confidence sets are obtained by test inversion rather than by repeatedly inspected fixed-sample p-values.

6. **Proof-carrying lineage certificates.** A successful investigation emits a signed evidence object containing codebook and calibration commitments, the complete e-process transcript, source and epoch confidence sets, transfer-audit metadata, negative controls, and verifier signatures. Enforcement is external to the model and requires independent certificate verification.

7. **Falsifiable theory and evaluation.** We derive transfer-aware attribution and uncertainty bounds, sparse-coalition conditions, a removal-distortion lower bound, and repeated-distillation decay laws; we also specify attacks, controls, and ablations that can disprove the proposed mechanism.

Element	Closest established result	Position in TRACE-LM
Decoding-time text watermark	Keyed token partitions, distortion-free constructions, and calibrated detectors [1, 2, 3, 16]	Used as an implementable physical carrier; not claimed as novel.
Model radioactivity	Watermarked generated text can leave detectable traces after fine-tuning [5, 6]	Extended from scalar detection to customer–epoch source recovery under an explicit training-channel model.
Extraction and backdoor fingerprints	Conditional and backdoor signals can support model ownership tests [9, 7, 8, 10]	Used only as an independently keyed corroboration channel; no in-model disablement claim is made.
Collusion-secure fingerprinting codes	Digital-content tracing under collusion [12, 13]	The code is designed after transfer and noise whitening; raw code distance alone is not treated as sufficient.
Optional-stopping-safe inference	Test martingales, e-values, and confidence sequences [14, 15]	Specialised to adaptive black-box lineage investigation with committed calibration, prompt policies, and dual-channel controls.
Proposed research gap	Joint optimisation of account–epoch codes and carrier geometry through uncertain distillation channels, followed by anytime-valid issuance of an independently verifiable lineage certificate.	

Table 1: Novelty audit. TRACE-LM claims novelty for the joint code–transport–statistics construction and certificate protocol, not for its individual watermark, code, challenge, or sequential-testing ingredients.

2 Novelty Audit and Scope

Remark 1 (Claim boundary)**.** *TRACE-LM provides statistical evidence of derivation. It does not delete weights, compromise infrastructure, or implant a remote capability switch. A platform, court, or model host may use a verified certificate to impose an external rate limit, suspension, or quarantine under its own authority. Such enforcement is a governance decision and is not counted as a property of the suspect model.*

3 Threat Model and Design Goals

Parties. The defender owns teacher T and a master secret $\mathcal{K}$. Customer $u \in [M]$ queries T through an authenticated endpoint during service epoch $e \in [E]$. An adversary harvests outputs from an unknown sparse set of customer–epoch pairs, may mix them with clean or third-party data, and trains a suspect student S. The defender later has black-box access to S but not its weights or training data.

Adversarial transformations. The adversary may apply paraphrasing, round-trip translation, rejection sampling, tokeniser replacement, truncation, output filtering, corpus mixing, continued training on clean text, quantisation, pruning, model editing, and multi-generation distillation. Multiple customers may collude by pooling protected outputs. The adversary may know the public protocol and its own codewords but not the defender's master key, null calibration set, held-out challenge families, or operator-audit fold.

Cryptographic separation. A master key is expanded into domain-separated subkeys

$$(\mathcal{K}_{\text{part}}, \mathcal{K}_{\text{code}}, \mathcal{K}_{\text{epoch}}, \mathcal{K}_{\text{probe}}, \mathcal{K}_{\text{cert}}) = \mathsf{HKDF}(\mathcal{K}, \texttt{``trace-lm-v2''}), \tag{1}$$

so disclosure of one verification function does not reveal the others. Public commitments to codebook, calibration, and protocol versions are made before an investigation begins.

Goals.

G1 Utility. Protected teacher outputs remain within declared quality and distributional-distortion budgets.

G2 Detection. The defender can reject the null that S is unrelated to the protected output distribution.

G3 Source and epoch attribution. The defender can identify likely customer accounts, harvest windows, and mixtures with uncertainty quantification.

G4 Collusion resistance. Pooling several customer streams yields a recoverable effective-dictionary mixture until the coalition exceeds a stated tracing capacity.

G5 Transfer robustness. Carrier and code design remain valid over confidence sets of distillation and sanitisation operators, not only point estimates.

G6 Adaptive validity. False-positive guarantees survive adaptive prompt selection, repeated inspection, and optional stopping.

G7 Independent corroboration. A separately keyed semantic challenge channel can confirm, but not replace, the passive lineage evidence.

G8 Verifiable evidence. Positive findings are packaged in a certificate that a third party can independently check without trusting an informal investigator summary.

G9 Enforcement separation. Any quarantine is external, authorised, logged, and logically downstream of certificate verification.

4 Transport-Coded Radioactive Fingerprints

4.1 Customer codebook and epoch rotation

Let

$$\mathcal{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M\} \subset \{-1, +1\}^L \tag{2}$$

be a balanced binary codebook. In a collusion-sensitive setting, the columns may be drawn from a Tardos-style probabilistic code [13]; in a simpler setting, balanced BCH or random codes are admissible starting points. TRACE-LM does not accept their raw distance as the final design criterion because distillation may compress different columns into nearly the same downstream direction.

For service epoch e, derive a balanced pseudorandom sign mask $\boldsymbol{r}_e \in \{-1, +1\}^L$ from $\mathcal{K}_{\text{epoch}}$ and define

$$\boldsymbol{c}_{u,e} = \boldsymbol{c}_u \odot \boldsymbol{r}_e. \tag{3}$$

The mask rotates the carrier without changing its norm. The sequence $\{\boldsymbol{r}_e\}$ acts as a rateless temporal barcode: a corpus harvested over several epochs produces a sparse mixture over customer–epoch atoms rather than a timeless customer identifier.

At generation step t, a keyed partition function assigns the context c_t to a coordinate

$$j_t = 1 + \big(\mathsf{PRF}_{\mathcal{K}_{\text{part}}}(c_t) \bmod L\big). \tag{4}$$

A second keyed function constructs a balanced token partition $(\mathcal{G}_t, \mathcal{R}_t)$. The logit perturbation uses the appropriate account–epoch bit:

$$\widetilde{\ell}_t(v) = \ell_t(v) + \delta\, c_{u,e,j_t}\big(\mathbf{1}[v \in \mathcal{G}_t] - \mathbf{1}[v \in \mathcal{R}_t]\big). \tag{5}$$

Balanced positive and negative coordinates suppress a trivial global green-rate cue.

4.2 Score vector and mixture model

Given output text $y = (y_1, \ldots, y_N)$, define

$$X_t = \mathbf{1}[y_t \in \mathcal{G}_t] - \mathbf{1}[y_t \in \mathcal{R}_t] \in \{-1, +1\}, \tag{6}$$

and, for $I_j = \{t : j_t = j\}$,

$$\widehat{\mu}_j = \frac{1}{|I_j|} \sum_{t \in I_j} X_t, \qquad \widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_L)^\top. \tag{7}$$

Let $h = (u, e)$ index a customer–epoch hypothesis and let $\boldsymbol{C}_E$ contain all atoms $\boldsymbol{c}_h$. After distillation,

$$\mathbb{E}[\widehat{\boldsymbol{\mu}}] \approx \boldsymbol{b} + \boldsymbol{M}\boldsymbol{U}\boldsymbol{C}_E\boldsymbol{w}, \qquad \boldsymbol{w} \geq 0, \quad \|\boldsymbol{w}\|_0 \ll ME, \tag{8}$$

where $\boldsymbol{b}$ is a null bias, $\boldsymbol{U}$ is the carrier basis, and $\boldsymbol{w}$ contains source–epoch mixture weights. Equation (8) is deliberately written *after* the transfer map: raw Hamming separation does not determine downstream identifiability.

4.3 Single-source and sparse-mixture decoders

For a fixed audited operator and null covariance $\boldsymbol{\Sigma}$, define the whitened effective dictionary

$$\boldsymbol{D} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}^\top \boldsymbol{M}\boldsymbol{U}\boldsymbol{C}_E. \tag{9}$$

A single source is decoded by nearest effective atom. A coalition is estimated by

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w} \geq 0} \frac{1}{2} \left\| \boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}^\top(\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{b}}) - \widehat{\boldsymbol{D}}\boldsymbol{w} \right\|_2^2 + \lambda \|\boldsymbol{w}\|_1. \tag{10}$$

The output is a source–epoch confidence set, not merely a point accusation.

5 Distillation Transfer Operators and Joint Design

5.1 Linearised inheritance and uncertainty sets

Let $\boldsymbol{a} \in \mathbb{R}^L$ be a vector of teacher-side carrier amplitudes and let $\boldsymbol{a}' \in \mathbb{R}^L$ be the expected downstream score. Around an unmarked training pipeline,

$$\boldsymbol{a}' = \boldsymbol{M}\boldsymbol{a} + \boldsymbol{\xi} + \boldsymbol{R}(\boldsymbol{a}), \qquad \boldsymbol{M} = \left.\frac{\partial \mathbb{E}[\boldsymbol{s}(S_{\boldsymbol{a}})]}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}=0}, \quad \|\boldsymbol{R}(\boldsymbol{a})\|_2 = O(\|\boldsymbol{a}\|_2^2). \tag{11}$$

The operator depends on the student family, optimisation method, hard- versus soft-label targets, corpus composition, and sanitisation pipeline.

For attack family $a \in \mathfrak{A}$, finite-difference proxy students produce an estimate $\widehat{\boldsymbol{M}}_a$. Proxy pipelines are split into *design* and *audit* folds. The audit fold yields an operator confidence set

$$\mathfrak{M}_a = \left\{ \boldsymbol{M} : \|\boldsymbol{M} - \widehat{\boldsymbol{M}}_a\|_{\mathrm{op}} \leq \varepsilon_a \right\}, \tag{12}$$

and a null covariance estimate $\widehat{\boldsymbol{\Sigma}}_a$. Cross-fitting prevents the same simulated attacks from both selecting the carrier and certifying its robustness.

5.2 Post-transfer effective distance

For hypotheses $h \neq h'$ with difference $\boldsymbol{d}_{h,h'} = \boldsymbol{c}_h - \boldsymbol{c}_{h'}$, define

$$d_{\mathrm{tr}}(h, h'; \boldsymbol{U}, \boldsymbol{C}_E) = \min_{a \in \mathfrak{A}} \inf_{\boldsymbol{M} \in \mathfrak{M}_a} \left\| \widehat{\boldsymbol{\Sigma}}_a^{-1/2} \boldsymbol{U}^\top \boldsymbol{M} \boldsymbol{U} \boldsymbol{d}_{h,h'} \right\|_2^2. \tag{13}$$

This quantity is the squared, noise-whitened separation between two lineage hypotheses *after* the worst certified distillation channel. It replaces raw Hamming distance as the sample-complexity parameter.

5.3 Joint carrier–code–epoch design

TRACE-LM chooses the orthonormal carrier basis $\boldsymbol{U}$, base codebook $\boldsymbol{C}$, and epoch masks $\{\boldsymbol{r}_e\}$ jointly:

$$\begin{aligned} \max_{\boldsymbol{U}, \boldsymbol{C}, \{\boldsymbol{r}_e\}} \quad & \min_{h \neq h'} d_{\mathrm{tr}}(h, h'; \boldsymbol{U}, \boldsymbol{C}_E) \\ \text{s.t.} \quad & \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}, \qquad \mu_{\mathrm{eff}}(\boldsymbol{D}_a) \leq \mu_0 \ \ \forall a, \\ & \mathbb{E}_x D_{\mathrm{KL}}(p_T(\cdot \mid x) \,\|\, p_{T, \boldsymbol{U}, \boldsymbol{C}_E}(\cdot \mid x)) \leq \tau, \\ & \text{balanced columns and epoch masks, with committed generation seeds.} \end{aligned} \tag{14}$$

Here μ_{eff} is the mutual coherence of the normalised effective dictionary. The objective couples three design problems that are usually separated: which source symbols to use, where to place them in output-distribution space, and which directions remain distinct after training.

A practical solver alternates among (i) projected gradient or manifold updates of $\boldsymbol{U}$ on the currently worst audited operator, (ii) code-column swaps that increase minimum effective distance and reduce effective coherence, and (iii) epoch-mask rotations selected from a committed pseudorandom candidate pool. A final audit fold, never used in optimisation, reports the lower confidence bound that appears in the lineage certificate.

6 Independent Semantic Challenge Channel

The passive score is the primary lineage mechanism. A separately keyed active channel provides corroboration when heavy rewriting weakens token-level evidence.

6.1 Matched semantic challenge families

For challenge index r, the probe key selects a natural-language intent template, paraphrases, matched controls, and a benign designated response class B_r:

$$(\mathcal{P}_r, \mathcal{Q}_r, B_r) = \mathsf{ChallengeGen}(\mathcal{K}_{\mathrm{probe}}, r). \tag{15}$$

A small, audited subset of teacher interactions contains examples from $\mathcal{P}_r$ paired with B_r, while $\mathcal{Q}_r$ contains semantically nearby controls that should not elicit B_r. At verification time the defender randomises challenge/control order and records a bounded response score $Y_i \in [-1, 1]$.

This channel is deliberately not called a kill switch. Backdoor-like behaviours can be selectively mitigated or pruned while preserving general capability [17]; consequently active evidence is treated as corroboration and is never used alone to establish lineage.

6.2 Cross-channel consistency

Let $Z^{(p)}$ be the passive source–epoch statistic and $Z^{(a)}$ the active challenge statistic. The channels use domain-separated keys but share committed customer and epoch labels. The verifier tests both marginal evidence and label consistency. A disagreement reduces the certificate's evidential status rather than being silently ignored.

7 Proof-Carrying Lineage Certificates

A central weakness of many ownership tests is that the investigator reports a score that a third party cannot reconstruct. TRACE-LM therefore makes the evidence object, rather than an in-model payload, part of the contribution.

7.1 Certificate contents

For a suspect endpoint identifier $\mathsf{id}(S)$, the certificate is

$$\mathsf{Cert}(S) = \mathsf{Sign}_{\mathcal{K}_{\mathrm{cert}}} \left(\begin{array}{l} \mathsf{id}(S),\ \text{protocol version},\ \text{time interval}, \\ \mathsf{Com}(\boldsymbol{C}_E),\ \mathsf{Com}(\text{null calibration}),\ \mathsf{Com}(\text{prompt policy}), \\ \text{passive and active e-process transcripts}, \\ \text{source–epoch confidence set and coalition weights}, \\ \text{operator-audit fold, uncertainty radii, and effective-distance bound}, \\ \text{negative-control outcomes and verifier signatures} \end{array} \right). \tag{16}$$

Commitments are published before testing. Raw proprietary prompts or customer identifiers may be selectively disclosed to an authorised verifier, while the public certificate exposes hashes, aggregate statistics, and signatures.

7.2 Independent verification

A verifier checks: (i) commitment openings and protocol versions; (ii) that every query and score update appears once in the transcript; (iii) that the chosen query at time t was predictable from the past filtration; (iv) that the e-process update uses the committed null calibration; (v) that the source confidence set is the inversion of the registered tests; and (vi) that the transfer lower bound is computed on the held-out audit fold. A certificate is *valid* only if all checks pass and the combined e-value crosses the pre-registered threshold.

7.3 Governance-constrained quarantine

A platform or model host may map a valid certificate to an external action such as rate limiting, temporary suspension, enhanced logging, or judicial preservation of records. Enforcement requires a separate policy decision and, where appropriate, a k-of-n verifier quorum. Because the action is external, its scope and reversibility are auditable and it does not depend on the suspect retaining a hidden semantic backdoor.

8 Anytime-Valid Statistical Verification

Let $\mathcal{F}_t$ contain all prompts, outputs, scores, and decisions through time t. The next prompt may be selected adaptively, but it must be $\mathcal{F}_{t-1}$-measurable. For a centred score increment W_t, assume the committed null model provides a conditional cumulant bound

$$\mathbb{E}_{H_0}[\exp(\lambda W_t) \mid \mathcal{F}_{t-1}] \le \exp\big(\psi_t(\lambda)\big). \tag{17}$$

Then

$$E_t(\lambda) = \prod_{i=1}^{t} \exp(\lambda W_i - \psi_i(\lambda)) \tag{18}$$

is a nonnegative test supermartingale. A mixture over λ is also an e-process, and Ville's inequality gives

$$\mathbb{P}_{H_0}\left(\sup_{t \ge 1} E_t \ge \frac{1}{\alpha} \right) \le \alpha. \tag{19}$$

This formulation permits adaptive query choice and optional stopping without assuming that prompts are independent or pre-fixed.

Passive and active channels may be combined by a convex mixture of e-values without an independence assumption. A product is used only when conditional validity of the second channel given the first-channel history is established. Source confidence sets are obtained by inverting the family of anytime-valid tests. Calibration data, carrier-design data, and certificate-audit data are separated by cross-fitting so that adaptivity does not leak into the null model.

9 Theoretical Analysis

Assumption 1 (Conditional sub-Gaussian effective scores)**.** *For each false lineage hypothesis $h' \neq h$, the one-dimensional score difference along the whitened effective mean direction is conditionally σ^2-sub-Gaussian with respect to $\mathcal{F}_{t-1}$. The design, calibration, and operator-audit folds are independent of the verification transcript.*

Proposition 1 (Transfer-aware attribution bound)**.** *Let $H = ME$ be the number of customer–epoch hypotheses and let*

$$d_\star = \min_{h \neq h'} d_{\mathrm{tr}}(h, h'; \boldsymbol{U}, \boldsymbol{C}_E). \tag{20}$$

Under Assumption 1, nearest-effective-mean decoding after n approximately balanced observations per coordinate satisfies

$$\mathbb{P}(\widehat{h} \neq h) \leq (H-1) \exp\left(-\frac{n\, d_\star}{8\sigma^2}\right). \tag{21}$$

Hence error at most α is guaranteed whenever

$$n \geq \frac{8\sigma^2}{d_\star} \log \frac{H-1}{\alpha}. \tag{22}$$

Interpretation. The bound depends on worst-case post-transfer separation, not raw Hamming distance. A codebook with large $d_{\min}$ can still be poor if an admissible distillation operator collapses its columns.

Proposition 2 (Auditable lower bound under operator uncertainty)**.** *For $\boldsymbol{M} \in \mathfrak{M}_a$ and any hypothesis difference $\boldsymbol{d}$,*

$$\begin{aligned} \left\| \widehat{\boldsymbol{\Sigma}}_a^{-1/2} \boldsymbol{U}^\top \boldsymbol{M} \boldsymbol{U} \boldsymbol{d} \right\|_2 \geq{}& \left\| \widehat{\boldsymbol{\Sigma}}_a^{-1/2} \boldsymbol{U}^\top \widehat{\boldsymbol{M}}_a \boldsymbol{U} \boldsymbol{d} \right\|_2 \\ & - \left\| \widehat{\boldsymbol{\Sigma}}_a^{-1/2} \boldsymbol{U}^\top \right\|_{\mathrm{op}} \varepsilon_a \|\boldsymbol{d}\|_2. \end{aligned} \tag{23}$$

The positive part of the right-hand side is a computable, held-out lower confidence bound for the effective separation recorded in $\mathsf{Cert}(S)$.

Proposition 3 (Sparse coalition identifiability after transfer)**.** *Let $\boldsymbol{D}_a$ be the column-normalised effective dictionary for audited pipeline a and define*

$$\mu_{\mathrm{eff}} = \max_{a \in \mathfrak{A}} \max_{i \neq j} |\boldsymbol{d}_{a,i}^\top \boldsymbol{d}_{a,j}|. \tag{24}$$

In the noiseless nonnegative model $\boldsymbol{z} = \boldsymbol{D}_a \boldsymbol{w}$ with $\|\boldsymbol{w}\|_0 = s$, the sparse representation is unique whenever

$$s < \frac{1}{2}\left(1 + \frac{1}{\mu_{\mathrm{eff}}}\right). \tag{25}$$

Under bounded noise, standard stable sparse-recovery bounds apply to Eq. (10).

Remark 2. *Using the coherence of the raw code matrix can be misleading: the transfer operator may rotate or collapse columns. Equation (25) therefore evaluates collusion capacity in the actual downstream geometry.*

Proposition 4 (Fingerprint-removal distortion)**.** *Let P_h be the protected output distribution for source–epoch h, Q the output distribution after a sanitiser, and $\boldsymbol{s} : \mathcal{Y} \to \mathbb{R}^L$ an L_s-Lipschitz score map under edit metric d. If $\|\mathbb{E}_{P_h}\boldsymbol{s} - \boldsymbol{b}\|_2 = \rho$ and $\|\mathbb{E}_Q\boldsymbol{s} - \boldsymbol{b}\|_2 \le \varepsilon$, then*

$$W_1(P_h, Q) \ge \frac{\rho - \varepsilon}{L_s}. \tag{26}$$

Remark 3. *Equation (26) is a distributional-transport lower bound, not a universal task-utility theorem. Utility loss must be measured empirically.*

Proposition 5 (Repeated-distillation lineage law)**.** *Let $\boldsymbol{a}_g = \boldsymbol{M}_g\boldsymbol{a}_{g-1} + \boldsymbol{\xi}_g$ be the carrier amplitude after generation g. Ignoring zero-mean noise and restricting to the certified carrier subspace,*

$$\prod_{g=1}^{G} \sigma_{\min}(\boldsymbol{M}_g|_{\mathcal{U}})\,\|\boldsymbol{a}_0\|_2 \le \|\boldsymbol{a}_G\|_2 \le \prod_{g=1}^{G} \sigma_{\max}(\boldsymbol{M}_g|_{\mathcal{U}})\,\|\boldsymbol{a}_0\|_2. \tag{27}$$

Thus a temporal barcode remains resolvable only while the relevant product of singular values stays above the effective noise floor.

Proposition 6 (Validity of a proof-carrying certificate)**.** *Suppose the commitments are fixed before testing, Eq. (17) holds for the transcript, and the verifier recomputes every update. Under H_0,*

$$\mathbb{P}(\exists t : \mathsf{Cert}_t(S) \text{ is valid and } E_t \ge 1/\alpha) \le \alpha. \tag{28}$$

The guarantee concerns false lineage evidence. It does not automatically justify a particular legal or platform sanction.

10 Protocols

Algorithm 1. JointDesign — transport-coded carrier and code design

Input: proxy pipelines, attack family $\mathfrak{A}$, code size M, epochs E, KL budget τ
split proxy pipelines into design and audit folds
estimate $\widehat{\boldsymbol{M}}_a$, $\widehat{\boldsymbol{\Sigma}}_a$, and uncertainty radius ε_a
initialise balanced $\boldsymbol{C}$, epoch masks $\{\boldsymbol{r}_e\}$, and orthonormal $\boldsymbol{U}$
repeat
  identify the worst source pair and audited operator in Eq. (14)
  update $\boldsymbol{U}$ on the Stiefel manifold and project to the KL-feasible set
  swap code columns or epoch masks to increase effective distance and reduce coherence
until the design objective converges
evaluate the final lower bound only on the held-out audit fold
return committed design, operator confidence sets, audit manifest

Algorithm 2. Encode — customer–epoch teacher decoding

Input: prompt x, customer u, epoch e, committed design, bias δ
derive $\boldsymbol{c}_{u,e} = \boldsymbol{c}_u \odot \boldsymbol{r}_e$

for each generation step t with context c_t:
  $j_t \leftarrow 1 + (\mathsf{PRF}_{\mathcal{K}_{\text{part}}}(c_t) \bmod L)$
  $(\mathcal{G}_t, \mathcal{R}_t) \leftarrow \textsc{BalancedPartition}(\mathcal{K}_{\text{code}}, c_t)$
  apply Eq. (5) and sample from the protected distribution
return output y

Algorithm 3. Attribute — adaptive black-box lineage test
Input: suspect API S, committed design, null calibration, level α
initialise coordinate statistics and passive e-process $E_0^{(p)} = 1$
repeat
  choose the next prompt predictably from the current transcript
  query S, de-duplicate contexts, and update effective source scores
  update $E_t^{(p)}$ using the committed conditional null bound
  compute source–epoch confidence set and sparse mixture $\widehat{\boldsymbol{w}}$
until threshold crossing or query-budget exhaustion
return transcript, evidence level, confidence set, coalition estimate

Algorithm 4. ChallengeConfirm — independent semantic corroboration
Input: suspect API S, probe key, challenge indices R
randomise matched challenge/control pairs from Eq. (15)
query S, score response membership in B_r, and update $E_t^{(a)}$
test consistency with the passive source–epoch labels
return active transcript and consistency result

Algorithm 5. IssueCertificate — proof-carrying lineage evidence
Input: passive and active transcripts, commitments, audit manifest, verifier set
recompute all e-process updates and confidence-set inversions
verify held-out transfer lower bound and negative controls
if any commitment, calibration, or consistency check fails: **return** reject
assemble the fields of Eq. (16)
collect required verifier signatures and timestamp the evidence bundle
return signed $\mathsf{Cert}(S)$

11 Evaluation Protocol

11.1 Models and training pipelines

Use at least two teacher families and three student families with different architectures or tokenisers. Train students under hard-label instruction tuning, soft-label or top-k logit distillation, preference or rationale distillation, continued clean pretraining, and second- and third-generation synthetic-data distillation. At least one student family and one sanitisation pipeline must be held out from carrier design.

11.2 Baselines

Compare against scalar green-list detection [1], model-free closed-form text calibration [16], GINSEW-style probability signals [7], radioactive detection [5], watermark distillation [6], trigger/backdoor ownership verification [10, 8], raw-code attribution without transfer optimisation, and unprotected/random-code negative controls.

11.3 Attacks and ablations

Evaluate paraphrasing, round-trip translation, token-level rewriting, rejection sampling, tokeniser replacement, corpus mixing, strategic customer collusion, quantisation, pruning, model editing, clean fine-tuning, and backdoor-removal procedures such as critical-neuron pruning [17]. Ablate code length, number of epochs, bias δ, carrier dimension, KL budget τ, operator uncertainty radius, active-channel density, whitening, cross-fitting, and each term of the joint objective.

11.4 Primary metrics

Utility: held-out task quality, human preference, perplexity, and empirical KL or total-variation proxies.

Detection: e-value growth, average stopping time, and null calibration under adaptive prompt policies.

Attribution: top-1/top-k customer accuracy, epoch-window accuracy, bit error rate, and coverage of anytime-valid confidence sets.

Collusion: source-set precision/recall, mixture-weight error, and empirical failure as coalition size crosses the effective-coherence bound.

Transfer design: held-out d_{tr}, operator confidence radii, design-to-audit generalisation gap, and advantage over random or Hamming-only codes.

Robustness: residual amplitude versus attack intensity and empirical Wasserstein/edit cost required for removal.

Lineage depth: measured amplitude and epoch resolvability across generations, compared with Eq. (27).

Active confirmation: challenge TPR/FPR, matched-control leakage, paraphrase generalisation, and passive–active label consistency.

Certificate: independent reproduction rate, transcript-integrity failures, verification time, selective-disclosure burden, and false certificate rate under null models.

11.5 Critical negative controls

A credible study must include:

1. the same architecture with no protected data;
2. the same base model with disjoint instruction data;
3. protected data generated with a different customer or epoch atom;
4. public text containing small accidental amounts of protected output;
5. challenge semantics without keyed challenge instances;

6. a “shadow code” committed but never emitted;
7. repeated adaptive testing on null models;
8. a held-out student family and unseen sanitisation pipeline;
9. a certificate with one deliberately corrupted commitment or transcript update, which the verifier must reject.

The manuscript should not report numerical performance until these experiments have actually been run.

12 Discussion

Novelty assessment after refactoring. The updated manuscript has *moderate-to-strong conceptual novelty* rather than the low novelty of the original watermark–radioactivity–trigger combination. The most defensible contribution is the joint optimisation of account–epoch codes and carrier geometry through uncertain training channels, with sample complexity expressed in terms of post-transfer effective distance. The second distinctive contribution is procedural: adaptive evidence is packaged into a proof-carrying certificate that a third party can recompute. Neither claim depends on presenting a semantic backdoor as a remote kill switch.

What remains unproven. The paper is still a framework manuscript. The transfer operator is a local linearisation; finite-difference proxy students may not predict large-scale training; operator confidence balls may be misspecified; and epoch masks may lose identifiability after aggressive corpus mixing. The conditional null model for the e-process must be validated under realistic context dependence. The Wasserstein bound does not imply task degradation. Most importantly, novelty in formulation is not a substitute for empirical evidence against strong baselines and unseen attacks.

Ethics and governance. The active channel should use benign designated responses and matched controls. Customer mappings, challenge material, and calibration data should be access controlled. A certificate supports an evidentiary claim, not an automatic allegation of infringement. External quarantine should be proportionate, reviewable, logged, and reversible, with independent verification and an appeal process where affected users or organisations are identifiable.

13 Related Work

Text watermarking and calibration. Decoding-time watermarking modifies token selection using a secret keyed rule [1, 2, 3]. Recent work also studies model-free closed-form calibration and robustness under editing [16]. TRACE-LM uses such mechanisms only as physical carriers; its target is downstream source recovery after training.

Radioactivity and learnability. Radioactive data traces training influence in vision models [4]. For LLMs, watermarked generated text can leave a detectable trace after fine-tuning [5], and student models can learn decoding watermarks [6]. TRACE-LM treats this inheritance as an uncertain linear channel and designs source codes in the resulting downstream geometry.

IP protection, extraction fingerprints, and backdoors. Conditional, probability-vector, lexical, and backdoor signals have been proposed for language-model and embedding services [9, 7, 8, 10]. Backdoor-removal work demonstrates why a semantic trigger should not be equated with durable enforcement [17]. TRACE-LM therefore makes the active channel secondary and moves quarantine outside the model.

Fingerprinting codes and sparse tracing. Collusion-secure fingerprinting codes trace redistributed digital content [12, 13]. TRACE-LM differs by optimising code separation after a learned distillation operator and by decoding a source–epoch mixture in the effective dictionary.

Sequential evidence. E-values and test martingales provide validity under optional stopping [14, 15]. TRACE-LM applies conditional e-processes to predictable adaptive API queries and binds the transcript, calibration, and confidence-set inversion into a signed evidence object.

14 Conclusion

The updated approach no longer relies on claiming novelty for a green-list watermark, radioactivity, a trigger, or a renamed kill switch. TRACE-LM formulates model lineage as robust communication through an uncertain distillation channel. Customer and epoch codes, carrier geometry, null whitening, and collusion constraints are designed jointly to maximise certified post-transfer separation. Adaptive passive and active tests then issue a proof-carrying lineage certificate whose Type-I guarantee can be independently recomputed. This refactoring increases novelty while narrowing the claims to what the mathematics can support. The decisive next step is an empirical study with unseen student families, strong semantic sanitisation, strategic collusion, backdoor-removal attacks, and independent certificate verification.

A Proofs and Certificate Schema

A.1 Proof sketch of Proposition 1

Fix a true hypothesis h and an alternative h'. After whitening, let $\boldsymbol{\Delta}_{h,h'}$ be their effective mean difference. The nearest-mean decoder selects h' only if the accumulated noise projected along $\boldsymbol{\Delta}_{h,h'}/\|\boldsymbol{\Delta}_{h,h'}\|_2$ exceeds half the mean separation. Under Assumption 1, the standard Chernoff bound for a sum of n conditionally sub-Gaussian increments gives

$$\mathbb{P}(h \to h') \le \exp\left(-\frac{n\|\boldsymbol{\Delta}_{h,h'}\|_2^2}{8\sigma^2}\right).$$

By Eq. (13), $\|\boldsymbol{\Delta}_{h,h'}\|_2^2 \ge d_\star$ for every certified attack family and operator in its uncertainty set. A union bound over the $H-1$ false hypotheses yields Eq. (21); solving for n gives Eq. (22).

A.2 Proof of Proposition 2

Write $\boldsymbol{M} = \widehat{\boldsymbol{M}}_a + \boldsymbol{\Delta}$ with $\|\boldsymbol{\Delta}\|_{\mathrm{op}} \le \varepsilon_a$. By the reverse triangle inequality,

$$\left\|\widehat{\boldsymbol{\Sigma}}_a^{-1/2}\boldsymbol{U}^\top \boldsymbol{M}\boldsymbol{U}\boldsymbol{d}\right\|_2 \ge \left\|\widehat{\boldsymbol{\Sigma}}_a^{-1/2}\boldsymbol{U}^\top \widehat{\boldsymbol{M}}_a\boldsymbol{U}\boldsymbol{d}\right\|_2 - \left\|\widehat{\boldsymbol{\Sigma}}_a^{-1/2}\boldsymbol{U}^\top \boldsymbol{\Delta}\boldsymbol{U}\boldsymbol{d}\right\|_2.$$

Submultiplicativity, $\|\boldsymbol{U}\|_{\mathrm{op}} = 1$, and the uncertainty radius imply

$$\left\|\widehat{\boldsymbol{\Sigma}}_a^{-1/2}\boldsymbol{U}^\top \boldsymbol{\Delta}\boldsymbol{U}\boldsymbol{d}\right\|_2 \le \left\|\widehat{\boldsymbol{\Sigma}}_a^{-1/2}\boldsymbol{U}^\top\right\|_{\mathrm{op}} \varepsilon_a \|\boldsymbol{d}\|_2,$$

which gives Eq. (23).

A.3 Minimum certificate schema

Field	Required verification
Protocol identity	Version, hash of implementation, and declared significance level.
Suspect identity	Endpoint or model-artifact hash and observation interval.
Design commitment	Hashes of $\boldsymbol{U}$, $\boldsymbol{C}_E$, epoch seeds, and KL budget.
Calibration commitment	Null models, score normalisation, covariance estimate, and conditional cumulant bounds.
Transfer audit	Held-out proxy pipelines, $\widehat{\boldsymbol{M}}_a$, ε_a, and certified lower bound on $d_\star$.
Transcript	Ordered prompts, outputs or hashes, bounded scores, and every e-process update.
Attribution result	Source–epoch confidence set, coalition estimate, and stopping rule.
Active controls	Challenge/control randomisation, response scores, and cross-channel consistency.
Negative controls	Shadow-code, innocent-model, and corrupted-certificate outcomes.
Signatures	Investigator, independent verifier, and optional quorum signatures.

Table 2: Minimum fields for a proof-carrying lineage certificate.

References

[1] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. *A Watermark for Large Language Models.* International Conference on Machine Learning (ICML), 2023.

[2] R. Kuditipudi, J. Thickstun, T. Hashimoto, and P. Liang. *Robust Distortion-Free Watermarks for Language Models.* arXiv:2307.15593, 2023.

[3] M. Christ, S. Gunn, and O. Zamir. *Undetectable Watermarks for Language Models.* Cryptology ePrint Archive, 2023.

[4] A. Sablayrolles, M. Douze, C. Schmid, and H. Jégou. *Radioactive Data: Tracing Through Training.* International Conference on Machine Learning (ICML), 2020.

[5] T. Sander, P. Fernandez, A. Durmus, M. Douze, and T. Furon. *Watermarking Makes Language Models Radioactive.* Advances in Neural Information Processing Systems (NeurIPS), 2024.

[6] C. Gu, X. L. Li, P. Liang, and T. Hashimoto. *On the Learnability of Watermarks for Language Models.* International Conference on Learning Representations (ICLR), 2024.

[7] X. Zhao, Y.-X. Wang, and L. Li. *Protecting Language Generation Models via Invisible Watermarking.* International Conference on Machine Learning (ICML), 2023.

[8] W. Peng, J. Yi, F. Wu, S. Wu, B. Zhu, L. Lyu, B. Jiao, T. Xu, G. Sun, and X. Xie. *Are You Copying My Model? Protecting the Copyright of Large Language Models for EaaS via Backdoor Watermark.* Annual Meeting of the Association for Computational Linguistics (ACL), 2023.

[9] X. He, Q. Xu, Y. Zeng, L. Lyu, F. Wu, J. Li, and R. Jia. *CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarks.* Advances in Neural Information Processing Systems (NeurIPS), 2022.

[10] Y. Adi, C. Baum, M. Cissé, B. Pinkas, and J. Keshet. *Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring.* USENIX Security Symposium, 2018.

[11] H. Ma, T. Chen, T.-K. Hu, C. You, X. Xie, and Z. Wang. *Undistillable: Making a Nasty Teacher That CANNOT Teach Students.* International Conference on Learning Representations (ICLR), 2021.

[12] D. Boneh and J. Shaw. *Collusion-Secure Fingerprinting for Digital Data.* IEEE Transactions on Information Theory, 44(5):1897–1905, 1998.

[13] G. Tardos. *Optimal Probabilistic Fingerprint Codes.* Proceedings of the 35th Annual ACM Symposium on Theory of Computing, 2003.

[14] J. Ville. *Étude Critique de la Notion de Collectif.* Gauthier-Villars, 1939.

[15] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. *Time-Uniform, Nonparametric, Nonasymptotic Confidence Sequences.* The Annals of Statistics, 49(2):1055–1080, 2021.

[16] C. Li-Chen and K. Kim. *ChainMark: Model-Free LLM Watermarking with Closed-Form Calibration.* arXiv:2607.18445, 2026.

[17] Y. Li, Z. Zhang, K. Wang, X. Han, L. Shi, and H. Wang. *Defense Against LLM Backdoors using Critical Neuron Isolation Pruning.* arXiv:2607.19894, 2026.