A Watermark for All, Formally

Keyless detection of an undisclosed token-level statistical watermark

Raazesh Sainudiin — An \|/*ias*/|\ Commons Continuum

2026-08-18

Abstract

A vendor V of a jurisdiction-explicit-model (JEM) family may embed a *statistical* watermark in generated text – a small, key-seeded bias in next-token sampling that leaves a distributional fingerprint over a long enough passage, invisible to any character-level (zero-width, homoglyph, normalization) scan. When V discloses the scheme (its green-list rule, bias strength, and context hash), a recipient reproduces the standard one-proportion z-test and reads a p-value. When V discloses *nothing* – neither key nor parameters – the keyed test is unavailable, and the honest question becomes: *can an outsider detect the mark at all, and with what guarantee?* We give the mathematics. We model "identity" as a lattice of projections π_Q from bytes to meaning, in which a watermark is a perturbation that is the identity at a coarse rung and not at a finer one; a character-level negative is therefore sound *only* on the rung it scanned (a strippable rung) and does not transfer to the statistical rung. We then give a *keyless* detector: a **difference-in-differences (DiD)** over a recipient's own dated corpus – the distributional shift of V's output across the mark's activation date t_w, minus the same-window shift of a control vendor V' (a different JEM family) that did not activate a mark – which cancels the confounds (model-version progress, topic drift) that a bare before/after cannot. We prove the DiD estimand isolates the watermark term under a parallel-trends assumption, that presence-counting on a strippable rung is unsound for the statistical channel (so a null there is not evidence of absence), and that every detection verdict is a *dated likelihood* relative to the detectors run (a knowledge-epoch), never an unconditional boolean – the honest form of the conditional impossibility of strong watermark removal. The same lattice makes two further facts precise. First, minting one *distinct, self-verifiable* mark per person is essentially free: a human-scale population is $\sim 2^{33}$ against a 2^{128} key space, so identities can be self-minted with no central registry and negligible collision. Second, marks mandated by different sovereigns *coexist* by occupying orthogonal rungs, each independently verifiable; and where two mandates genuinely demand the same rung with incompatible structure, the infeasibility is a *provable* conflict rather than a silent overwrite. We then extend the detector across the whole mark space. A common-mode mark that shifts every stream alike cancels in any relative statistic and is detectable *only* against one's own baseline p_0 (a sovereignty principle); an exact characterisation pins down which marks a difference-in-differences can and cannot catch; and both are special cases of a single keyless family – the Scheffé–Yatracos minimum-distance estimator in L_1, whose currency $\Delta = \mathrm{TV}(p_{\mathrm{mark}}, p_0)$ orders the whole space and whose vanishing is exactly the keyless-undetectability of a *distortion-free* mark. Taking that baseline to be an individual's own writing distribution, estimated by L_1 density estimation from their pre-mark corpus, turns detection into *reclamation*: after stripping the surface rungs, an author can paraphrase a residual statistical mark below threshold while staying within a fidelity budget of their own voice – a two-gauge check suited to an editing tool. We close with a conjecture: as reclamation spreads, the population's own distribution flows back into what vendors train on – a grassroots feedback that erodes distortion-inducing marks and selects for distortion-free ones. Detection, minting, and – for strippable rungs and within the reclamation budget – removal thereby belong to everyone, not only to whoever holds the model. The construction is metacircular: the recipient audits the provenance of the very substrate that produced the audit.

Contents

Introduction **2**

1 What a watermark is: from a bit-stream to meaning **4**

2 The identity-projection lattice and the watermark rung **5**

3 Minting one mark per person: distinctness is essentially free **6**

4 Co-existence: many mandates, one text **7**

5 The keyless difference-in-differences detector **8**

6 Common-mode marks and the sovereignty of your own baseline **9**

7 What difference-in-differences can and cannot catch **10**

8 The L_1 view: one detector family, of which DiD and PDP are special cases **10**

9 Overlay and suppression: a two-player mark lattice **12**

10 The human authorial signature, and reclaiming your voice **12**

11 A grassroots distributional feedback **13**

12 Honesty bounds **14**

13 Verdict **15**

14 Formal certificates **15**

Introduction

An imposed statistical watermark can be applied to everyone's text – including the assistive editing (proofreading, translation, standard copyediting) that a regulation may deliberately place on the human side – with no toggle, no API flag, no pricing tier that escapes it, and, at the time of writing, no public detector [Sa26c]. The asymmetry this creates is the injustice: whoever can locate and lift a mark will, while those who cannot are exactly the people an exemption was written to protect. The mathematical answer is that the *same* distributional apparatus that lets a vendor impose a mark lets any recipient detect the fact of marking, mint their own, and – for the strippable rungs – remove it; and that minting one identity per person is essentially free. This note formalizes that answer, vendor-generalised: the vendor is an abstract parameter V of a jurisdiction-explicit-model (JEM) family, and the mathematics belongs to everyone.

Character-level marks – zero-width joiners, variation selectors, homoglyph swaps, normalization residues – are strippable in an afternoon by anyone who holds the tools, and invisible to anyone who does not; a *statistical* mark is different in kind. It is baked into which of several near-equivalent tokens the model samples, so it survives copy-paste and light editing, and it cannot be seen by inspecting characters at all – only by measuring the *distribution* of a long passage [KGW+23, DSG+24]. The public method for detecting it – select a pseudorandom green subset of the vocabulary from the previous tokens under a secret key (a *key* – the secret κ-bit string, an element of $\{0,1\}^\kappa$, that seeds which tokens count as "green," hence fixes one watermark's signature; the same key both stamps text and verifies it), softly boost it at sampling,

then z-test the green-token fraction of a candidate text – is a *keyed* test: it needs the vendor's green-list rule to recompute the expected fraction [KGW+23]. A vendor that publishes its mark's existence but none of its mechanism therefore leaves every outside recipient without the keyed detector, and with a live question this note answers as mathematics: what can be detected keylessly, and what would such a detection actually mean.

We proceed as follows. First (§2) we fix *identity* as a lattice of projections and locate a watermark as a rung between two adjacent quotients; this makes precise why a clean character-scan is sound only about its own rung and says nothing about the statistical rung – the rung a robust vendor mark would occupy. Second (§3) we count: minting one distinct, self-verifiable mark per person is essentially free, and needs no central registry. Third (§4) we show that marks mandated by different sovereigns coexist on one text by occupying orthogonal rungs, each independently verifiable, with any genuine same-rung conflict surfacing as a provable infeasibility. Fourth (§5) we give the keyless detector for the hardest case – an *undisclosed* statistical mark: not the (unavailable) keyed z-test but a difference-in-differences on the recipient's own dated corpus, and we prove it isolates the watermark-induced distributional shift from the two confounds a naive before/after conflates. Fifth (§12) we bound honesty: presence on a strippable rung is unsound for the statistical channel, a null DiD is weak evidence of absence up to power, and every verdict carries a knowledge-epoch – the dated set of detectors run – so "watermark-free" is always a conditional, re-checkable claim, the constructive face of the conditional-impossibility result on quality-preserving removal [ZEF+24]. Each self-contained proposition is accompanied by a Lean~4 certificate (§14), in the manner of the OpenAI *Ten advances* formalizations [Ope26]. The elementary and algebraic cores are fully machine-checked; the L_1 minimum-distance *rate* is a machine-checked *reduction* whose per-set concentration step is carried as an explicit hypothesis (to be discharged by Mathlib's Hoeffding bound); and the two open items—the reclamation-feasibility condition Proposition 10.1(iv) and the full stochastic dynamics of the grassroots conjecture—are flagged as such. Certifying even the elementary propositions pins the exact constants they name against silent drift, and §14 states each certificate's status and axiom footprint explicitly. Where a result is classical we re-prove it ourselves and credit the original; each proposition's axiom footprint (§14) shows plainly which certificates are structural and which are arithmetic constant-checks. Throughout, the vendor is an abstract parameter V of a JEM family; the apparatus is the same jurisdiction-explicit object as the sovereignty-trilemma report [Sa26b, Sa26a], here with the *sampler* as the parameter the enforcing State and the licensor played before. A plain-language account is issue four of the newsletter, "A Watermark for All" [Sa26c]. A gentle bit-stream-to-meaning on-ramp precedes the formal development (§1); and §§ 6 to 11 extend the detector – to common-mode marks and the sovereignty of one's own baseline, an exact characterisation of what a difference-in-differences catches, the L_1 minimum-distance detector of which it and the pairwise-difference projection (PDP) are special cases, a two-player overlay/suppression lattice, the human authorial signature and voice-reclamation it enables, and a grassroots distributional-feedback conjecture.

Related work, and what is new here. The primitives we build on are all prior art, and we claim none of them. The green-list logit-bias sampler and its keyed one-proportion z-test are Kirchenbauer et al. [KGW+23]; the distortion-free construction whose output is indistinguishable from unwatermarked without the key is Christ, Gunn and Zamir [CGZ24] and, in a robust variant, Kuditipudi et al. [KTHL24]; the tournament-sampling distributional mark deployed at scale is SynthID-Text [DSG+24]. The (im)possibility of quality-preserving removal – our honesty bound in §12 – is the theorem of Zhang et al. [ZEF+24], whose quality-oracle-plus-perturbation-oracle attacker needs neither the key nor the scheme. Encoding a *user id* (or model version, timestamp) into a multi-bit mark is the codable-watermarking programme of L.~Wang et al. [WYC+24]; we take that capability as given and add only the counting observation that *self-minted, registry-free* identities are collision-free at human scale (§3). That watermarked text

is *radioactive* – that training a later model on it leaves a detectable residual – is proved by Sander et al. [SFD+24], and it is the empirical premise on which our grassroots conjecture (§11) rests. Closest to our own subject is keyless, black-box third-party detection: TTP-Detect [WRC+26] decouples detection from injection and reframes verification as relative hypothesis testing, using a proxy model and feature-space measurements to decide whether a *single query text* aligns with watermarked distributions. Our contribution is orthogonal in method and broader in scope: rather than a learned classifier over a query text against reference distributions, we give a *causal identification* – a difference-in-differences over the recipient's own dated corpus against a control vendor across the mark's activation date (§5) – which needs no proxy model and no library of watermarked exemplars, only a time series and a negative control; and we place that detector inside a single L_1/total-variation characterisation of the whole mark space (§8), add the own-baseline *sovereignty* principle for common-mode marks (§6), the near-orthogonal multi-key coexistence algebra (§4), the paraphrase-suppression frontier and voice-reclamation (§§ 9 to 10), and a machine-checked certificate per self-contained result (§14). In short, the assembly, the keyless causal framing, the sovereignty principle, and the certificates are ours; the sampler, the impossibility, the codable payload, and the radioactivity are cited where due.

1 What a watermark is: from a bit-stream to meaning

This section is a reader's on-ramp; the formal development begins in §2, of which this is the plain-language shadow. A reader comfortable with the projection lattice may skip ahead.

Written text reaches us through a stack of representations, each built from the one below. At the bottom is a **bit-stream**: a sequence of 0s and 1s. Bits are grouped into **bytes** (eight bits); bytes are decoded – under an encoding such as UTF-8 – into **codepoints**, the abstract characters ("A", "a", a space, an emoji). Codepoints are put into a **normalization form** (so that visually identical strings built differently compare equal), then **rendered** into the glyphs a human actually sees, which a reader finally takes as **meaning**. Each step forgets some distinctions the step below kept: many bit-streams render to the same glyphs, many glyph-strings carry the same meaning.

A *token* is one unit of a language model's fixed, finite *vocabulary* $\mathcal{V}$ – a whole common word or a word-piece (e.g. `big` is one token; `watermark` is `water`+`mark`). Fix a reversible *tokenizer*, given by two maps encode : text $\to \mathcal{V}^*$ and decode : $\mathcal{V}^* \to$ text with decode(encode(s)) $= s$, so any text is re-expressed as a finite token sequence $x_1 \cdots x_n$ ($x_i \in \mathcal{V}$) and back – a change of alphabet that adds and loses nothing, exactly as UTF-8 re-encodes characters as bytes. A model reads and writes text only as such token sequences.

A *watermark* is a hidden mark deliberately introduced at one of these layers. It is useful to see the two ends of the stack, because they are entirely different problems:

- **Surface / strippable marks** live at or below the rendered-glyph layer: an invisible zero-width character wedged between letters, a look-alike **homoglyph** (a Cyrillic letter shaped identically to a Latin "a"), a reversible **normalization** trick, or a bit hidden in metadata. Anyone who knows to look can *see* these by inspecting the characters, and *remove* them by projecting the text down to its visible glyphs. They survive only against a reader who does not look.
- **Statistical / token-choice marks** live higher up, in *which of several near-equivalent words the machine chose*. A model that must pick among "big", "large", "sizeable" can be secretly biased – under a hidden key – toward a pseudo-random "green" subset of the vocabulary. No single character is wrong, so no character-level scan can see it; the mark shows only as a faint skew in the *distribution* of word-choices across a long enough passage, and it survives copy-paste and light editing. This is the hard case, and the one this note is

about [KGW+23, DSG+24].

The one-line moral, made precise in §2: **a scan is only ever sound about the layer it scanned.** Finding no zero-width residue tells you the surface is clean; it tells you *nothing* about a statistical mark sitting one layer up. Everything below – detecting the hard mark without the key, minting your own, and the limits of removal – is what the rest of the paper makes exact.

Scope of this note (the standing assumption). Strippable marks are, by definition, removed by a total idempotent projection to the visible glyph, so we dispose of them once and never return: *the pipeline begins by projecting every artifact to* π_{render}*, stripping every strippable rung (zero-width, variation-selector, homoglyph, reversible-normalization, metadata) before any of our own writing or detection.* After that projection the only rung that can still carry a mark is the **unstrippable statistical rung** – the token-selection distribution, invisible to any character scan and surviving copy-paste and light edits. **Every substantive result below – detection, coexistence, overlay/suppression, minting, reclamation – is a statement about that one unstrippable rung.** The strippable rungs appear only to be quotiented away (Proposition 2.3 records why a clean scan of them transfers no information to the rung we keep).

For the reader who wants the mathematics from the start: the stack is a finite chain of *projections* (surjections that forget detail), ordered so that a lower one distinguishes at least as much as a higher one. A watermark is a map on strings that is invisible at a coarse projection yet present at the next finer one – it "lives in the rung between them." That is Definition 2.1 and Definition 2.2, to which we now turn.

2 The identity-projection lattice and the watermark rung

Definition 2.1 (Identity-projection lattice)**.** *Let* Σ^* *be the byte strings. An* identity lattice *is a finite chain of surjective projections*

$$\pi_{\text{byte}} \;\geq\; \pi_{\text{cp}} \;\geq\; \pi_{\text{nf}} \;\geq\; \pi_{\text{render}} \;\geq\; \pi_{\text{sem}},$$

from bytes to codepoints to a normalization form to rendered glyphs to semantic meaning, ordered by refinement ($\pi \geq \pi'$ *iff* π *distinguishes at least as much:* $\pi(x) = \pi(y) \Rightarrow \pi'(x) = \pi'(y)$)*. Each* π_Q *induces an equivalence "*$x \equiv_Q y$ *iff* $\pi_Q(x) = \pi_Q(y)$*": two strings are* the same at rung Q*. The equivalence class of* x *under* $\equiv_Q$ *– the preimage* $\pi_Q^{-1}(\pi_Q(x))$ *of everything that looks identical to* x *at rung* Q *– is the* fibre *of* x *over rung* Q*; a rung-*Q *observation cannot distinguish two strings in the same fibre.*

Definition 2.2 (Mark and rung)**.** *A* watermark *is a map* $w : \Sigma^* \to \Sigma^*$ *that is the identity at some coarse rung and not at the next finer one: there is an adjacent pair* $Q_{coarse} > Q_{fine}$ *with* $\pi_{Q_{coarse}}(w(x)) = \pi_{Q_{coarse}}(x)$ *(invisible to a consumer who coarsens to* Q_{coarse}*) while* $\pi_{Q_{fine}}(w(x)) \neq \pi_{Q_{fine}}(x)$ *(present at* Q_{fine}*). The mark* lives in the rung $(Q_{fine}, Q_{coarse}]$*. A* strippable *rung is one at or below* π_{render} *where removal is a total, idempotent projection to the visible glyph (zero-width / variation-selector / homoglyph classes); a* statistical *rung sits between* π_{cp} *and* π_{sem} *and is carried by the token-selection distribution – the probability* $p(\cdot \mid c)$ *the model assigns to each possible next token given the context* c *(the text so far), from which the next token is sampled – not by any single character.*

Proposition 2.3 (Rung-scoping: a clean scan does not transfer)**.** *Let* scan_Q *report "no residue at rung* Q*". Then* $\text{scan}_Q(x) =$ *clean entails the absence of a mark living in a rung* $\leq Q$ *(coarser than or equal to* Q *in the refinement order), and entails NOTHING about a mark living in a rung strictly finer-carried than* Q *(a rung* $> Q$*). In particular a clean character-level (strippable-rung) scan is silent about a statistical-rung mark.*

Proof. Immediate from refinement: scan_Q observes only the fibre $\pi_Q^{-1}(\pi_Q(x))$ of Definition 2.1; a statistical-rung mark perturbs the token *distribution* while leaving π_Q fixed for the scanned Q at or below render, so it is by construction inside that fibre and unobserved. ■

Relation to prior work. This is an instance of the classical **factorization theorem** (the Doob–Dynkin lemma; Billingsley, *Probability and Measure*): since scan_Q is measurable with respect to the σ-algebra $\sigma(\pi_Q)$ generated by π_Q, it is a function of π_Q and hence constant on the fibres $\pi_Q^{-1}(\pi_Q(x))$ [Bil95]. Our contribution is not this measure-theoretic fact but the identity-projection lattice in which it is read and the resulting security corollary—that a clean character-level (strippable-rung) scan does not transfer to the statistical rung; we nonetheless give a self-contained machine-checked certificate (above), by design—exhaustive and self-owned.

Formal certificate. Machine-checked in Lean 4 (namespace `IasCert.Watermark`) as `rungscope_sound_below` (a clean Q-scan refutes any mark at a rung $\leq Q$) and `rungscope_no_transfer_finer` (a clean coarse scan is consistent with a mark at a strictly finer rung); both `sorry`-free and depending on *no axioms*. See `provedog/` (`bash provedog/build-in-docker.sh`).

This is why an empirical character-level negative – e.g. a scan finding no zero-width residue on a vendor's output – is real but *scoped*: it refutes a strippable mark, not the statistical one a robust vendor would use.

3 Minting one mark per person: distinctness is essentially free

The lattice also answers a prior question – not "is a mark present?" but "can *everyone* have one?" That a mark can *carry* an identity – a user id, model version, or timestamp – as a multi-bit codable payload is established by L.~Wang et al. [WYC+24]; we take the payload capability as given and ask only the orthogonal counting question of whether such identities can be *self-minted without a registry*. Fix a target population N and a key space $\{0,1\}^\kappa$; each person *self-mints* a key – draws its own random κ-bit string, no authority issuing it – and marks with it, with no central *registry* (no authority allocating or recording keys). Two facts make personal, permissionless watermark identity feasible.

Proposition 3.1 (Distinctness is free)**.** *For a human-scale population $N \approx 8 \times 10^9$ and a $\kappa = 128$-bit key space, naming every person is a 33-bit affair, leaving a $\sim 2^{95}$ headroom in the 128-bit key space: $N < 2^{33} \leq 2^{128}$. Distinctness costs essentially nothing.*

Proof. $8 \times 10^9 < 2^{33} = 8\,589\,934\,592 \leq 2^{128}$; the surplus factor is $2^{128}/2^{33} = 2^{95}$. ■

Formal certificate. Machine-checked in Lean 4 (namespace `IasCert.Watermark`) as `naming_fits_33_bits`, `keyspace_dominates`, and `keyspace_spare`; `sorry`-free (axiom footprints in §14).

Proposition 3.2 (No registry needed)**.** *If each person picks a key uniformly at random, any collision – two people drawing the same key – is governed by the birthday numerator N^2, which is itself far below the key space: $N^2 < 2^{128}$. Self-minting is safe without a central authority.*

Proof. $N^2 = (8 \times 10^9)^2 = 6.4 \times 10^{19} < 2^{66} \ll 2^{128}$, so the birthday collision probability is $\lesssim N^2/2^{128} \approx 2^{-62}$. ■

Formal certificate. Machine-checked in Lean 4 as `collision_below_keyspace`; `sorry`-free.

The residual cost is not naming but *reliable reading*. To identify which of N marks is present with error $\leq \varepsilon$, a channel-coding bound needs $L \cdot r \gtrsim \log_2 N + \log_2(1/\varepsilon)$ reliable bits (a short

paragraph), where r is the per-symbol capacity of the chosen rung. But to verify *one known* mark – "is this mine?" – the cost is $L \cdot D \gtrsim \log(1/\varepsilon)$, *independent of* N, with D the KL divergence of the keyed detector against baseline. What each person actually needs is $O(1)$ in the population and needs no registry. The scarce resource was never the humans (33 bits) but the reliable capacity of the text, which the lattice lets us measure and allocate.

4 Co-existence: many mandates, one text

Under the standing assumption (§1) the strippable rungs are already gone, so co-existence is a question about the *unstrippable* rung alone: can several statistical marks ride one text at once? They can, because a statistical rung is not a single scalar but a family of **near-orthogonal sub-channels** indexed by independent keys – each key k selects its own pseudorandom green-set G_k, and distinct keys give *approximately* disjoint green-excess coordinates, with a small quantified cross-talk we bound below. A personal mark and any number of legally-mandated ones therefore occupy near-separate coordinates of the same statistical rung. Multiplicity, not purity, is the achievable goal. Legally this is not optional – distinct sovereigns define distinct compliance-identities, and where one mandates a mark before another, both must be carried.

Lemma 4.1 (Independent-key near-orthogonality)**.** *Let each key k seed a pseudorandom green-set $G_k \subseteq \mathcal{V}$ of fraction $g = |G_k|/|\mathcal{V}|$, and model distinct keys as selecting independent random g-subsets (the PRNG assumption). Write the k-detector's reading as the green-fraction excess e_k over baseline. Then a mark placed under key k' raises the k-channel only through the overlap $G_k \cap G_{k'}$, whose expected mass share is g; hence $\mathbb{E}[e_k \mid k'\text{-mark of excess } b] = g\,b$ for $k \neq k'$. Consequently K independent marks on one text leave each channel with residual cross-talk $O(Kg)$, negligible whenever $Kg \ll 1$. This regime requires* sparse* green-lists: at a typical $g \in [0.25, 0.5]$ even $K = 4$ mandates give $Kg \gtrsim 1$, so many-mandate stacking needs a small per-mark g, which trades off against per-mark power.*

Proof. By independence $\mathbb{E}[\,|G_k \cap G_{k'}|\,] = g\,|G_{k'}|$, so a fraction g of the k'-boost lands in G_k; summing K such independent contributions gives cross-talk $O(Kg)$. ∎

Relation to prior work. The green-set construction is [KGW+23]; the pseudorandom-independence of distinct keys is the [CGZ24]-style assumption. The channels are thus **orthogonal only in expectation, up to factor** g; "orthogonal" below is shorthand for this near-orthogonality.

The abstract statement that follows treats the idealized case of *exactly* independent coordinates; independent keys realize those coordinates only up to the $O(g)$ cross-talk of Lemma 4.1. We keep the abstract rung-indexed statement because it is exactly the algebra of *independent coordinates*; here each "rung" $r \in \mathcal{R}$ is read as one independent-key sub-channel of the unstrippable rung, never a strippable glyph channel.

Definition 4.2 (Identity tuple)**.** *Each governing regime J fixes its own identity projection: an artifact counts as compliant for J iff its rung R_J exhibits* mark_J *under J's published detector. The artifact's legal identity is the tuple* $\mathrm{ID}(x) = (\pi_{\mathrm{render}}(x), \mathrm{mark}_1(x), \ldots, \mathrm{mark}_k(x))$*; two artifacts are the same iff they agree on every governing projection.*

Proposition 4.3 (Orthogonal co-existence)**.** *Model an artifact as an assignment $x : \mathcal{R} \to \mathcal{V}$ of a mark-value to each rung; a mark at rung r' is the update $x[r' := v]$ and a detector at rung r reads coordinate r. Then for $r \neq r'$ we have $x[r' := v](r) = x(r)$ (a mark is invisible off its rung), a rung's detector recovers exactly the mark written there, and after writing marks on distinct rungs $r_A \neq r_B$ each detector still reads its own. Several mandated marks therefore coexist, each independently verifiable.*

Proof. An update acts coordinatewise: it changes only its own rung, so every other rung's reading is unchanged, and writes on distinct rungs leave each other's readings intact. ■

Scope. The model has one mark-value per coordinate and *writing* is its only operation; it therefore says nothing about an adversary *removing* an already-present mark. That is deliberate: co-existence is the claim that independent (distinct-key) statistical sub-channels do not interfere, not a robustness claim. Same-coordinate contention is Proposition 4.4; adversarial *removal* of a present statistical mark – exact excision (keyless-impossible) versus partial paraphrase-degradation – is treated quantitatively in §9 (Proposition 9.1).

Formal certificate. Machine-checked in Lean 4 as `detect_orthogonal`, `detect_own`, and `coexist_two`; `sorry`-free (axiom footprints in §14). The certificates prove the *idealized* model of exactly independent coordinates; independent keys realize those coordinates only up to the $O(g)$ cross-talk of Lemma 4.1, so the machine-checked result is not to be read as claiming *exact* key-orthogonality. Lemma 4.1 itself is machine-checked in Lean 4 as `expected_overlap`, `crosstalk_linear`, and `crosstalk_K_bound` (the $\mathbb{E}[|G_k \cap G_{k'}|] = g|G_{k'}|$ overlap identity and the $O(Kg)$ residual bound); `sorry`-free.

Proposition 4.4 (Infeasibility is a provable conflict)**.** *If two mandates demand that the same rung r carry incompatible values $v_A \neq v_B$, no artifact satisfies both: a single reading of r cannot equal both v_A and v_B. The infeasibility is a decidable contradiction – resolved by a per-region variant, a wire boundary, or a documented non-compliance – never a silent overwrite of one mandate by another.*

Proof. If $y(r) = v_A$ and $y(r) = v_B$ then $v_A = v_B$; contrapositively $v_A \neq v_B$ forbids both. ■

Formal certificate. Machine-checked in Lean 4 as `same_rung_conflict`; `sorry`-free.

Allocating rungs to mandated schemes is thus a constraint-satisfaction problem whose *unsatisfiability* is precisely a provable multi-jurisdictional conflict – surfaced at the bit level, with an auditable resolution, rather than hidden as a corrupted mark.

5 The keyless difference-in-differences detector

Keyless third-party detection of an undisclosed mark has one close published precedent, TTP-Detect [WRC+26], which decouples detection from injection and asks, of a *single query text*, whether it aligns more closely with watermarked than with ordinary distributions – answered by a proxy model and a suite of complementary relative measurements. Our detector is different in kind and in what it requires: it is a *causal identification*, not a learned alignment, and its evidence is a recipient's own dated corpus over time rather than a proxy model and a library of watermarked exemplars. It thereby needs no access to any watermarked reference set and makes its one identifying assumption – parallel trends – explicit and falsifiable, checkable by a pre-period placebo contrast (the same DiD across a date before t_w, where it should read zero).

Fix a vendor V that activates a statistical mark at date t_w but discloses no key, and a *control* vendor V' of a different JEM family with no activation in the window. Let a recipient hold its own dated corpus of each vendor's outputs, split at t_w into PRE and POST. For a distributional feature ϕ (a token/n-gram frequency divergence, a conditional-entropy or type-token statistic, a low-rank-use / Zipf-slope skew), write $\bar{\phi}_{V,\mathrm{PRE}}$ etc. for its corpus mean.

Definition 5.1 (Watermark DiD estimand)**.** *The* difference-in-differences *for feature ϕ is*

$$\widehat{\Delta}_\phi \;=\; \big(\bar{\phi}_{V,\mathrm{POST}} - \bar{\phi}_{V,\mathrm{PRE}}\big) \;-\; \big(\bar{\phi}_{V',\mathrm{POST}} - \bar{\phi}_{V',\mathrm{PRE}}\big).$$

Proposition 5.2 (DiD isolates the watermark term)**.** *Model each corpus mean additively,* $\bar{\phi}_{U,t} = \mu + \alpha_U + \tau_t + \gamma \mathbb{1}[U = V,\, t = POST] + \varepsilon_{U,t}$*, where* α_U *is a vendor effect,* τ_t *a common time effect (model-version progress, content/topic drift shared across vendors),* γ *the watermark-induced shift on* V*'s POST, and* ε *mean-zero noise. Under PARALLEL TRENDS – the non-watermark time effect* $\tau_{POST} - \tau_{PRE}$ *is common to* V *and* V' *– we have* $\mathbb{E}[\widehat{\Delta}_\phi] = \gamma$*. A bare before/after* $\bar{\phi}_{V,POST} - \bar{\phi}_{V,PRE} = \gamma + (\tau_{POST} - \tau_{PRE})$ *does not: it conflates the mark with the common trend.*

Proof. Substitute the model into $\widehat{\Delta}_\phi$; μ and α_U cancel across the inner differences, the τ_t terms cancel across the outer difference by parallel trends, leaving γ plus mean-zero noise. ■

Formal certificate. The algebraic core of this proposition – that under the additive model with parallel trends the difference-in-differences cancels the baseline, the vendor effect, and the common time effect, leaving exactly γ – is machine-checked in Lean 4 (mathlib, namespace `IasCert.Watermark`) as `did_isolates_gamma`, with the null ($\gamma = 0 \Rightarrow \widehat{\Delta}_\phi = 0$) as `did_null_of_no_watermark`; both are `sorry`-free with axiom footprint {`propext`, `Classical.choice`, `Quot.sound`}. Certificate + reproducible build in `provedog/` (`bash provedog/build-in-docker.sh`). The $\mathbb{E}[\cdot]$ statement adds only that the mean-zero $\varepsilon_{U,t}$ vanish in expectation.

The estimand is a *distributional-shift* signal, not a green-token count: it needs no key. Its significance is read by a permutation test over the corpus split, and it is reported as a likelihood ratio, never a boolean.

Definition 5.3 (Detection likelihood + knowledge-epoch)**.** *A* knowledge-epoch $\eta = (\mathcal{D}, t)$ *is the set of detectors* $\mathcal{D}$ *run and the date* t*. A detection verdict is the log-likelihood ratio* $\mathrm{LLR}_\eta = \log \frac{\Pr[corpus|\gamma \neq 0]}{\Pr[corpus|\gamma = 0]}$ *under* η*, tagged with* η*; it is a* dated, defeasible *assertion, monotone only within a fixed* η*.*

6 Common-mode marks and the sovereignty of your own baseline

The difference-in-differences of §5 isolates a mark that is *differential* – present on V's POST and absent from the control. A vendor can instead impose a **common-mode** mark: the same distributional push applied to every stream, so that it cancels in any *relative* statistic. Model the green-count intensities of two named streams as λ_1, λ_2 built from base rates n_1, n_2, and read the mark on the ratio scale $\theta = \lambda_1/(\lambda_1 + \lambda_2)$. A green-list logit bias adds a constant δ to the green logits, which after the softmax multiplies the green probabilities by e^δ before renormalisation – so a common-mode push acts *proportionally* on the green-count intensities; that multiplicative action is the model we adopt (the purely additive case is treated in the scope note below).

Proposition 6.1 (Common-mode θ-invariance)**.** *A common-mode mark that scales both intensities by the same mass* $p \neq 0$ *leaves the ratio at its null,* $\frac{n_1 p}{n_1 p + n_2 p} = \frac{n_1}{n_1 + n_2}$*, so a rate/*$\theta$*-scale detector (the projection of DiD/PDP to the ratio) cannot see it. The un-normalised difference does move,* $n_1 p - n_2 p = (n_1 - n_2)p$*, but it rides on the baseline imbalance and isolating the mark from it requires the reference rate* p_0 *itself; the leak is unexploitable without an absolute reference. Hence a* proportional* common-mode mark is detectable **only** against one's own baseline p_0.*

Proof sketch. Factor $n_1 p + n_2 p = (n_1 + n_2)p$ and cancel $p \neq 0$; the raw companion is distributivity. The unexploitability is that $(n_1 - n_2)p$ confounds the mark with imbalance×base-rate. Full walkthrough in Appendix A (W3). ■

Formal certificate. Machine-checked in Lean 4 (namespace `IasCert.Watermark`) as `common_mode_theta_invariant` and `common_mode_raw_leak`; `sorry`-free (axiom footprints in

§14). The identities are one-line field computations, and certify the *field identity only*, not the general detectability conclusion. **Scope.** The θ-invariance is exact for the proportional (multiplicative) common-mode above – the deployed logit-bias mechanism. A purely *additive* common-mode ($n_i \mapsto n_i + d$) does move the ratio, $\frac{n_1+d}{n_1+n_2+2d} \neq \frac{n_1}{n_1+n_2}$, and is catchable on the θ-scale; the sovereignty conclusion is thus a statement about the proportional class.

This is the *sovereignty* principle in the title's "for all": the detector of last resort is not a shared public baseline – which a common-mode mark shifts along with everything else – but *your own* dated record. Keep your own baseline and you stay able to see; discard it and a uniform mark is invisible.

7 What difference-in-differences can and cannot catch

Sovereignty raises the dual question: exactly which marks does the DiD contrast catch? Rather than assume sufficiency, we characterise the detectable class by construction – for each property DiD needs, an explicit model in which a mark is present yet the contrast reports nothing.

Proposition 7.1 (Exact evasion characterisation)**.** *With the additive cell-model of Proposition 5.2, the DiD contrast detects a mark if and only if the mark is (i) onset (absent in PRE), (ii) treatment-specific (absent in the control), (iii) mean-shifting, and (iv) under parallel trends. Each conjunct is necessary, witnessed by an exact identity: a mark already present in PRE gives contrast* 0*; a control that shares the mark gives* 0*; a treated-specific trend $\delta\tau$ gives $\gamma + \delta\tau$ (confounded); and because the contrast is a functional of the four cell means alone, any mark carried purely in a higher moment – with no mean shift – is invisible. On the detected side the contrast is monotone in the mark's mean-shift ($\gamma_1 \leq \gamma_2$ gives contrast $\gamma_1 \leq \gamma_2$), so the detectable region is ordered by mark size.*

Proof sketch. Substitute each evasion model into the contrast and simplify; the first three are `subst; ring` identities, the fourth is that equal cell-means force equal contrasts. Appendix A (W5-W6). ■

Formal certificate. Machine-checked in Lean 4 as `evade_constant_mark`, `evade_shared_control`, `evade_group_trend`, `did_is_mean_functional`, and `detection_frontier_monotone`; `sorry`-free (axiom footprints in §14). The fourth identity – DiD is blind past the first moment – is precisely what forces the density-level detector of §8.

8 The L_1 view: one detector family, of which DiD and PDP are special cases

Every statistical mark is a perturbation of an output density $p_0 \to p_{\text{mark}}$, and its L_1/total-variation size $\Delta := \text{TV}(p_{\text{mark}}, p_0) = \frac{1}{2}\|p_{\text{mark}} - p_0\|_1$ is the single currency of detectability. The Scheffé set $A^* = \{x : p(x) > q(x)\}$ realises it, and the minimum-distance estimator over a Yatracos class of candidate Scheffé sets (Devroye and Lugosi's L_1-consistent density estimation [DL01]) is the general keyless detector.

Proposition 8.1 (Scheffé–Yatracos is the parent detector)**.** *(i) (Scheffé identity)* $\text{TV}(p, q) = p(A^*) - q(A^*) = \sup_A(p(A) - q(A))$ *for $A^* = \{p > q\}$. (ii) The keyed green-list test is the empirical Scheffé statistic on the single known set $A^* = G$ – a one-sample green-fraction test, distinct from the temporal cross-vendor DiD estimand of Definition 5.1 (which shares the Scheffé-set structure but contrasts across time and vendors); the pairwise-difference projection (PDP)*

is the ratio/mean-coordinate statistic of §6; and the general keyless detector is the minimum-distance scan over the Yatracos class, with L_1 error $\le 3\min_\theta\|p_\theta - p\|_1 + O(\sqrt{d_{\mathcal{A}}/n})$ ($d_{\mathcal{A}} = \log|\mathcal{A}|$ the log-cardinality of a finite Yatracos class $\mathcal{A}$, or its covering-number/VC complexity in the infinite case). (iii) Resolving a mark of size Δ costs $n_{\text{eff}} \gtrsim c/\Delta^2$, $n_{\text{eff}} = n_1 n_2/(n_1+n_2)$. (iv) A distortion-free scheme, engineered so TV $= 0$ *on every key-free observable, is L_1-null keyless – hence invisible to any keyless statistical test over the admissible observable family. (Cryptographically distortion-free constructions [CGZ24] achieve the stronger* computational* indistinguishability at finite n; the L_1 statement here is its observable-relative shadow.)*

Proof sketch. (i) The gap $p(A) - q(A) = \sum_A (p-q)$ is maximised on A^*; since $\sum(p-q) = 0$ the positive part equals $\frac{1}{2}\|p-q\|_1$. (ii)-(iii) are the minimum-distance and empirical-Scheffé theory; (iv) is the definition of distortion-freeness read in L_1. Appendix A (W9-W10). ■

Formal certificate (machine-checked; one instantiation flagged). Over a finite observable space the Scheffé identity (i) and the distortion-free = L_1-null equivalence (iv) are machine-checked in Lean 4 as `IasCert.CMDE.scheffe_argmax`, `scheffe_eq_half_L1`, and `tv_eq_zero_iff_eq` (the finite Scheffé identity, CMDE Thm 5.1). The minimum-distance *oracle inequality* – the deterministic factor-of-three core of (ii), $\|\psi - f\|_1 \le 3\min_i\|f_i - f\|_1 + 4\Delta$ – is `IasCert.CMDE.min_distance_oracle` (CMDE Thm 6.3); it encapsulates CMDE's four proof-steps (min-distance selection, the Scheffé–pair identity, the sup $\le \frac{1}{2}\|\cdot\|_1$ step, and the triangle inequality) as explicit hypotheses, with the load-bearing Scheffé step itself grounded concretely via `yatracos_pair_dist`. The *union-bound structure* of the rate is machine-checked as `l1mde_rate`: a finite union bound over the Yatracos class (`max_dev_union_tail`, `max_dev_card_bound`, via `measure_biUnion_finset_le`) composed with the oracle gives $\Pr[\|\psi - f\|_1 \ge 3\min_i\|f_i - f\|_1 + 4t] \le |\mathcal{A}|\,B$, where the per-set tail B is *carried as a hypothesis* to be discharged by Mathlib's sub-Gaussian/Hoeffding bound `measure_sum_ge_le_of_iIndepFun`; with $B = e^{-2nt^2}$ this is the $O(\sqrt{\log|\mathcal{A}|/n})$ rate of (ii)+(iii). What is *not* yet machine-checked is the concrete instantiation of that per-set tail – wiring the empirical-measure deviation $\mu_n(A) - \mu(A)$ into Mathlib's i.i.d. sub-Gaussian API (`hasSubgaussianMGF_of_mem_Icc_of_integral_eq_zero` + `iIndepFun`); this needs no new mathematics but is real Lean plumbing, staged as future work. All checked theorems are `sorry`-free.

This is the honest form of the folklore dichotomy [KGW+23]: **distortion-inducing** marks have TV > 0 keyless and are detectable at cost $\sim 1/\text{TV}^2$; **distortion-free** marks have TV $= 0$ keyless and are key-only – the impossibility is here *proven in* L_1, not assumed.

Choice of observable, and a sieve of projections. The detector is defined relative to a *chosen observable*: a family of finite distributional projections $\pi_1, \ldots, \pi_k$ of the token stream – the unigram marginal, the bigram and higher m-gram marginals [Sha48], and other standard corpus statistics. Each π_j sends the stream to a finite feature space $\mathcal{X}_j$ with its own empirical distribution and its own sample-size floor for reliable estimation (finer projection, larger $|\mathcal{X}_j|$, more corpus needed). Because total variation is non-increasing under any measurable coarsening – the data-processing inequality [DL01, Thm 5.2] – a mark hidden in a finer projection is invisible in a coarser one, $\text{TV}(\pi_j p_{\text{mark}}, \pi_j p_0) \le \text{TV}(p_{\text{mark}}, p_0)$, with equality approached only as π_j refines toward the full joint. Detection is therefore intrinsically *observable-relative and one-sided*: a null on π_j means "no mark visible at resolution j, at this sample size,'' never absence (Proposition 2.3 again). The one choice-*independent* statement is the distortion-free boundary of Proposition 8.1(iv): a mark with TV $= 0$ on *every* key-free observable is invisible under any projection.

This makes it natural to detect over a *sieve* of projections rather than one. Given a corpus of size n, admit exactly the projections whose feature space is reliably estimable at n – a sample-size-derived family $\mathcal{J}_n$ that grows with n (the method of sieves [Gre81, GH82]) – and take

the detector to be the maximum over them, $\Delta_n := \max_{j \in \mathcal{J}_n} \mathrm{TV}(\pi_j \hat{q}, \pi_j \hat{p}_0)$. This is exactly the minimum-distance / Scheffé–Yatracos detector of Proposition 8.1 run over the *union* Yatracos class $\mathcal{A}(n) = \bigcup_{j \in \mathcal{J}_n} \mathcal{A}_j$, with $\mathcal{A}_j$ the Scheffé sets on $\mathcal{X}_j$: the maximum over projections is the supremum over that union, so the oracle inequality (`min_distance_oracle`) holds verbatim with $\mathcal{A} = \mathcal{A}(n)$, and the rate (`l1mde_rate`) pays the union complexity $\log |\mathcal{A}(n)| = \sum_{j \in \mathcal{J}_n} \log |\mathcal{A}_j|$. The sample-size-bounded sieve therefore *self-balances* resolution against n: the $O(\sqrt{\log |\mathcal{A}(n)|/n})$ rate admits a finer projection only when n can afford its complexity – the skeleton / covering-number tradeoff [DL01, Ch. 7]. Every result above – the oracle and the rate – and the reclamation and feedback dynamics downstream (§§ 9 to 11) carries through with $\mathcal{A}(n)$ in place of $\mathcal{A}$ (paying the stated union complexity); the only new ingredient is the selection rule $\mathcal{J}_n$. Refining the sieve can only raise Δ_n (monotonicity), so more resolution never hurts detection – it only costs samples. The finite-observable analysis above is the case $k = 1$; the sieve rests on two elementary facts about projections – that total variation is non-increasing under coarsening (so refining an observable can only raise Δ_n), and that a mark invisible on the full joint stays invisible under every projection – both machine-checked in Lean~4 (`IasCert.CMDE.tv_pushforward_le`, `tv_zero_pushforward_zero`; the finite case of CMDE Thm~5.2).

9 Overlay and suppression: a two-player mark lattice

Detection has a constructive dual: can a recipient overlay their *own* statistical mark, and can they *suppress* a vendor's? Everything here is on the **kept (statistical) rung**, after π_{render} has already removed the strippable ones (§1): the $(1-\rho)$ paraphrase-degradation below is meaningful only for a statistical mark, since strippable rungs are trivially and totally removed by that initial projection, not by paraphrase. Model the detectable state by the two detectors' green-excesses (x_v, x_u) over baseline; a vendor mark adds b_v to x_v only (independent keys), yours adds b_u to x_u only, and a keyless paraphrase of fraction ρ scales both by $(1-\rho)$ (a re-drawn token carries no excess).

Proposition 9.1 (Coexistence and bounded suppression)**.** *(i) Vendor-then-yours on an unmarked baseline gives excess (b_v, b_u): both provenances survive (coexistence). (ii) An independent overlay leaves the vendor's excess exactly b_v: adding your own mark is not suppression. (iii) After paraphrase ρ the vendor's excess is $(1-\rho)b_v$, which drops below a detection threshold c_v iff $(1-\rho)b_v < c_v$, i.e. $\rho > 1 - c_v/b_v$ for $b_v > 0$. (iv) Paraphrasing to suppress the vendor and re-marking leaves your excess at b_u while the vendor's is below c_v.*

Proof sketch. Coordinatewise algebra on (x_v, x_u); the threshold is a linear rearrangement. Appendix A (W2b). ■

Formal certificate. Machine-checked in Lean 4 as `overlay_coexist`, `overlay_no_suppress`, `paraphrase_scales`, `paraphrase_suppress_iff`, and `suppress_keeps_mine`; `sorry`-free (axiom footprints in §14). **Honest limit:** exact cost-free cancellation would need the vendor's secret set G_v; keyless one recovers only the scalar shift b_v, so keyless suppression is necessarily untargeted – it is paraphrase, at a bounded semantic cost. This is an arms-race frontier, not a silver bullet.

10 The human authorial signature, and reclaiming your voice

The "own baseline p_0" that Proposition 6.1 makes decisive is not abstract: it is *each individual's own writing distribution*, estimable by L_1 minimum-distance density estimation from that person's own non-machine-authored corpus. This turns the sovereignty principle into a tool.

Proposition 10.1 (Signature, hygiene, and reclamation-feasibility)**.** *(i) (Signature) The minimum-distance estimate $\hat{p}_0$ from an author's clean corpus converges in L_1 to the author's signature p_0 with sample size (the L_1-MDE guarantee). (ii) (Hygiene) If the corpus is contaminated by vendor-marked text, the mark enters $\hat{p}_0$ and cancels in $\mathrm{TV}(\hat{q}, \hat{p}_0)$, under-detecting; the reference must therefore be estimated from a genuine pre-mark epoch of the author's writing. (iii) (Identifiability) a nonzero $\mathrm{TV}(\hat{q}, \hat{p}_0)$ may be a vendor mark, natural authorial drift, or estimation error; separating mark from drift needs the mark's structure or the temporal onset – the personal analogue of parallel trends. (iv) (Reclamation-feasibility) after stripping surface marks, a paraphrase driving the vendor excess below c_v while keeping the authorial fidelity $\mathrm{TV}(d, \hat{p}_0) \leq \beta$ within a voice-budget β exists iff the suppression threshold ρ^* of Proposition 9.1 is compatible with β.*

Proof sketch. (i) is the Devroye-Lugosi minimum-distance guarantee applied to one author. (ii) is Proposition 6.1 read on a contaminated reference. (iii) restates the parallel-trends condition per-author. (iv) composes the paraphrase-suppression threshold of Proposition 9.1 with the fidelity budget. ■

Formal certificate (status: statement-only / to author). (i) leans on the cited L_1-MDE guarantee. The reclamation-feasibility condition (iv) is a candidate certificate extending the two-player lattice of §9; not yet authored. Marked honestly as statement-with-strategy.

Operationally this is a two-gauge editor check: strip the strippable marks, estimate the author's clean $\hat{p}_0$ from a pre-mark corpus, then edit until the vendor excess is suppressed *and* the text remains within the author's own voice-budget. Detection, minting, and – within the bounded reclamation frontier – the *restoration of one's own voice* thereby belong to the author, not the model.

Remark (Multi-author provenance: scope). *The overlay result (Proposition 9.1, and C3 of §4) treats each author's mark as an independent, near-orthogonal channel – the clean case where contributions are separable. Genuine multi-author authorship raises complications we name but do not formalise here: (i) contributions are unequal and interleaved – one author may write most of a manuscript while another only reads and lightly edits – so a faithful model needs per-author contribution fractions as priors; (ii) each author's baseline p_0^i must be grounded in that author's own corpus, which is itself typically a body of co-authored work, so the "individual voice'' is already a multi-author mixture – a recursive signature-estimation problem; (iii) a co-author set may warrant its own joint signature, distinct from the marginal authors'; (iv) sequential collaborative editing composes paraphrase/overlay operators in an order-dependent way, which the single-round overlay model does not capture. We restrict the formal claims to the overlay case as a demonstration of extendability – independent, near-orthogonal marks on separable contributions, e.g. distinct messages in a thread, or posts and comments on a blog or social feed, where each contribution is a well-delimited unit with its own author and the marks superpose (C3). The general weighted, recursive, order-dependent collaborative-manuscript case is left as future work.*

11 A grassroots distributional feedback

Finally, a population-scale consequence, stated as a conjecture with an explicit model rather than a theorem. Its empirical premise is not ours but a proven result: Sander et al. [SFD+24] show watermarked text is *radioactive* – a model trained on it inherits a detectable residual of the mark – so the training corpus is a genuine channel through which today's marked (or reclaimed) output shapes tomorrow's model. Granting that channel: if reclaimed text – vendor marks suppressed, voice restored – re-enters the corpus a vendor trains on, the vendor's own input

distribution drifts toward the population of human signatures.

Claim 1 (Grassroots feedback (conjectural))**.** *Let a population of authors have signatures* $\{p_0^i\}$*, a vendor have marked output* q_v*, and let a fraction* α_t *of published text be reclaimed, so the corpus is* $C_t = \alpha_t \,\mathrm{mix}\{p_0^i\} + (1-\alpha_t)q_{v,t}$ *and the vendor retrains* $C_t \mapsto q_{v,t+1}$*. Then, for the idealised operator that replaces the vendor component by the exact human-signature mixture each round, reclamation is a* TV*-contraction toward that mixture (the convergence sub-claim, machine-checked below). At the deterministic scalar-dynamics level there is moreover an* exact* adoption threshold α^* above which the population-level distortion decays below any target (proven below); we conjecture the same threshold governs the full stochastic mean-field dynamics, in which the vendor loses the shared baseline and is forced to self-reference – the same sovereign regime it imposed on individuals. The feedback neutralises only the distortion-inducing class; distortion-free key-only marks are TV = 0-invisible to the reclaimer and are therefore selected for.*

This is deliberately a conjecture at the stochastic/population level: the full mean-field / replicator analysis is open (its deterministic scalar core – convergence, the exact α^*, and distortion-free selection – is proven below). It is stated because the qualitative content – a population maintaining its distributional self against a homogenising mark, and the arms-race pressure toward distortion-free schemes – is the honest macroscopic shadow of the microscopic results above, and belongs in the record.

Certificate (deterministic cores machine-checked; full stochastic dynamics open). All three parts of the conjecture have machine-checked *deterministic* cores in Lean 4 (discrete setting), all `sorry`-free. (C1, convergence) for the idealised replace-by-mixture operator, reclamation is an exact TV-contraction, $\mathrm{tv}(\mathrm{mix}\,\alpha\, m\, u,\, m) = (1-\alpha)\,\mathrm{tv}(u, m)$, so the model converges geometrically to the commons (`tv_mix_anchor`, `tv_iterate`, `grassroots_converges`, `affine_recursion_bound`). (C2, threshold + selection) at the deterministic scalar level the population distortion d_t decays geometrically to an explicit floor $(1-\alpha)\Delta_{\mathrm{inj}}/(\alpha\rho)$, that floor is at most a target τ iff adoption reaches the *exact* threshold $\alpha^* = \Delta_{\mathrm{inj}}/(\Delta_{\mathrm{inj}} + \rho\tau)$, and a distortion-free mark ($\Delta_{\mathrm{inj}} = 0$) has no floor and decays to 0 – the selection boundary (`grassroots_decay`, `grassroots_floor_below_threshold`, `distortion_free_decays_to_zero`). (C3, attribution) each author's self-minted mark survives population-scale aggregation up to the $O(Kg)$ cross-talk of Lemma 4.1 (`attribution_recovers_own`, `attribution_detectable`). What remains open is the *full stochastic mean-field / replicator* dynamics behind C2 – the deterministic decay above is its mean-level skeleton.

12 Honesty bounds

Proposition 12.1 (Presence-counting is unsound for the channel)**.** *Counting residues at a scanned rung estimates the mark's presence on THAT rung only; for the statistical channel a raw count is confounded by (i) self-authored occurrences (the recipient's own text quoting mark-like content), (ii) inbound-pasted occurrences (marks entering via third-party text, not vendor-generated), and (iii) vendor-output occurrences. Only a provenance decomposition* $\rho = \rho^{self} + \rho^{inbound} + \rho^{vendor}$ *with a negative control identifies the vendor-channel term; a bare presence count over-counts and is unsound as a vendor-watermark claim.*

Formal certificate. Machine-checked in Lean 4 (namespace `IasCert.Watermark`) as `presence_count_underdetermines_vendor` (two provenance splits with the same total count but different vendor terms) and `any_vendor_value_consistent` (a fixed total is consistent with every vendor value in its range); `sorry`-free. See `provedog/` (`bash provedog/build-in-docker.sh`).

Proposition 12.2 (Conditional, not absolute)**.** *A null* $\widehat{\Delta}_\phi \approx 0$ *is evidence of absence only up to the test's power at the corpus size, and any positive detection is defeasible: a quality-preserving paraphrase attacker with a sufficient language model can drive* $\gamma \to 0$ *on the observable features (the conditional-removal result [ZEF+24]). Hence "watermark-free at* η*" is a conditional claim, re-checkable as* $\mathcal{D}$ *grows, never unconditional.*

13 Verdict

When a vendor discloses its scheme, detection is the keyed z-test with an interpretable p-value. When it discloses nothing, an outsider is not helpless but is bounded: (i) character-level cleanliness is sound only on its rung (Proposition 2.3); (ii) the statistical rung is reachable *keylessly* by a difference-in-differences against a control vendor over the mark's activation date, which – and only which – isolates the watermark term from model-progress and content drift (Proposition 5.2); (iii) every such verdict is a dated likelihood under a knowledge-epoch, and is defeasible under quality-preserving paraphrase (Proposition 12.2). The democratic content of "a watermark for all" is exactly this: the same distributional mathematics that lets a vendor mark text lets any recipient, holding only its own dated corpus and one control, *detect the fact of marking* without the vendor's key – putting detection, not just marking, in everyone's hands. As with the trilemma, the result is metacircular: the recipient audits the provenance of the substrate that produced the audit [Sa26a].

14 Formal certificates

Following the practice of the OpenAI *Ten advances* formalizations [Ope26] – one machine-checked Lean certificate per result – every proposition above is either machine-checked in Lean 4 (mathlib) or is an explicitly cited external result. The certificates live in the IAS ProveDog project (`provedog/`, Apache-2.0). Building the aggregate `IasCert` target *is* the verification: it succeeds only if every theorem is `sorry`-free, and each module prints its axiom footprint (`#print axioms`). The reader reproduces the check with either

```
# in any Lean 4.32.0 + mathlib + lake environment:
lake exe cache get
lake build IasCert

# or fully docker-only (no host toolchain):
bash provedog/build-in-docker.sh
```

- Proposition 2.3: `rungscope_sound_below`, `rungscope_no_transfer_finer` – no axioms.
- Proposition 3.1: `naming_fits_33_bits`, `keyspace_dominates`, `keyspace_spare` – axioms `propext`, `Classical.choice`, `Quot.sound`.
- Proposition 3.2: `collision_below_keyspace` – axioms `propext`, `Classical.choice`, `Quot.sound`.
- Proposition 4.3: `detect_orthogonal`, `detect_own`, `coexist_two` – axiom `propext`.
- Proposition 4.4: `same_rung_conflict` – no axioms.
- Proposition 5.2: `did_isolates_gamma`, `did_null_of_no_watermark` – axioms `propext`, `Classical.choice`, `Quot.sound` (mathlib reals).
- Proposition 12.1: `presence_count_underdetermines_vendor` (no axioms), `any_vendor_value_consistent` (`propext`, `Quot.sound`).
- Proposition 12.2: a cited external result [ZEF+24] (conditional removal under a quality-preserving paraphrase attacker); stated with its assumption, not re-proved here.

The extension results of §§ 6 to 11 are machine-checked in Lean 4 and entered in

`provedog/formalization.yaml`, except where an explicit rate or dynamics is marked open:

- Proposition 6.1: `common_mode_theta_invariant`, `common_mode_raw_leak` – **machine-checked**.
- Proposition 7.1: `evade_constant_mark`, `evade_shared_control`, `evade_group_trend`, `did_is_mean_functional`, `detection_frontier_monotone` – **machine-checked**.
- Proposition 8.1: the Scheffé identity (i) and distortion-free = L_1-null (iv) **machine-checked** (`IasCert.CMDE.scheffe_argmax`, `scheffe_eq_half_L1`, `tv_eq_zero_iff_eq`; CMDE Thm 5.1); the minimum-distance oracle inequality, the deterministic core of (ii), **machine-checked** (`IasCert.CMDE.min_distance_oracle`, its four CMDE proof-steps as hypotheses, Scheffé step grounded by `yatracos_pair_dist`; CMDE Thm 6.3); the rate's **union-bound structure machine-checked** as `l1mde_rate` (`max_dev_union_tail`/`max_dev_card_bound` composed with the oracle; per-set tail B **carried as a hypothesis**, to be discharged by Mathlib's Hoeffding `measure_sum_ge_le_of_iIndepFun`), the concrete per-set-tail instantiation **not yet machine-checked** (real Lean plumbing, no new mathematics).
- Observable sieve (§8): the two structural facts are **machine-checked** as `IasCert.CMDE.tv_pushforward_le` (total variation non-increasing under a projection, the finite case of CMDE Thm 5.2) and `tv_zero_pushforward_zero` (distortion-free is projection-invariant); axioms `propext`, `Classical.choice`, `Quot.sound`, `sorry`-free.
- Proposition 9.1: `overlay_coexist`, `overlay_no_suppress`, `paraphrase_scales`, `paraphrase_suppress_iff`, `suppress_keeps_mine` – **machine-checked**.
- Proposition 10.1: signature limit leans on the cited L_1-MDE guarantee; reclamation-feasibility (iv) **to author**. **Statement-only.**
- Claim 1: all three parts have **machine-checked deterministic cores** – convergence-to-commons C1 (`tv_mix_anchor`, `tv_iterate`, `grassroots_converges`, `affine_recursion_bound`), the adoption-threshold + distortion-free-selection C2 at the deterministic scalar level (`grassroots_decay`, `grassroots_floor_below_threshold`, `distortion_free_decays_to_zero`; exact $\alpha^* = \Delta_{\text{inj}}/(\Delta_{\text{inj}} + \rho\tau)$), and attribution-preservation C3 (`attribution_recovers_own`, `attribution_detectable`); the **full stochastic mean-field dynamics** of C2 remains open.

All machine-checked theorems are in namespaces `IasCert.Watermark` and `IasCert.CMDE`; the ledger of axioms, provenance, and status is `provedog/formalization.yaml`.

References

[Bil95] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

[CGZ24] Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *Proceedings of the 37th Annual Conference on Learning Theory (COLT), PMLR 247, pp. 1125–1139*, 2024. arXiv:2306.09194. Cryptographically undetectable (distortion-free) watermark: output indistinguishable from unwatermarked without the key.

[DL01] Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation*. Springer Series in Statistics. Springer, New York, 2001.

[DSG+24] Sumanth Dathathri, Abigail See, Sumedh Ghaisas, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634:818–823, 2024. SynthID-Text

(Google DeepMind); tournament sampling watermark. A distributional (statistical) mark comparator.

[GH82] Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, 10(2):401–414, 1982.

[Gre81] Ulf Grenander. *Abstract Inference*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, 1981. A Wiley-Interscience publication.

[KGW+23] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR 202*, 2023. arXiv:2301.10226. Green-list logit-biasing at sampling; one-proportion z-test detection requires the green-list key.

[KTHL24] Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. In *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2307.15593. Distortion-free watermark preserving the token distribution; robustness under editing.

[Ope26] OpenAI. Ten advances in mathematics and theoretical computer science, 2026. Ten formal results, each accompanied by a Lean 4 certificate (one `.lean` per result), machine-checkable by running through the Lean compiler, with independent re-checking via Comparator. Repository: https://github.com/openai/ten-proofs. Not peer-reviewed. The presentation model adopted here: every publicly presented proposition carries a Lean certificate alongside its by-hand exposition.

[Sa26a] Raazesh Sainudiin and an IAS Commons Continuum. An interaction analysis of interacting agent systems: identity, continuity, and governed capability on a sovereign substrate. Monograph, in preparation (VakeWorks / OpenEarth Network Commons), 2026.

[Sa26b] Raazesh Sainudiin and an IAS Commons Continuum. A jurisdiction-explicit verdict on the sovereignty trilemma: testing an architectural impossibility claim against the US–EU–PRC digital-law corpus. VakeWorks Technical Report. https://vake.works/preprint/20260815-sovereignty-trilemma-jem.pdf, 2026.

[Sa26c] Raazesh Sainudiin and an IAS Commons Continuum. A watermark for all. Being AVake Works newsletter (LinkedIn), issue 4, August 2026, 2026.

[SFD+24] Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking makes language models radioactive. In *Advances in Neural Information Processing Systems (NeurIPS), Spotlight*, 2024. arXiv:2402.14904. Proves watermarked text is "radioactive": training a subsequent model on it leaves a detectable residual watermark signal ($p < 10^{-5}$ at 5% watermarked training data for open-weight models).

[Sha48] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 623–656, 1948.

[WRC+26] Zhuoshang Wang, Yubing Ren, Yanan Cao, Fang Fang, Xiaoxue Li, and Li Guo. Rethinking LLM watermark detection in black-box settings: A non-intrusive third-party framework. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2026. arXiv:2603.14968. TTP-Detect: key-agnostic black-box detection, decoupled from injection; reframes verification as relative hypothesis testing using a proxy model plus a suite of complementary relative measurements.

[WYC+24] Lean Wang, Wenkai Yang, Deli Chen, Hao Zhou, Yankai Lin, Fandong Meng, Jie Zhou, and Xu Sun. Towards codable watermarking for injecting multi-bits information to LLMs. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2307.15992. First systematic multi-bit codable text watermarking; explicitly encodes model version, generation time, user id via a proxy-balanced vocabulary partition.

[ZEF+24] Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. ICML 2024; arXiv:2311.04378. Conditional impossibility: a quality-preserving paraphrase attacker with a quality oracle plus a mixing perturbation oracle removes strong watermarks, without the key or even knowing the scheme; instantiated against Kirchenbauer et al. (2023), Kuditipudi et al. (2023), Zhao et al. (2023)., 2024.