SELECTIVE WATERMARKING: SECURE MULTI-TASK ATTRIBUTION OF LANGUAGE MODEL OUTPUT

Anonymous authors
Paper under double-blind review

ABSTRACT

Watermarking for language models answers a binary question – *was this text machine-generated?* – yet a single model now serves many pipelines and tenants, where the operative questions are *which task produced a text* and whether auditing one task leaks information about another. We introduce *Selective Watermarking*, which upgrades detection to secure multi-class attribution: distinct cryptographic keys and heterogeneous embedding channels are routed to disjoint compartments of a single document, and a fail-closed rule attributes a text to a task key or abstains. Our system, SAFIZ, composes an undetectable cryptographic in-sampling core with a zoned lexical channel and a global, invertible orthographic channel, fusing calibrated per-channel p-values by Fisher's method. The guarantee is exact: under the pseudorandomness of the keyed function, every compartment's detection score follows a $\mathrm{Gamma}(\ell, 1)$ law under any non-matching key, giving a family-wise false-attribution rate of at most α with no distributional approximation and no fusion weights to tune. The post-hoc channels supply evidence precisely where no undetectable scheme can embed – short, low-entropy outputs – at the declared cost of system-level undetectability. Across six models, three tasks and $8{,}100$ documents, SAFIZ attributes 99.2% with zero misattributions, and $872{,}310$ null trials stay within the bounds (0.050% family-wise, 0.053% per key). The cryptographic core leaves text quality untouched: under a third-party model, paired perplexity differences between watermarked and unwatermarked generations is balanced in sign on every model. A controlled base/instruct pair further shows instruction tuning doubling the share of documents on which the undetectable core lacks entropy to embed. Under signal-removal attacks, the system degrades to abstention rather than error: soundness and isolation hold even where recoverability is lost.

1 INTRODUCTION

Identifying and tracing machine-generated text is central to the responsible deployment of language models and, under regulation such as the EU AI Act, is becoming a legal obligation for providers; watermarking is the principal technical response, an imperceptible signal embedded at generation and certifiable by a verifier holding a secret key. Yet essentially all existing schemes answer a single *binary* question – *was this text produced by a machine?* – while one model increasingly serves many pipelines and many tenants under one provider. Consider a provider whose model serves three tasks: a public code assistant, a code-review service that patches user-supplied code, and an authorised security-research sandbox. A working piece of malware surfaces online and the watermark confirms that the provider generated it; the binary verdict ends there, and it is the wrong question, because what matters is *for which purpose* the text was generated: produced by the public assistant, it is a policy breach in production; by the review service, a patched variant of code the user brought in; by the sandbox, an expected research artefact. The operative questions are thus *which task produced a given text* and whether a verifier holding one task's key can gather any evidence, spurious or genuine, from the output of another; a watermark keyed once for the whole service answers neither.

We close this gap with *Selective Watermarking*, which upgrades detection from a binary test into a secure multi-class attribution problem: given a text, the system outputs a task label from a predefined set or a fail-closed NONE. By "secure" we mean *soundness* (a bounded family-wise false-positive rate), *isolation* (a non-matching key recovers only noise), and *fail-closed* operation (abstention rather than misattribution); robustness to signal removal is deliberately not among the guarantees and is

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

characterised, not claimed. A *task* is an operational label – the serving pipeline or tenant contract to which the provider bound a key at generation – not a semantic category to be recognised: the candidate set is closed and defined by the provider's routing. We realise the paradigm by routing distinct, key-separated watermarks to disjoint compartments of a single document. Its defining property is *key-selectivity*: a task's signature is recoverable only with its key, while under any other key the evidence is indistinguishable from unmarked text. Our instantiation, SAFIZ (Selective Watermark Attribution by Fisher-fused Isolated Zoning), embeds evidence through two complementary carriers: an undetectable cryptographic channel bounded by text entropy, and an entropy-independent orthographic channel, effective everywhere but fragile by design. We also retain a zoned lexical channel for its architectural role; the guarantees hold regardless of which channel carries the evidence.

Our contributions are: (i) the **Selective Watermarking paradigm**, multi-task attribution by key separation over disjoint compartments formalised with cryptographic security definitions (Definitions 1–2), to our knowledge the first such formalisation with family-wise soundness and abstention as an explicit outcome; (ii) **calibrated multi-channel fusion with an exact guarantee**: heterogeneous channels are combined by Fisher's method into one calibrated p-value whose per-compartment null is $\mathrm{Gamma}(\ell, 1)$ under the pseudorandomness of the keyed function, with no distributional approximation and no fusion weights to tune (Theorem 1); (iii) **linear-time, model-free verification**, $O(Kn)$ over K keys: the compartment layout is logged at generation, so the verifier re-derives all evidence from the suspect text in one pass, with no model call and none of the super-linear search over layouts that a detector without the log requires; (iv) **an empirical study** across six models and $8{,}100$ documents (Section 4), to our knowledge the largest evaluation of multi-key watermark attribution to date and the first to verify both the family-wise and the per-key false-positive bounds, over $872{,}310$ null trials: 99.2% attribution with zero misattributions, false-accusation rates of 0.050% family-wise and 0.053% per key, the false-detection inflation of uncorrected multi-key scoring reproduced and removed by construction, ablations against a single-channel baseline in the style of Christ et al. (2024), and a characterisation under truncation, token edits, Unicode normalisation and paraphrase in which failures are abstentions, not misattributions.

2 Background and Related Work

Two families, and the cryptographic core. Watermarks divide into *distribution-modifying* schemes, which perturb sampling (the green/red logit bias of Kirchenbauer et al. (2023)), and *distortion-preserving* ones, which act on the randomness of sampling without changing the distribution (Kuditipudi et al., 2023; Hu et al., 2024). The strongest notion in the second family is the cryptographically *undetectable* watermark of Christ et al. (2024), the reference *zero-bit* construction (the watermark carries no message; detection tests presence, not content): without the key, watermarked and unwatermarked output are computationally indistinguishable, so quality is provably preserved. It rests on a *pseudo-random function* (PRF), indistinguishable from random without the key, and on *empirical entropy*, which bounds how much pseudo-random signal a text can hide (Section 3.3). Undetectability is a *binary* notion; *which* key marked a text is a separate question, the subject of this paper. Later constructions build on it (Cohen et al., 2025), and SAFIZ takes it as its point of departure, lifting it from a single-key binary test to multi-task attribution under key separation.

Surface-form watermarks and the low-entropy challenge. Post-hoc methods edit the surface form instead: Unicode homoglyph substitution (Rizzo et al., 2016; 2019) replaces ASCII characters with visually identical code points, leaving the rendered text unchanged, and lexical methods encode a signal through controlled synonym choice; SAFIZ uses keyed, deterministic versions of both, the orthographic channel being the entropy-independent counterpart of the cryptographic one. This matters because in-generation watermarks depend on entropy: low-entropy output (terse answers, summaries, code) sharply limits what can be embedded, and dedicated schemes therefore watermark only the high-entropy tokens (Lee et al., 2024); SAFIZ's orthographic channel instead supplies evidence exactly where the core falls silent, which our ablations show is decisive on short outputs.

Robustness and the adversary. Much work hardens watermarks against edits: Zhao et al. (2023) give provable robustness to bounded perturbations, and Kuditipudi et al. (2023) make a distortion-free watermark robust to token-level edits by aligning the text to a key-derived sequence under an edit-tolerant distance. These schemes target robust binary detection of a single watermark, an axis orthogonal to attribution. We build on the undetectable core instead because its detection statistic

has a null in closed form, which is what multi-key attribution needs (Section 3.4) – in any case an edit-robust core would not remove the need for an entropy-independent channel, since robust detectors also fail where the evidence they align to was never embedded.

Beyond binary: multi-bit and multi-user watermarks. Closest to our setting, Cohen et al. (2025) trace output to users or colluding groups by embedding a fingerprinting codeword as an L-bit payload on top of an undetectable zero-bit scheme; independently, Fu & Russell (2025) formalise the *false detection problem*: max-score thresholding over K keys without correction inflates the false-positive rate as K grows, and most sharply when the text is short. Their remedy reserves part of the tokens for a shared, key-independent detection watermark, so that detection no longer maximises over keys; SAFIZ instead keeps every statistic key-specific and corrects the threshold, since a shared detection signal is readable by anyone holding it and would forfeit isolation (Section 4.3). The two regimes have a quantitative boundary. The payload must be carried by the generated text: in the multi-user setting L is the length of a fingerprinting code, growing with the collusion bound and the logarithm of the population, and the construction of Cohen et al. (2025) extracts it from $O(L\lambda)$ watermarkable blocks, each closing only when the model has spent enough entropy. The $\log K$ cost thus lands on the generation side, as a budget of high-entropy text that short outputs cannot supply, whereas SAFIZ pays it on the detector side, where the Bonferroni threshold demands per-key evidence growing like $\ln(K/\alpha)$ that any calibrated channel may supply, including the orthographic channel on a 50-token answer. Zero-bit key separation also keeps isolation cryptographic rather than a decodable field in the text, and lets classes be added or revoked with a key. Our price is detection cost linear in the candidate set, $O(Kn)$, against $2L$ calls to the underlying detector in their construction, and no collusion resistance. Key selection with Bonferroni correction therefore dominates when K is small (a fixed set of tasks) and the scheme must work even on short or entropy-starved outputs; a payload-and-code construction dominates when K is large and outputs are reliably long enough to amortise the codeword.

Ensembles, semantics, and the wider landscape. Ensembles combine per-feature p-values for a single, *binary* provenance decision (Niess & Kern, 2025); SAFIZ differs in purpose, fusing channels *per candidate key* into a multi-class, fail-closed verdict with exact family-wise control. Semantic watermarks move the signal to a level that survives paraphrase (Liu et al., 2024a; Hou et al., 2024); SAFIZ's evidence lives at the token, character, and lexical levels, and semantic channels are a complementary direction. Post-hoc detectors identify machine text from statistical artefacts with no planted signal (Mitchell et al., 2023; Gehrmann et al., 2019; Tang et al., 2024) but offer no cryptographic guarantee, while watermarking is deployed at scale (Dathathri et al., 2024) with toolkits and surveys mapping the area (Pan et al., 2024; Liu et al., 2024b; Liang et al., 2024); SAFIZ's novelty in this landscape is *secure task-selective attribution*, not detection robustness.

3 METHOD

SAFIZ (Selective Watermark Attribution by Fisher-fused Isolated Zoning) routes heterogeneous watermarking techniques and distinct cryptographic keys to non-overlapping compartments of a single document: each task is bound to its own key and its signature confined to its own compartments, so that attribution becomes selective and cross-task leakage is provably controlled.

3.1 NOTATION, REGISTRY, AND THE INFORMED-AUDIT SETTING

A *document* is an n-character string $X \in \Sigma^*$ over the Unicode alphabet, with $X[c_s{:}c_e]$ the substring at offsets $[c_s, c_e)$. A fixed tokenizer and a binary encoding of the vocabulary put embedding and detection on three coupled levels: characters, tokens and bits. When X was produced by the system, $(X, \mathcal{R}(X)) = \mathsf{Wat}_{K_i}(\pi)$ for some prompt π; the detector treats X identically either way. Embedding partitions X into B contiguous *compartments* $s_1, \ldots, s_B$ with boundaries $0 = c_0 < \cdots < c_B = n$, each holding a *role* in $\{\texttt{crypto}, \texttt{syn}, \texttt{plain}\}$ (sampling-time watermark, keyed lexical channel, or neither), while the orthographic channel is global, that is a key-dependent *invertible* homoglyph map over the whole document, undone by normalisation before any other channel is scored. Each of the K tasks holds an independent key $K_i \in \{0,1\}^\lambda$, λ being the security parameter (the key length in bits), and $F : \{0,1\}^\lambda \times \{0,1\}^* \to (0,1)$ is a pseudorandom function (HMAC-SHA256 mapped to the unit interval): we write $F_{K_i}(b, \mathrm{seed}_b, j)$ for its evaluation on the compartment index b, the compartment seed, and the bit index j, packed at fixed width into one binary string. Including b

guarantees that two compartments which happen to share a seed still receive different PRF values (*domain separation*, Remark 2). $\alpha \in (0, 1)$ is the family-wise target false-positive rate.

SAFIZ operates in an *informed-audit* regime: the generating provider, which holds the task keys, logs the structural skeleton of every document in a *registry* $\mathcal{R}(X)$ recording where each compartment lies, its role, and which seed it uses. The registry contains neither keys nor any of the bits the detector scores; it only says where to look and how to index the PRF, while every scored bit is re-derived from the suspect text alone, with no call to the model, so the verdict is a function of the text. With the registry, verification is one linear pass, $O(n)$ per key and $O(Kn)$ over the candidate set; without it, SAFIZ operates in the *registry-free mode*, where the core remains detectable in the blind mode of Christ et al. (2024), at super-linear cost (up to $O(n^3)$) and without the zone structure the lexical channel needs (Remark 5). Registries are authenticated (a MAC under a provider key), since an adversary free to *choose* the registry could search layouts until a p-value clears the threshold; corruption of an authentic registry only degrades evidence toward abstention, never toward a wrong key (Remark 4).

3.2 Syntax and Security Definitions

Definition 1 (Selective watermarking scheme)**.** A selective watermarking scheme for K classes is a triple $\mathsf{Setup}(1^\lambda, K) \to \mathcal{K} = \{K_1, \ldots, K_K\}$, $\mathsf{Wat}_{K_i}(\pi) \to (X, \mathcal{R}(X))$, $\mathsf{Verify}(X, \mathcal{R}, \mathcal{K}) \to \hat{k} \in \mathcal{K} \cup \{\varnothing\}$: Setup samples K independent keys, Wat_{K_i} generates a document for prompt π under K_i with its registry, and Verify attributes the document to a key or abstains ($\varnothing$). The null hypothesis $H_0^{(i)}$ states that X was not generated under K_i.

Definition 2 (Security notions)**.** Let $\epsilon = \mathrm{negl}(\lambda)$. *(a) Soundness.* The scheme is (α, ϵ)-*sound* if for every document X and registry $\mathcal{R}$ fixed independently of the key set $\mathcal{K}$ (any human-written or differently-keyed text), $\Pr\left[\mathsf{Verify}(X, \mathcal{R}, \mathcal{K}) \neq \varnothing\right] \leq \alpha + \epsilon$ over $\mathsf{Setup}(1^\lambda, K)$: the probability that *any* of the K keys is falsely attributed is bounded. *(b) Isolation.* The scheme is ϵ-*isolating* if for every pair $i \neq j$ and every efficient distinguisher with access to documents generated under K_i, the evidence that Verify accumulates for K_j on those documents is computationally indistinguishable from the evidence accumulated on unwatermarked text: no efficient distinguisher separates the two better than guessing, up to ϵ (i.e., cross-task evidence is noise). *(c) Fail-closed.* Verify is *fail-closed* if it outputs some $\hat{k} \in \mathcal{K}$ only when the corresponding combined p-value clears the family-wise threshold (Equation 4) and $\varnothing$ otherwise (insufficient evidence, corrupted registry, conflicting evidence), so that evidence degradation can cause missed attribution, never forced misattribution.

Adversary model. The adversary probes a document under keys other than the one that produced it, seeking a misattribution or cross-task leakage; it holds no secret key and cannot forge the authenticated registry. Signal-*removal* attacks (truncation, character noise, paraphrase) are outside the security claim and are treated as a robustness *characterisation* (Section 4.5): where an edit destroys evidence the rule degrades to abstention, so soundness and isolation hold even when recoverability is lost.

3.3 Generation in Three Stages, Three Channels, One Key

Generation in three stages. Generation composes one in-sampling and two post-hoc channels in a fixed order (Figure 1a). The cryptographic core (below) generates the text, opening a new block whenever accumulated empirical entropy crosses a threshold; the boundaries define the compartments, each with its offsets, bit range, and seed. A deterministic rule then assigns roles so that cryptographic and lexical evidence lie on *disjoint* spans: the seed compartment stays plain and the first C_{budget} scorable compartments retain `crypto`, later compartments too short to vote (fewer than $\ell_{\min}$ scorable bits) are ceded to the lexical channel, and a fraction ϕ_{zone} of the surplus is promoted to `syn`, all reproducible at detection (Appendix E). Finally the lexical channel substitutes inside the synonym zones, the orthographic channel substitutes homoglyphs globally at rate ρ, and partition, roles, and seeds are recorded in the authenticated registry. C_{budget}, $\ell_{\min}$, and ϕ_{zone} are design choices fixed before any experiment (see Appendix E); they decide which channel carries the verdict on a given document and do not affect soundness or isolation, which Theorem 1 and Proposition 1 guarantee for any setting (a worked example: Appendix B). All three channels draw on the same keyed function F_K but ask it different questions – *which bit* to emit (cryptographic), *which synonym* to place (lexical), *which positions* to substitute (orthographic) – and at detection each re-asks its question under the candidate key and counts how often the text agrees.

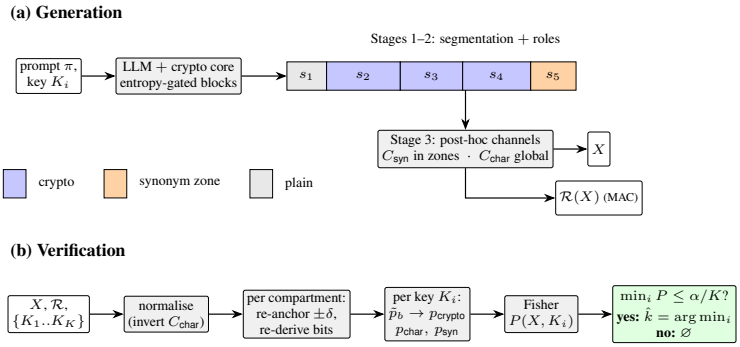

Figure 1: The SAFIZ pipeline. **(a)** The core segments the document into compartments, assigns roles, applies the post-hoc channels (lexical in-zone, orthographic globally), and logs the structure in the registry. **(b)** Invert the orthographic map; re-anchor each compartment ($\pm\delta$) and re-derive its bits; fuse p-values by Fisher per key; attribute the minimum iff it clears α/K, else abstain.

Cryptographic channel $\mathcal{C}_{\text{crypto}}$ (in-sampling, zoned by entropy). Following Christ et al. (2024), tokens are generated bit by bit over the binary encoding of the vocabulary, each bit chosen by inversion sampling: if the model assigns probability p to the bit being 1 given the bits already chosen, the bit is 1 exactly when the PRF value $u = F_K(b, \text{seed}_b, j)$ satisfies $u \le p$, so the model's distribution is sampled faithfully with the key supplying the randomness; the PRF re-seeds at every block boundary, the seed of block b being the bits emitted since the previous boundary. The core is active only on spans of high empirical entropy; low-entropy and ultra-short spans carry no cryptographic evidence by the *impossibility result* and are exactly the spans the orthographic channel covers.

Lexical post-hoc channel $\mathcal{C}_{\text{syn}}$ (zoned by role). A post-generation substitution operator over hand-checked synonym classes, keyed by F_K on the zone index, the slot position, and the 8 preceding base-form words, confined to the synonym zones; it is not invertible, which is why its support must be disjoint from the crypto compartments.

Orthographic post-hoc channel $\mathcal{C}_{\text{char}}$ (global). An exact bijective map within the Unicode character space (Rizzo et al., 2016), substituting ASCII graphemes with visually indistinguishable lookalikes at key-driven positions at rate ρ; the PRF reads the position with the 16 preceding normalised characters, so the substitution mask is a fresh draw for every document rather than a constant of the key (Remark 2). Being a bijection, it is inverted by normalisation at the start of detection.

Composition. Synonym substitution rewrites the very tokens the core scores, and homoglyph substitution alters the bytes re-tokenisation depends on; SAFIZ reconciles them by *zoning* (the non-invertible lexical channel is confined to spans disjoint from the crypto blocks) and *invertibility* (the orthographic channel may overlap them because its effect is undone before scoring), the general condition for interference-free composition (Remark 1).

What composition buys, and what it costs. With their *impossibility result*, Christ et al. (2024) prove that no undetectable scheme can watermark output whose empirical entropy is too low (empirically, the scheme of Kuditipudi et al. (2023) reports $\sim$25% detectability at $p \le 0.01$ on short instruction-following output); SAFIZ keeps the undetectable core where the theory permits it and covers the provably unreachable regime with a post-hoc channel that does not depend on entropy. The cost is system-level undetectability (homoglyphs are flaggable by a key-less observer, a fragility this family of techniques declares by design), and it is bounded: detecting *that* a text is marked reveals nothing about *which* key marked it, so soundness and isolation are unaffected, and any scheme covering the low-entropy regime must give up undetectability there. The two channels are thus complementary *by construction* – the core alone suffices on long output and base models, the orthographic channel works on any text but can be stripped – and Theorem 1 holds whichever carries the evidence (Appendix D).

3.4 Detection: Re-anchoring, Re-derivation, and Calibrated Fusion

Given a suspect document, the registry, and the candidate keys, the detector first inverts the orthographic map by normalisation, then evaluates each key in four steps (Figure 1b; Appendix C), each

producing a *calibrated* p-value: the probability, under a wrong key, of evidence at least as strong as the one observed, stochastically no smaller than $\mathrm{Unif}(0,1)$ under $H_0^{(i)}$.

Step 1: local re-anchoring. For each `crypto` compartment the detector reads $X[c_s{:}c_e]$ at the logged offsets and re-tokenises the substring locally, exactly as generation did (a span tokenised on its own is not guaranteed to match the same span tokenised inside the whole document), repeating the read at the $2\delta+1$ character shifts to absorb drift. Because every compartment is read at its own absolute offsets, an edit displaces only the compartments after it and never mixes one compartment's bits with another's: damage stays confined by the data layout.

Step 2: bit re-derivation and alignment score. At each shift, the substring's bit string $\beta \in \{0,1\}^\ell$ is re-derived and scored against the PRF stream keyed by the candidate:

$$S \;=\; \sum_{j=1}^{\ell} -\ln v_j, \qquad v_j = \begin{cases} u_j & \text{if } \beta_j = 1, \\ 1-u_j & \text{if } \beta_j = 0, \end{cases} \tag{1}$$

where $u_j = F_{K_i}(b, \mathrm{seed}_b, j)$ and ℓ is the number of scored bits; v_j is the probability mass the PRF left on the side of the bit actually observed, and $-\ln v_j$ turns it into a positive score, large when v_j is small. Its null distribution is exact (Lemma 1, Appendix A): if the u_j are i.i.d. $\mathrm{Unif}(0,1)$ and independent of β (the ideal-PRF null), then the v_j are i.i.d. uniform, $S \sim \mathrm{Gamma}(\ell, 1)$ *exactly*, and $p = Q(\ell, S)$ is exactly uniform, Q being the regularised upper incomplete gamma function. Under the right key, inversion sampling biases each v_j downward in proportion to the entropy spent on that bit, inflating S; the per-compartment p-value is an exact Gamma tail, not a Gaussian approximation, so it is valid for compartments of any size.

Step 3: per-compartment p-value. The detector keeps the most significant shift and multiplies its p-value by the number of shifts tried, $\tilde{p}_b = \min\{1, (2\delta+1)\min_d Q(\ell, S_{b,d})\}$, so that trying several alignments cannot inflate significance (Lemma 2); compartments with fewer than $\ell_{\min}$ bits abstain.

Step 4: channel p-values and two-level fusion. *Cryptographic:* the Fisher combination (Equation 3) of the $\tilde{p}_b$ across voting compartments. *Orthographic:* at the n positions that K_i predicts to be substituted, k carry the prescribed homoglyph; the channel abstains when $n < n_{\min}$, and otherwise

$$p_{\mathrm{char}}(K_i) = \mathrm{SF}(k; n, q), \qquad \mathrm{SF}(k; n, q) \;=\; \Pr\big[\mathrm{Bin}(n, q) \geq k\big], \qquad q = \max(\rho, \hat{d}), \tag{2}$$

where SF is the survival function, $\hat{d}$ is the homoglyph density measured on the eligible positions *not* predicted by K_i, disjoint from those entering the count k and hence independent of it; the adaptive term q guards the null against documents whose homoglyph density has been inflated by editing, since under a matching key $\hat{d} \approx 0$ and power is unchanged, while $\hat{d} > \rho$ makes the p-value conservative (Lemma 4). *Lexical:* among the n checkable tokens inside the synonym zones, k match the choice prescribed by K_i, and $p_{\mathrm{syn}}(K_i) = \mathrm{SF}(k; n, q_{\mathrm{syn}})$ with q_{syn} the mean inverse class size; the channel votes only when its zones are present, and both binomial tails are conservative by discreteness (Lemma 3). Let $\mathcal{S}(K_i) \subseteq \{\mathrm{crypto}, \mathrm{char}, \mathrm{syn}\}$ be the channels active for K_i and $m = |\mathcal{S}(K_i)|$; SAFIZ combines their p-values by Fisher's method,

$$X^2(K_i) \;=\; -2 \sum_{c \,\in\, \mathcal{S}(K_i)} \ln p_c(K_i) \;\overset{H_0}{\sim}\; \chi^2_{2m}, \qquad P_{\mathrm{combined}}(X, K_i) \;=\; \Pr\big[\chi^2_{2m} > X^2(K_i)\big]. \tag{3}$$

Because each p_c is calibrated, P_{combined} has a *known* null law and is itself a calibrated p-value: no fusion weights exist to tune, hence none can silently inflate the false-positive rate.

3.5 Multi-Task Attribution and Guarantees

Given X, the candidate keyspace $\mathcal{K}$, the target rate α, and the registry, the attribution operator $\mathcal{A}\colon \Sigma^* \times \mathcal{K} \to \mathcal{K} \cup \{\varnothing\}$ returns the best-supported key if it clears a Bonferroni-corrected threshold and abstains otherwise: with $K_\star = \arg\min_{K_i \in \mathcal{K}} P_{\mathrm{combined}}(X, K_i)$,

$$\mathcal{A}(X, \mathcal{K}) = \begin{cases} K_\star & \text{if } P_{\mathrm{combined}}(X, K_\star) < \alpha/K, \\ \varnothing & \text{otherwise} \end{cases} \tag{4}$$

(Algorithm 1, Appendix C).

Two assumptions underlie the guarantees: ***PRF security*** (Assumption 1), that no efficient adversary distinguishes F_K from a random function with advantage better than $\mathrm{negl}(\lambda)$; and *channel independence under the null* (Assumption 2), enforced structurally (disjoint PRF domains, zones disjoint from crypto compartments, orthographic map inverted before scoring) and verified empirically in Section 4.3.

Theorem 1 (Soundness)**.** *Under Assumptions 1 and 2, the scheme of Definition 1 with the rule of Equation 4 is* $(\alpha, \mathrm{negl}(\lambda))$*-sound: for any* X *and* $\mathcal{R}$ *fixed independently of* $\mathcal{K}$, $\Pr\left[\mathcal{A}(X, \mathcal{K}) \neq \varnothing\right] \leq \alpha + \mathrm{negl}(\lambda)$.

Proof idea. In the ideal-PRF model each per-compartment p-value is exactly uniform (Lemma 1), every correction is conservative (Lemmas 2–4), Fisher preserves conservativeness (Lemma 5), and Bonferroni union-bounds the K keys; replacing the random function by F costs at most the PRF advantage (Appendix A, which also proves the propositions below).

Proposition 1 (Isolation)**.** *Under Assumption 1 the scheme is* $\mathrm{negl}(\lambda)$*-isolating: the statistics computed for a non-generating key* K_j *depend on* K_j *only through PRF evaluations on inputs fixed by the text and registry, so a distinguisher between cross-key and null scores breaks* F.

Proposition 2 (Fail-closed)**.** *The rule of Equation 4 is fail-closed: any evidence-destroying edit, including registry corruption and truncation, can only increase the combined* p*-values and so moves the outcome toward* $\varnothing$*, never toward a key.*

4 Evaluation

We evaluate attribution accuracy, false-positive rates capturing soundness and key-selectivity (including scaling in K), per-channel ablations, and behaviour under signal removal, reporting counts alongside rates and Wilson 95% intervals for the small rates on which the security claims rest.

4.1 Experimental Setup

Generation. Text is generated by six models spanning three families and two post-training regimes: Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct and Mistral-7B-Instruct, and Llama-3.1-8B, Qwen3-8B-Base and GPT-2-medium (base, prompted in completion style with the same prompt content); Llama-3.1-8B appears in both regimes, a controlled within-family contrast, and GPT-2-medium is a high-entropy reference point. Prompts come from ELI5 (`qa`), XSum (`summary`) and C4 RealNewsLike (`news`), frozen in a versioned file; each model generates 50 documents per task at three lengths (50, 100, 200 tokens) under three seeds, $8{,}100$ in all. **Configuration.** The configuration is *homogeneous*, every task carrying all three channels with identical parameters; keys are independent 256-bit CSPRNG values, versioned with the code; $\alpha = 0.01$, so $\alpha/K = 0.0033$. Channel parameters were fixed before the campaign and never adjusted: $\rho = 0.25$, $n_{\min} = 5$ with the adaptive null, $q_{\mathrm{syn}} = 0.409$, $\ell_{\min} = 48$, $C_{\mathrm{budget}} = 3$, $\phi_{\mathrm{zone}} = 0.5$, $\delta = 1$ (role and rule behind each value in Appendix E, sensitivity of the detection-side ones in Appendix F.7). A *heterogeneous* per-task configuration is evaluated in Section 4.4. **Baselines.** No published system performs multi-key attribution with a family-wise guarantee (no implementation or empirical evaluation of Cohen et al. (2025) is reported), so we use three internal baselines: the *crypto-only* configuration, per-key attribution with the undetectable core of Christ et al. (2024) alone (Section 4.4); the *uncorrected max-score* rule, which attributes to the best-scoring key at level α without the α/K correction, the naive baseline whose inflation Fu & Russell (2025) analyse (Section 4.3); and, for robustness, the binary single-key detection metric of Kuditipudi et al. (2023) (Section 4.5).

4.2 Multi-Task Attribution

Across the six models SAFIZ attributes 99.19% of the $8{,}100$ documents with *zero* misattributions; the 66 errors are all abstentions, stable across tasks, lengths and seeds (Appendix F.1), and detection and attribution coincide by construction of the rule (Remark 3). The two rightmost columns of Table 1 measure the phenomenon the paper is designed around, and the Llama pair isolates its cause: at fixed family, tokenizer and size, instruction tuning raises the zero-block fraction from 22.6% to 47.2%, *doubling* the share of documents on which the core has no evidence to give; across the instruct models half to two thirds of documents have no scorable block, against 3.9% on GPT-2-medium. On that

Table 1: Attribution on clean text ($N = 1{,}350$ per model; $K = 3$, $\alpha = 0.01$, three seeds); every error is an abstention (Appendix F.1). Blocks: mean scorable cryptographic blocks per document; zero-block: fraction of documents with none. Seed sd: accuracy std deviation (percentage points).

Model	Accuracy	Misattr.	Abstain	Blocks/doc	Zero-block	Seed sd (pp)
Llama-3.1-8B-Instruct	99.93%	0	1	2.27	47.2%	0.13
Qwen2.5-7B-Instruct	100.00%	0	0	0.52	71.6%	0.00
Mistral-7B-Instruct	99.85%	0	2	0.56	68.7%	0.13
Llama-3.1-8B (base)	97.63%	0	32	1.97	22.6%	0.13
Qwen3-8B-Base	99.33%	0	9	1.10	51.2%	0.59
GPT-2-medium	98.37%	0	22	4.05	3.9%	0.64
All	99.19%	0	66	1.75	44.2%	—

majority regime the core is structurally silent, and that accuracy is nonetheless 97.6–100% is the operating-envelope claim of Section 3.3 measured; the Llama base abstentions reflect degenerate or empty continuations produced by completion-style prompting.

4.3 Soundness, Key-Selectivity, and Scaling in the Number of Keys

Theorem 1 bounds the family-wise false-attribution rate by $\alpha + \mathrm{negl}(\lambda)$ and, via the Bonferroni threshold, any single key's accusation rate by α/K; we verify both over 872,310 null trials (Table 5, Appendix F.2): (a) $4{,}985$ human texts (CNN/DailyMail, AG News, WikiText) under six tokenizers and synthetic registries (29,910 full-attributor trials): 15 false accusations, 0.050% family-wise, twenty times below α; (b) the $8{,}100$ watermarked documents, each against 100 fresh keys, one key per trial at $\alpha/K = 0.33\%$ (810,000 trials): 427 accusations, 0.053%, six times below that bound; (c) corrupted registries on watermarked documents, counting only attributions to a non-generating key (32,400 trials): none. The four corruption modes (shifted offsets, random seeds, shuffled compartments, another document's registry) are the empirical counterpart of Proposition 2; under corruption the verdict usually remains *correct*, because the orthographic channel does not depend on the registry. Key-selectivity is probed with the other live keys (full rule at α/K: 5 of 16,200, 0.031%) and, for the cryptographic core alone at α, with fresh random keys (106 of $40{,}500$, 0.262%) and adjacent keys $K{+}1$ (23 of $8{,}100$, 0.284%): all Wilson intervals lie below threshold, and adjacent and random keys are indistinguishable, as PRF security predicts.

Fu & Russell (2025) prove that thresholding the maximum score over K keys without correction inflates the false-detection rate as K grows; we reproduce this on SAFIZ itself by re-scoring the corpus against nested key banks (three live keys plus random decoys) at $K \in \{3, 10, 30, 100\}$ with and without the Bonferroni threshold (Appendix F.3): uncorrected, the family-wise false-positive rate grows $0.28 \to 1.61 \to 4.96 \to 14.90\%$, with misattributions from $K{=}10$; corrected, it stays at or below 0.14% at every K with zero misattributions, at a power cost of 0.56 accuracy points between $K{=}3$ and $K{=}100$, and detection time is linear in K (4.2–4.3 ms per key per document).

4.4 Channel Ablation, the Registry, and Channel Allocation

The ablation (Appendix F.4) makes the division of labour explicit: the orthographic channel alone reaches 99.23%, the dominant carrier on this corpus, consistent with the zero-block fractions of Table 1; the crypto-only baseline reaches 42.36%, below the 55.8% of documents that close at least one block; the lexical channel contributes no measurable evidence (0.21%), and is also the only component that perturbs the text: perplexity under a third-party model is unchanged by the core, whose paired per-document difference from unwatermarked generation is balanced in sign on all six models, while the lexical channel fires on a minority of them and costs a few perplexity points when it does (Appendix F.8). On clean text the post-hoc channels, which need no registry, match the full system (99.23% vs. 99.19%), while under homoglyph stripping (Section 4.5) the surviving 37.7% is carried entirely by the registry-enabled core: the registry is an optimisation and defence in depth, not a requirement on clean text (Remark 5); we do not run blind detection of the core without a registry. Channels are also configurable per task, and the guarantees do not depend on the allocation: re-running the protocol on Llama-3.1-8B-Instruct with `qa` crypto+char, `news`

all three, and summary crypto+syn (the low-entropy task loses the entropy-independent channel) gives 100.0% on qa and 99.8% on news against 10.2% on summary, which without the entropy-independent channel depends on the core alone: 82.7% of its documents close no cryptographic block at all, and accuracy rises with length as more entropy accumulates (2.7, 4.7, 23.3% at 50, 100, 200 tokens; Appendix F.6). The 405 failures are all abstentions, and 135,000 fresh-key null trials give 0.056%, six times below α/K: removing the channel costs recall where evidence cannot exist, and nothing in soundness, which Theorem 1 guarantees for any set of channels.

4.5 Unicode Normalisation and Robustness Under Signal Removal

A finding that changes the shape of this attack surface (Table 8, Appendix F.5): *standard Unicode normalisation does not touch the channel*. NFC, NFKC, NFD and NFKD leave all 412,545 substituted characters in place and accuracy at 98.9–99.2%, because cross-script confusables are distinct characters, not compatibility variants; the one-line defence commonly assumed for homoglyph watermarks does not exist, and removal needs an explicit confusables mapping. With one the channel dies as designed: the strip targeted to the system's alphabet leaves 37.7%, carried by the cryptographic core, while the full UTS #39 map of 1,535 pairs (a superset, per a secondary source, of the alphabet of Rizzo et al., 2016) also rewrites punctuation the models generate natively, corrupting re-tokenisation and collapsing accuracy to 5.8% while visibly altering the document; misattributions stay within the bound (0 and 2 of 8,100).

Signal removal is outside the security claim (Section 3.2); we characterise it by scoring each of the 8,100 documents under 48 conditions: truncation (75–10% retained), character noise (1–20%), token substitutions, insertions and deletions to 70% following Kuditipudi et al. (2023), their mixture, and paraphrase in four variants. Truncation degrades gracefully (99.0 / 98.1 / 86.3 / 42.5% at 75/50/25/10% retained, with *zero* misattributions in 40,500 verdicts), because compartments anchor independently. Token-level edits are the brittle axis: 5% substitutions leave 11.8% attribution and by 10% the system abstains almost everywhere, whereas the alignment-based detector of Kuditipudi et al. (2023) is designed for this regime. The trade is measured: on their clean-text metric (binary single-key detection at $p \leq 0.01$) SAFIZ attains 99.95% on the three instruction-tuned models (99.3% over all six) where they report ∼25% for Alpaca-7B (models and tasks differ, so the figures are indicative). Character noise sits between the two, and its failure mode is diagnostic: at 5% noise 96.8% of the homoglyphs survive, yet attribution falls to 13.8%, because the context-bound PRF can no longer re-derive the positions; the loss is in reading, not in the mark. Paraphrase in four paired variants (original 99.2%, first sentence rewritten 8.6%, first 80 words rewritten with tail kept or dropped 0.1%) yields 12 misattributions in 32,400 verdicts (0.037%), where before the adaptive null this figure was 3.5–3.7% with intervals excluding α (Appendix F.5). Across all 388,800 verdicts, failures are overwhelmingly abstentions and per-condition misattribution stays within the bound.

5 Limitations and Conclusion

Channel concentration. On this corpus the orthographic channel carries most of the evidence, so clean-text accuracy is exposed to confusables stripping; this is the fragility the design declares, no undetectable channel can fill the gap, and the core retains 37.7% as defence in depth. *Context-window sensitivity.* Binding the post-hoc PRFs to local context makes the null document-fresh but turns character noise into a read failure before the marks are destroyed; a shorter or error-tolerant context is a direct next step. *No adaptive adversary.* As is standard in this literature the adversary does not adapt to the scheme; since soundness and isolation do not depend on its strategy, an adaptive one could gain recall, not a false attribution; the heterogeneous experiment covers one model.

We introduced Selective Watermarking, which upgrades watermark detection from a binary test to secure multi-class attribution, and realised it in SAFIZ: an undetectable core and two post-hoc channels, routed to disjoint compartments under distinct keys and fused by Fisher's method, with an exact guarantee (Theorem 1) and evidence exactly where the core cannot supply it, at the declared cost of system-level undetectability. Beyond this instantiation, the composition condition of Remark 1 tells when a channel can be added to any selective scheme, and the regime analysis of Section 2 delimits where key selection is the right tool (few classes, short outputs); next steps are semantic channels and an adaptive adversary.

Ethics statement

This work concerns provenance and attribution of machine-generated text, a capability with dual-use aspects. Attribution supports accountability, auditability, and compliance with transparency obligations for providers of generative models; the same machinery, if deployed with per-user rather than per-task keys, could support surveillance of individual users, which is precisely the regime we do not target and whose costs we make explicit (Section 2). The orthographic channel relies on Unicode confusables, a mechanism historically associated with phishing and filter evasion; we use it only on text generated by the provider's own pipelines, characterise its removability (standard Unicode normalisation does not remove it; an explicit confusables mapping does), and note that its presence is detectable by any key-less observer. No human subjects or personal data are involved; all corpora are model-generated or standard public text.

Reproducibility statement

All security definitions and complete proofs of the guarantees are in Section 3 and Appendix A, including the assumptions (PRF security, channel independence) under which they hold. Generation and detection parameters, key handling, and corpus composition are specified in Section 4.1; the decision rule and detection procedure are given in full as Algorithms 1 and 2. An anonymised implementation of the generation and verification pipeline, together with scripts reproducing every table, is provided as supplementary material.

AI use statement

In this work, we used generative AI tools to aid and polish parts of the manuscript for readability, to create scientific figures, to assist in the writing of proofs, to identify related literature, and to assist in implementing and debugging the experimental code. We did not use generative AI tools to generate synthetic data, to clean or process the experimental data, to derive the proofs, or to analyse the results. All AI-assisted content was reviewed by the authors, including literature attributions checked against the original sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1125–1139. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/christ24a.html.

Aloni Cohen, Alexander Hoover, and Gabe Schoenbach. Watermarking language models for many adaptive users. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 2583–2601, 2025. doi: 10.1109/SP61157.2025.00084.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Ali Cemgil, Zahra Ahmed, Kitty Stacpoole, and Pushmeet Kohli. Scalable watermarking for identifying large language model outputs. *Nature*, 634:818–823, 10 2024. doi: 10.1038/s41586-024-08025-4.

Zihao Fu and Chris Russell. Multi-use LLM watermarking and the false detection problem, 2025. arXiv:2506.15975.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In Marta R. Costa-jussà and Enrique Alfonseca (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3019. URL https://aclanthology.org/P19-3019/.

Abe Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. SemStamp: A semantic watermark with paraphrastic robustness for text generation. In Kevin Duh, Helena Gomez, and

Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4067–4082, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.226. URL https://aclanthology.org/2024.naacl-long.226/.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 45408–45436, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/c5b00c5bdcc6fe35907dbcca03d27652-Paper-Conference.pdf.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17061–17084. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust Distortion-free Watermarks for Language Models. *arXiv e-prints*, art. arXiv:2307.15593, July 2023. doi: 10.48550/arXiv.2307.15593.

Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, Hwaran Lee, Sangdoo Yun, Jamin Shin, and Gunhee Kim. Who wrote this code? watermarking for code generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4890–4911, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.268. URL https://aclanthology.org/2024.acl-long.268/.

Yuqing Liang, Jiancheng Xiao, Wensheng Gan, and Philip S. Yu. Watermarking Techniques for Large Language Models: A Survey. *arXiv e-prints*, art. arXiv:2409.00089, August 2024. doi: 10.48550/arXiv.2409.00089.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark for large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 6499–6519, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/1a2131ebe25bd55e4fc734126ea583ed-Paper-Conference.pdf.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. A survey of text watermarking in the era of large language models. *ACM Comput. Surv.*, 57(2), November 2024b. ISSN 0360-0300. doi: 10.1145/3691626. URL https://doi.org/10.1145/3691626.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24950–24962. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/mitchell23a.html.

Georg Niess and Roman Kern. Ensemble watermarks for large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2903–2916, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.145. URL https://aclanthology.org/2025.acl-long.145/.

Leyi Pan, Aiwei Liu, Zhiwei He, Zitian Gao, Xuandong Zhao, Yijian Lu, Binglin Zhou, Shuliang Liu, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. MarkLLM: An open-source toolkit for LLM watermarking. In Delia Irazu Hernandez Farias, Tom Hope, and

Manling Li (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 61–71, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-demo.7. URL https://aclanthology.org/2024.emnlp-demo.7/.

Stefano Giovanni Rizzo, Flavio Bertini, and Danilo Montesi. Content-preserving text watermarking through unicode homoglyph substitution. In *Proceedings of the 20th International Database Engineering & Applications Symposium (IDEAS)*, pp. 97–104. ACM, 2016. doi: 10.1145/2938503.2938510.

Stefano Giovanni Rizzo, Flavio Bertini, and Danilo Montesi. Fine-grain watermarking for intellectual property protection. *EURASIP Journal on Information Security*, 2019(10), 2019. doi: 10.1186/s13635-019-0094-2.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated text. *Commun. ACM*, 67(4):50–59, March 2024. ISSN 0001-0782. doi: 10.1145/3624725. URL https://doi.org/10.1145/3624725.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable Robust Watermarking for AI-Generated Text. *arXiv e-prints*, art. arXiv:2306.17439, June 2023. doi: 10.48550/arXiv.2306.17439.

A PROOFS

This appendix proves the guarantees of Section 3.5 under the two assumptions stated there in prose.

Assumption 1 (PRF security)**.** No probabilistic polynomial-time adversary distinguishes $F_K(\cdot)$, for uniform $K \in \{0,1\}^\lambda$, from a uniformly random function into $(0,1)$ with advantage better than $\mathrm{negl}(\lambda)$.

Assumption 2 (Channel independence under the null)**.** Under $H_0^{(i)}$, the active channel statistics for K_i are mutually independent.

Throughout, "the ideal model" replaces each F_{K_i} by a uniformly random function into $(0,1)$, sampled independently across keys; Assumption 1 transfers the results to the real PRF up to a negligible additive term.

Lemma 1 (Exact per-compartment null)**.** *Let $u_1, \ldots, u_\ell$ be i.i.d.* $\mathrm{Unif}(0,1)$*, independent of $\beta \in \{0,1\}^\ell$, and let $v_j = u_j$ if $\beta_j = 1$ and $v_j = 1 - u_j$ otherwise. Then $S = \sum_j -\ln v_j \sim \mathrm{Gamma}(\ell, 1)$ and $p = Q(\ell, S) \sim \mathrm{Unif}(0,1)$, where Q is the regularised upper incomplete gamma function.*

Proof. Condition on β. If $u \sim \mathrm{Unif}(0,1)$ then $1 - u \sim \mathrm{Unif}(0,1)$, so each $v_j \sim \mathrm{Unif}(0,1)$; the map $u \mapsto 1 - u$ is measure-preserving and applied coordinate-wise, so the v_j remain independent. Hence $-\ln v_j \sim \mathrm{Exp}(1)$ i.i.d. and their sum is $\mathrm{Gamma}(\ell, 1)$. Since the conditional law does not depend on β, the unconditional law is the same. Uniformity of p is the probability integral transform applied to the continuous survival function $Q(\ell, \cdot)$. □

Lemma 2 (Shift correction is conservative)**.** *Let p_d, $d \in \{-\delta, \ldots, \delta\}$, be p-values (each stochastically $\geq \mathrm{Unif}(0,1)$ under H_0, not necessarily independent) and let $\tilde{p} = \min\{1, (2\delta + 1) \min_d p_d\}$. Then $\Pr[\tilde{p} \leq t] \leq t$ for all $t \in [0,1]$.*

Proof. $\Pr[\tilde{p} \leq t] = \Pr[\exists d : p_d \leq t/(2\delta+1)] \leq \sum_d \Pr[p_d \leq t/(2\delta+1)] \leq (2\delta+1) \cdot t/(2\delta+1) = t$. □

Lemma 3 (Binomial tails are conservative)**.** *If $k \sim \mathrm{Bin}(n, q')$ with $q' \leq q$, then $p = \mathrm{SF}(k; n, q)$ satisfies $\Pr[p \leq t] \leq t$.*

Proof. For a discrete statistic, $\Pr[\mathrm{SF}(k; n, q) \leq t] = \Pr[k \geq k_t]$ where k_t is the smallest integer with $\mathrm{SF}(k_t; n, q) \leq t$; by definition of the survival function this probability equals $\mathrm{SF}(k_t; n, q) \leq t$ when $k \sim \mathrm{Bin}(n, q)$, and stochastic domination in q ($q' \leq q$) can only decrease it. □

Lemma 4 (Adaptive orthographic null)**.** *Let the n positions predicted by the candidate key be disjoint from the m eligible positions on which the density $\hat{d}$ is measured, and suppose that under H_0 each predicted position carries the prescribed homoglyph independently with probability q'. If $q' \leq \max(\rho, \hat{d})$, then conditionally on $\hat{d}$ the statistic $p = \mathrm{SF}(k; n, \max(\rho, \hat{d}))$ satisfies $\Pr[p \leq t \mid \hat{d}] \leq t$, and hence unconditionally $\Pr[p \leq t] \leq t$: the level is preserved.*

Proof. Disjointness makes k and $\hat{d}$ independent under H_0, so conditioning on $\hat{d}$ leaves $k \sim \mathrm{Bin}(n, q')$; the claim is then Lemma 3 with $q = \max(\rho, \hat{d}) \geq q'$, and the unconditional bound follows by taking expectation over $\hat{d}$. □

The hypothesis $q' \leq \max(\rho, \hat{d})$ holds exactly in the designed nulls: on clean or differently-keyed text the collision probability is at most ρ (the context-bound PRF makes the masks of distinct keys independent), and $\max(\rho, \hat{d}) \geq \rho$ always. Under tampering that inflates the document's homoglyph density, q' tracks the inflated density and $\hat{d}$ estimates it on m disjoint positions, so the hypothesis holds up to the estimation error of $\hat{d}$, which shrinks as $O(m^{-1/2})$; the empirical check of Section 4.5 (12 misattributions in 32,400 paraphrase verdicts) confirms the level in exactly this regime. An exact alternative is the conditional (hypergeometric) test given the document's total homoglyph count, which requires no density estimate; we retain the adaptive binomial for continuity with the deployed detector.

Lemma 5 (Fisher combination of conservative p-values)**.** *Let $p_1, \ldots, p_m$ be independent with $\Pr[p_c \le t] \le t$ for all t, and let $P = \Pr[\chi^2_{2m} > -2\sum_c \ln p_c]$. Then $\Pr[P \le t] \le t$.*

Proof. Couple each p_c with $p_c^\star \sim \mathrm{Unif}(0,1)$ such that $p_c^\star \le p_c$ almost surely: define $p_c^\star = G_c(p_c)$ pathwise via the inverse-CDF construction, using that the CDF of p_c is pointwise below the identity. Then $-2\sum \ln p_c \le -2\sum \ln p_c^\star$ pointwise, so $P \ge P^\star$ pointwise, where $P^\star$ is the Fisher p-value of the uniform p-values; $P^\star$ is exactly uniform, because the $-2\ln p_c^\star$ are i.i.d. χ^2_2, their sum is χ^2_{2m}, and the tail transform is the probability integral transform. Hence $\Pr[P \le t] \le \Pr[P^\star \le t] = t$. □

Proof of Theorem 1. Work first in the ideal model.

Step 1 (per-key calibration). Fix a candidate key K_i and condition on $(X, \mathcal{R})$, which are independent of K_i by hypothesis. Domain separation (Remark 2) makes all PRF inputs across compartments, shifts, and channels distinct, so the associated ideal outputs are i.i.d. $\mathrm{Unif}(0,1)$. By Lemma 1 each per-compartment p-value at each shift is exactly uniform; by Lemma 2 the shift-corrected $\tilde{p}_b$ are conservative; by Lemma 5 applied at the compartment level, the cryptographic channel p-value is conservative. The lexical and orthographic p-values are conservative by Lemma 3. By Assumption 2 the active channel p-values are independent, so by Lemma 5 applied at the channel level, the fused $P_{\mathrm{combined}}(X, K_i)$ is conservative: $\Pr[P_{\mathrm{combined}}(X, K_i) \le t] \le t$ for all t.

Step 2 (family-wise union bound).

$$\begin{aligned}\Pr\left[\mathcal{A}(X,\mathcal{K}) \neq \varnothing\right] &= \Pr\left[\min_i P_{\mathrm{combined}}(X, K_i) < \alpha/K\right] \\ &\le \sum_{i=1}^{K} \Pr\left[P_{\mathrm{combined}}(X, K_i) < \alpha/K\right] \;\le\; K \cdot \frac{\alpha}{K} \;=\; \alpha.\end{aligned}$$

Step 3 (from ideal to real). The event $\{\mathcal{A} \neq \varnothing\}$ is an efficiently computable predicate of the PRF outputs. If its probability under F exceeded its ideal-model probability by a non-negligible ϵ, the detector itself would constitute a distinguisher between F_K and a random function with advantage at least ϵ/K against some key, contradicting Assumption 1. Hence $\Pr[\mathcal{A} \neq \varnothing] \le \alpha + \mathrm{negl}(\lambda)$. □

Proof of Proposition 1. Fix $i \neq j$ and a document $(X, \mathcal{R}) = \mathsf{Wat}_{K_i}(\pi)$. Every statistic that Verify computes for the candidate key K_j is a deterministic function of $(X, \mathcal{R})$ and of PRF evaluations $F_{K_j}(\cdot)$ on inputs determined by $(X, \mathcal{R})$. Since K_j is sampled independently of K_i and hence of $(X, \mathcal{R})$, replacing F_{K_j} by a random function replaces the score vector for K_j by its exact null distribution (Step 1 above). An efficient distinguisher between the real cross-key scores and null scores therefore breaks the PRF with the same advantage. In particular, by Theorem 1 restricted to the single key K_j, the probability that X is attributed to K_j is at most $\alpha/K + \mathrm{negl}(\lambda)$. □

Proof of Proposition 2. The range of $\mathcal{A}$ is $\mathcal{K} \cup \{\varnothing\}$ by Equation 4, and a key is returned only when its fused p-value is below α/K, so the rule satisfies Definition 2(c) by construction; what remains is the direction of degradation. Every channel statistic is a count of agreements between the text and the answer the candidate key prescribes (Section 3.3), and every channel p-value is a non-increasing function of that count: the Gamma tail $Q(\ell, S)$ decreases in S, and the binomial tails decrease in k. An edit that destroys evidence—truncation removing compartments, corruption of offsets so that a compartment is read at the wrong place, a homoglyph stripped—removes or randomises agreements it does not create, since agreements under K_i on positions the edit touches are, under the PRF, as likely as under a wrong key; it therefore lowers the counts, or removes a voting compartment, or inflates $\hat{d}$ in Equation 2, each of which weakly increases the channel p-value. Fisher's statistic is monotone in its inputs, so $P_{\mathrm{combined}}(X, K_i)$ weakly increases for every K_i, and the minimum over keys can only cross the threshold α/K in the direction of $\varnothing$. An edit that instead *creates* agreements under some key does so with the probability of a wrong key, which is the event Theorem 1 bounds; it is not an evidence-destroying edit and lies outside the proposition. □

B Worked Example: A Detection in Six Bits

We follow one document end to end: the QA-task answer *"The Amazon river discharges about one fifth of the world's fresh water. Its basin spans nine countries and remains the planet's largest rainforest"*, generated under key K_{qa}. Segmentation (Section 3.3) splits it into $B = 5$ compartments: the seed compartment s_1, which is never scored, entropy-rich s_2, s_3 and s_4, and a short final s_5 (*"remains the planet's largest rainforest"*). Role assignment with $C_{\mathrm{budget}} = 3$ keeps s_2–s_4 as `crypto` and cedes s_5, too short to vote and beyond the budget, to the lexical channel, which then substitutes "largest" → "biggest" in s_5 as prescribed by $F_{K_{\mathrm{qa}}}$; the orthographic channel substitutes homoglyphs at rate ρ over the whole document. The following instantiates the detection pipeline of Section 3.4 on this document, with the exact statistics used by the system.

A detection in six bits

Candidate key K_{qa}, $K = 3$ tasks, $\alpha = 0.01$, $\delta = 1$. Compartment s_2 of the running example re-derives $\ell = 6$ bits; the PRF gives per-bit values $v = (0.08, 0.21, 0.05, 0.33, 0.12, 0.09)$, so

$$S = \textstyle\sum_j -\ln v_j = 12.72, \qquad p_{s_2} = Q(6, 12.72) = 0.0129,$$

and after the $2\delta{+}1 = 3$-shift correction $\tilde{p}_{s_2} = 3 \times 0.0129 = 0.039$. (Under a wrong key $\mathbb{E}[S] = \ell = 6$; the observed S is more than twice its null mean.) Suppose the crypto channel reduces to this single voting compartment, the orthographic channel finds $k = 19$ prescribed homoglyphs among $n = 40$ eligible positions ($p_{\mathrm{char}} = \mathrm{SF}(19; 40, 0.25) = 0.0017$), and the lexical zone is uninformative ($p_{\mathrm{syn}} = 0.40$). Fisher over the three channels:

$$X^2 = -2\big(\ln 0.039 + \ln 0.0017 + \ln 0.40\big) = 21.08, \qquad P = \Pr[\chi^2_6 > 21.08] = 0.0018.$$

Since $0.0018 \le \alpha/K = 0.0033$ and no other key comes close, the document is attributed to K_{qa}. Had the homoglyphs been stripped ($p_{\mathrm{char}} \to$ uninformative), the fused p-value would rise above the threshold and the system would abstain: evidence loss moves the outcome toward $\varnothing$, never toward another key.

C Pseudocode for the Cryptographic Channel

The attribution rule of Equation 4 is implemented by the loop of Algorithm 1; the per-key cryptographic p-value it invokes is computed by Algorithm 2.

Algorithm 1: SAFIZ detection and task attribution

Data : document X; candidate keys $\mathcal{K} = \{K_1, \ldots, K_K\}$; registry $\mathcal{R}$; re-anchoring window δ; level α
Result : attributed key $K_\star \in \mathcal{K}$, or $\varnothing$ (fail-closed)
for $K_i \in \mathcal{K}$ **do**
  $\mathcal{P} \leftarrow \emptyset$
  if $\mathcal{R}$ *has crypto compartments* **then** $\mathcal{P} \leftarrow \mathcal{P} \cup \{\texttt{CryptoP}(X, K_i, \mathcal{R}, \delta)\}$
  $(k_{\mathrm{chr}}, n_{\mathrm{chr}}, \hat{d}) \leftarrow$ matched / predicted homoglyph sites under K_i, and homoglyph density on other eligible positions
  if $n_{\mathrm{chr}} \ge n_{\min}$ **then** $\mathcal{P} \leftarrow \mathcal{P} \cup \{\texttt{SF}(k_{\mathrm{chr}}; n_{\mathrm{chr}}, \max(\rho, \hat{d}))\}$
  $(k_{\mathrm{syn}}, n_{\mathrm{syn}}) \leftarrow$ matched / checkable synonym slots under K_i (in zones)
  if $n_{\mathrm{syn}} > 0$ **then** $\mathcal{P} \leftarrow \mathcal{P} \cup \{\texttt{SF}(k_{\mathrm{syn}}; n_{\mathrm{syn}}, q_{\mathrm{syn}})\}$
  if $\mathcal{P} = \emptyset$ **then**
    $P(K_i) \leftarrow 1$
  else
    $X^2 \leftarrow -2 \sum_{p \in \mathcal{P}} \ln p$; $P(K_i) \leftarrow \mathrm{SF}_{\chi^2_{2|\mathcal{P}|}}(X^2)$
$\mathcal{F} \leftarrow \{K_i \in \mathcal{K} : P(K_i) < \alpha/K\}$
if $\mathcal{F} = \varnothing$ **then return** $\varnothing$
return $\arg\min_{K_i \in \mathcal{F}} P(K_i)$

Algorithm 2: Per-key cryptographic p-value $p_{\text{crypto}}(X, K_i)$

Procedure `CryptoP` $(X, K_i, \mathcal{R}, \delta)$ **:**

$\Pi \leftarrow \emptyset$ `// per-compartment p-values`

foreach *crypto compartment* $b = (c_s, c_e, \text{seed}_b, L_b)$ *in* $\mathcal{R}$ **do**

  $p_{\min} \leftarrow 1$

  for $d \leftarrow -\delta$ **to** δ **do**

    $\text{sub} \leftarrow X[c_s{+}d : c_e{+}d]$ `// re-anchor on character offsets`

    $\beta \leftarrow$ `bits(retokenize(sub))` `// bits re-derived from the text`

    $\ell \leftarrow \min(|\beta|,\ L_b)$; **if** $\ell < \ell_{\min}$ **then continue**

    $S \leftarrow 0$

    for $j \leftarrow 1$ **to** ℓ **do**

      $u \leftarrow F_{K_i}(b, \text{seed}_b, j)$; $v \leftarrow u$ **if** $\beta_j{=}1$ **else** $1-u$; $S \leftarrow S - \ln v$

    $p_{\min} \leftarrow \min\left(p_{\min},\ Q(\ell, S)\right)$ `// exact Gamma(ℓ,1) tail (Lemma 1)`

  if $p_{\min} = 1$ **then continue**

  $\Pi \leftarrow \Pi \cup \{\, \min(1,\ (2\delta{+}1)\, p_{\min}) \,\}$ `// union bound over the 2δ+1 shifts`

if $\Pi = \emptyset$ **then return** 1

$X^2 \leftarrow -2 \sum_{p \in \Pi} \ln p$

return $\text{SF}_{\chi^2_{2|\Pi|}}(X^2)$ `// Fisher over compartments`

Algorithm 2 realises Steps 1–4 for one candidate key: the outer loop walks the logged crypto compartments (Step 1); the inner shift loop re-tokenises and re-derives β at each anchoring, accumulating the score of Equation 1 (Step 2); each score is converted to its exact Gamma tail and the most significant shift is kept under the union-bound correction (Step 3); the surviving per-compartment p-values are Fisher-combined into the channel p-value (Step 4). $\text{SF}_{\chi^2_{2|\Pi|}}$ denotes the chi-square survival function of Equation 3 with $2|\Pi|$ degrees of freedom.

D Operating Envelope

Figure 2 renders the argument of Section 3.3 schematically: the region in which the cryptographic core can embed evidence, and the region – provably unreachable by any undetectable scheme – covered by the post-hoc channels.

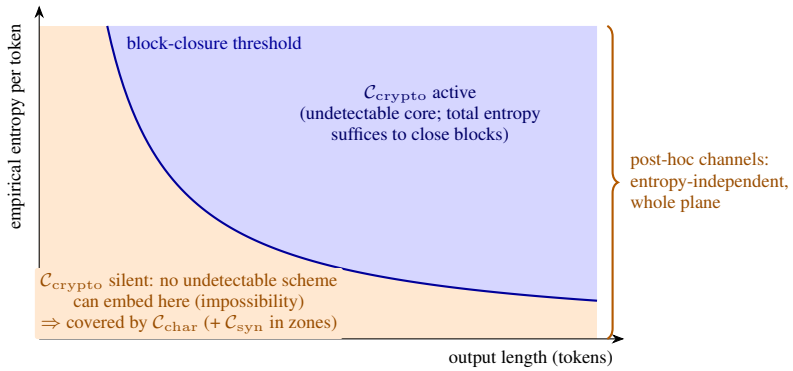

Figure 2: The operating envelope of SAFIZ (conceptual). The cryptographic core embeds evidence only where accumulated empirical entropy suffices to close blocks (above the threshold curve); below it, no undetectable scheme can embed at all (Christ et al., 2024), and attribution evidence is supplied by the entropy-independent post-hoc channels, which cover the whole plane. Composition extends coverage into the provably unreachable regime at the declared cost of system-level undetectability (Section 3.3).

E Additional Remarks

Remark 1 (Composition by disjointness or invertibility). Embedding operators interfere when one rewrites what another scores. SAFIZ avoids interference with two devices: *zoning* confines the non-invertible lexical channel to spans disjoint from the crypto blocks, and *invertibility* lets the bijective orthographic channel overlap them, its effect undone before scoring. Disjoint support, or inversion before scoring, is the general condition under which further channels can be added to any selective scheme without breaking the calibration of the existing ones.

Remark 2 (Domain separation). Including the compartment index b in the PRF input is not cosmetic: seeds seed_b are bit strings emitted by preceding blocks and may collide across compartments, in which case the inputs (seed_b, j) would repeat and the u_j of Equation 1 would no longer be independent across compartments, invalidating Lemma 1. Domain separation restores distinct inputs—with an unambiguous, length-prefixed serialisation of the tuple—and hence exact independence in the ideal model. The requirement has a second axis: separation *across documents*. The cryptographic core satisfies it by construction, because its seed is the bit string emitted by the document itself; a post-hoc channel whose PRF reads only the position index does not—its substitution mask becomes a constant of the key, identical on every document, so that N documents contribute a single effective observation to any cross-key statistic and the per-document freshness that the binomial null assumes is lost. Binding the post-hoc PRFs to local context (16 preceding normalised characters; 8 preceding base-form words) restores per-document freshness while preserving locality: the context looks backward only, so truncation still degrades gracefully.

Remark 3 (Two-tier verdicts). The rule of Equation 4 returns a key or $\varnothing$—never "machine-generated, task unknown". This is by construction: the natural binary test for "generated under *some* key in $\mathcal{K}$" rejects exactly when $\min_i P_{\mathrm{combined}}(X, K_i) < \alpha/K$, the same event that triggers attribution, so detection and attribution coincide under this rule (a coincidence our experiments confirm empirically). A coarser intermediate tier nonetheless exists without any key: the orthographic channel is flaggable by inspection (Section 3.3), so an anomalous density of Unicode confusables marks a text as carrying *some* SAFIZ signature while key-selectivity keeps *which one* protected. Evidence destroyed by edits thus degrades in two steps—attributed, then marked-but-unattributed under keyless screening, then unmarked—though the formal guarantees of this paper concern only the first tier; calibrating the screening tier is left to future work.

Remark 4 (Registry integrity). Soundness (Definition 2(a)) quantifies over registries fixed independently of the key set. An adversary who can *choose* the registry after seeing the text can search over compartment layouts and seeds until some fused p-value clears α/K—a chosen-registry attack that voids any statistical guarantee. We therefore require registries to be authenticated (a MAC or signature under a provider key); Verify rejects unauthenticated registries. Under an authentic registry, corruption of offsets degrades evidence and, by Proposition 2, yields abstention rather than misattribution.

Remark 5 (The registry is an optimisation, not a requirement). The cryptographic channel is detectable without any metadata: the detector of Christ et al. (2024) recovers blocks by scanning candidate (start, length) windows and testing each as a seed, at super-linear cost—up to $O(n^3)$ for a naive scan. SAFIZ remains compatible with this blind mode for the crypto channel. Logging the compartment structure replaces the search with a single linear pass, $O(n)$ per key and $O(Kn)$ over the candidate set, and additionally enables the zone structure that the lexical channel and the role-aware fusion require. We refer to the fallback as the *registry-free mode*: cryptographic detection alone, at super-linear cost, with the zone structure and the lexical channel disabled. The registry therefore buys a $O(n^3) \to O(n)$ verification and the zoned composition; what is lost without it is speed and the lexical channel, not detectability of the core. Section 4.4 measures what the post-hoc channels achieve without a registry; the blind detection of the core is not run in this paper.

Interpretation of key-selectivity

Remark 6 (Interpretation). Without the matching key, every channel collapses to its null, the fused statistic of Equation 3 is indistinguishable from noise, and the attributor abstains except on an α-fraction of documents by chance. The signature is therefore invisible to anyone who does not hold the correct task key: this is the key-selectivity that defines Selective Watermarking.

Table 2: System parameters, their effect, and how each value was chosen. None affects soundness (Theorem 1); the structural ones trade coverage among channels, the detection ones trade power against conservativeness.

Parameter	Value	Role	Choice
λ	256	key length (security parameter)	HMAC-SHA256 output size
α	0.01	family-wise false-attribution target	conventional audit level
C_{budget}	3	compartments guaranteed to the cryptographic channel before surplus is ceded	fixed when adaptive zoning was introduced, from the observed distribution of compartments; structural
ϕ_{zone}	0.5	fraction of surplus compartments promoted to synonym zones	even split of the surplus, fixed with C_{budget}; structural
$\ell_{\min}$	48	minimum scored bits for a compartment to vote	inherited from the Gaussian approximation the exact Gamma null replaced; kept because compartments are far longer than the threshold (Section F.7); structural
δ	1	re-anchoring window in characters	absorbs one-character drift at cost $(2\delta{+}1)$ on the p-value; detection-side
ρ	0.25	target homoglyph substitution rate	also the binomial null of the channel: higher ρ gives more carriers but a more permissive null, and $\rho^5 < \alpha/K$ makes five predicted positions sufficient on short text
$n_{\min}$	5	minimum predicted positions for the orthographic channel to vote	smallest n at which a perfect match (ρ^n) can clear α/K; detection-side
q_{syn}	0.409	lexical null collision probability	mean inverse class size of the dictionary
context	16 chars / 8 words	PRF context for the orthographic / lexical channels	per-document freshness (Remark 2)

THE INFORMED-AUDIT REGIME IN FULL

Two properties of this regime determine what the auditor can and cannot do. First, verification is *model-free*: the generating model $\mathcal{M}$ is never invoked at detection, so audits are cheap and reproducible. Second, and crucially, the registry is used only to locate compartments and to index the pseudo-random function; the bit-level evidence itself is *never used for the verdict*. At detection, every scored bit is re-derived from the suspect text X alone. The decision is therefore a function of X, which is precisely why it remains responsive to tampering: an adversary who edits the text changes the re-derived evidence. The registry fixes *where* to look and which seed indexes the pseudo-random function; it does not encode the mark, which is inaccessible without the matching key. We summarise this as metadata-assisted, text-grounded verification: the auditor is given structure, not secrets.

PARAMETERS

Table 2 lists every parameter of the system with its role and the rule by which its value was fixed. All values were set before the campaign and were not adjusted on the results. Only $n_{\min}$ and δ act purely at detection, so they can be swept on the stored generations without re-running any model (Section F.7); $\ell_{\min}$, C_{budget} and ϕ_{zone} enter role assignment and therefore the zoning itself, so varying them would require regenerating the corpus.

ROLE ASSIGNMENT IN FULL

The deterministic role-assignment rule of Section 3.3, in full:

Table 3: Per-model breakdown of clean-text attribution: accuracy by task and by target length, and confusion-matrix rows (true task → counts). All off-diagonal counts are zero for every model; errors are abstentions.

Model		qa	summary	news	50 tok	100 tok	200 tok
Llama-3.1-8B-Instruct	acc.	99.8%	100.0%	100.0%	100.0%	100.0%	99.8%
	abstain	1	0	0			
Qwen2.5-7B-Instruct	acc.	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	abstain	0	0	0			
Mistral-7B-Instruct	acc.	100.0%	100.0%	99.6%	100.0%	99.6%	100.0%
	abstain	0	0	2			
Llama-3.1-8B (base)	acc.	98.9%	95.6%	98.4%	96.7%	98.2%	98.0%
	abstain	5	20	7			
Qwen3-8B-Base	acc.	99.6%	99.8%	98.7%	98.7%	100.0%	99.3%
	abstain	2	1	6			
GPT-2-medium	acc.	98.9%	97.1%	99.1%	98.7%	98.2%	98.2%
	abstain	5	13	4			

Table 4: Cryptographic coverage by task: mean scorable cryptographic blocks per document and fraction of documents with none. The entropy collapse concentrates on `summary` for every model except GPT-2-medium, and on 50-token outputs within each task (not shown).

	qa		summary		news	
Model	Blocks	Zero-block	Blocks	Zero-block	Blocks	Zero-block
Llama-3.1-8B-Instruct	3.00	29.3%	0.35	86.9%	3.46	25.3%
Qwen2.5-7B-Instruct	0.32	78.2%	0.02	97.6%	1.22	39.1%
Mistral-7B-Instruct	0.63	63.6%	0.00	100.0%	1.06	42.4%
Llama-3.1-8B (base)	2.47	8.9%	1.73	34.7%	1.73	24.2%
Qwen3-8B-Base	1.27	40.4%	0.16	89.6%	1.87	23.6%
GPT-2-medium	4.20	2.0%	4.25	4.4%	3.72	5.1%

- the initial seed compartment, sampled without watermark to seed the chain, is never scored and never becomes a synonym zone;
- the first C_{budget} scorable compartments retain the role `crypto` – a guaranteed cryptographic budget, so that on short inputs the cryptographic evidence coincides with the crypto-only configuration;
- compartments too short to vote – fewer than $\ell_{\min}$ scorable bits – and lying beyond the budget are ceded to the lexical channel, since they cannot contribute to the cryptographic quorum in any case;
- among the remaining *surplus* compartments, a fraction ϕ_{zone} is promoted to the role `syn`; candidates are ranked by their count of substitutable, dictionary-eligible words, ties broken by a key-seeded hash, both reproducible at detection.

F Additional Experimental Results

This appendix reports the per-model, per-task, and per-length breakdowns summarised in Section 4, together with the secondary experiments referenced there.

F.1 Full Attribution Breakdown

Table 3 reports the per-task and per-length accuracy for each model, with the abstention counts by true task; Table 4 disaggregates the cryptographic coverage of Table 1 by task. Of the 32 abstentions of Llama-3.1-8B (base), 11 are empty completion-style continuations (3 in `qa`, 8 in `summary`); GPT-2-medium has 8 empty documents among its 22. Off-diagonal confusion counts are zero for every model and every cell: no document is ever attributed to a task other than its own.

Table 5: The complete null-check ($K = 3$, $\alpha = 0.01$). "Rule" states what is scored in each trial and the bound the rate is compared with: the full attributor ("full", family-wise, α), a single key against the Bonferroni threshold ($\alpha/K = 0.33\%$), or the cryptographic core alone ("core", α). In (c) all four corruption modes yield zero. The three key-selectivity probes below the rule are not part of the 872,310 trials.

Family	Construction	Rule / bound	Trials	FP	Rate	Wilson 95% CI
(a) human text	CNN/DailyMail, AG News, WikiText, deduplicated; synthetic registry; six tokenizers	full, α	29,910	15	0.050%	[0.030%, 0.083%]
(b) random keys	each of the 8,100 marked documents against 100 fresh keys, one key per trial	single key, α/K	810,000	427	0.053%	[0.048%, 0.058%]
(c) corrupted registry	four modes, 8,100 trials each (see text)	full, α	32,400	0	0.000%	[0.000%, 0.012%]
Total (pooled)		mixed	872,310	442	0.051%	[0.046%, 0.056%]
Cross-key	other live task keys	full, α/K	16,200	5	0.031%	[0.013%, 0.072%]
Random keys	fresh keys	core, α	40,500	106	0.262%	[0.216%, 0.316%]
Adjacent keys	$K+1$: one or two bits from the true key	core, α	8,100	23	0.284%	[0.189%, 0.426%]

Table 6: Scaling in K: family-wise false-positive rate on the same corpus scored against nested key banks, with and without the Bonferroni correction. The uncorrected max-score rule exhibits exactly the false-detection inflation predicted by Fu & Russell (2025).

	Bonferroni (α/K)			Uncorrected (max score, α)	
K	FW-FPR	Misattr.	Accuracy	FW-FPR	Misattr.
3	0.062%	0	99.19%	0.284%	0
10	0.136%	0	98.99%	1.605%	1
30	0.111%	0	98.79%	4.963%	1
100	0.123%	0	98.63%	14.901%	2

F.2 False-Positive Rates: Structure of the Null-Check

Table 5 gives the composition of the 872,310 null trials of Section 4.3. In (a) and (c) each trial is one verdict of the full attributor ($K = 3$, $\alpha = 0.01$): in (a), on human text paired with a synthetic registry so that the cryptographic channel is exercised rather than silent, any attribution is false; in (c), on a watermarked document with a corrupted registry, only an attribution to a key other than the generating one is false. In (b) each trial scores one fresh key against the per-key threshold α/K, the event that would make the attributor accuse that key; the pooled rate in the last row of the table therefore mixes the two rules and is reported for completeness only. The 29,910 trials of (a) are 4,985 distinct texts scored under the six models' tokenizers, so they are not independent replicates; the Wilson intervals are computed on the trial count and should be read accordingly. The 15 false accusations on human text fall on the `news` key (9), `summary` (5) and `qa` (1); the 427 under random keys on `news` (186), `qa` (155) and `summary` (86). Under registry corruption the verdict remains correct in 8,027 / 8,030 / 8,029 / 8,027 of 8,100 documents for shifted offsets, random seeds, shuffled compartments and another document's registry respectively, the remainder being abstentions: the orthographic channel, which does not read the registry, carries the verdict. Within (a) the rate varies by source but stays far below α throughout: 1 accusation in 9,918 trials on CNN/DailyMail (0.010%), 10 in 9,996 on AG News (0.100%), 4 in 9,996 on WikiText (0.040%).

Table 7: Channel ablation on the full corpus ($N = 8{,}100$ per configuration). Misattributions across all ablated configurations total 6 of $56{,}700$ verdicts, within the α bound; the operational configuration has zero.

Configuration	Attribution	Misattr.
Full (crypto+char+syn)	99.19%	0
Crypto+char	99.19%	0
Char only	99.23%	0
Char+syn (registry-free)	99.23%	0
Crypto only	42.36%	3
Crypto+syn	42.37%	3
Syn only	0.21%	0

Table 8: The Unicode attack ladder ($N = 8{,}100$ per row). "Lookalikes" counts the homoglyphs surviving in the corpus. The targeted strip removes exactly the system's alphabet; the full UTS #39 map also rewrites characters the models generate spontaneously (curly quotes, dashes), damaging the cryptographic channel as well.

Attack	Attr.	50 tok	100 tok	200 tok	Misattr.	Abstain	Lookalikes
No attack	99.2%	99.0%	99.3%	99.2%	0	66	412,545
NFC / NFKC	99.2%	99.0%	99.3%	99.2%	0	66–67	412,545
NFD / NFKD	98.9%	98.8%	99.1%	98.9%	0	88–90	412,545
Targeted strip (system map)	37.7%	17.9%	39.7%	55.5%	0	5,046	0
UTS #39 map (1,535 pairs)	5.8%	2.3%	4.8%	10.3%	2	7,630	0

F.3 Scaling in the Number of Keys

Detection time is measured, not assumed: scoring one document against the full bank costs 12.9 / 42.4 / 126.7 / 417.2 ms at $K = 3/10/30/100$—4.2–4.3 ms per key per document, constant across two orders of magnitude in K, confirming the $O(Kn)$ scaling stated in Section 2 (single-core CPU; the per-key cost is dominated by the linear pass of Remark 5). The full grid of family-wise rates and accuracies underlying Table 6 is in the released results.

F.4 Per-Channel Ablations

Table 7 gives attribution for each channel subset on the full corpus. The orthographic channel alone matches the full system, the cryptographic core alone recovers the entropic minority, and the lexical channel is inert; the three misattributions in the crypto-only configurations are within the α bound.

F.5 Unicode Normalisation and Robustness Curves

Table 8 gives the Unicode attack ladder discussed in Section 4.5, with the per-length breakdown: under confusables stripping the surviving attribution grows with length (17.9 / 39.7 / 55.5% at 50/100/200 tokens for the targeted strip), because longer outputs close more cryptographic blocks.

Figures 3 and 4 give the per-model robustness curves behind Section 4.5. The six models behave alike on every axis: the curves differ by a few points on truncation and on character noise (GPT-2-medium, the most entropic model, retains 24.3% at 5% noise against 9–16% for the others, because its cryptographic blocks survive where the context-bound orthographic reads do not) and coincide on token-level edits and paraphrase, where no model retains more than 12% once the first sentence is rewritten. Per model and per condition, the misattribution rate never exceeds 0.3%.

The orthographic null mis-specification, and its fix. Before the adaptive null of Lemma 4, paraphrased documents produced misattribution rates of 3.5% and 3.7% on the two heavy paraphrase variants (237 and 250 of 6,750 each, on the five-model corpus available at the time)—confidence intervals excluding α, a genuine soundness violation of the implementation. Diagnosis attributed 139 of 140 inspected cases to the orthographic channel and identified the cause: in the triggering documents the measured homoglyph density among eligible positions averaged 0.80, against 0.25 on

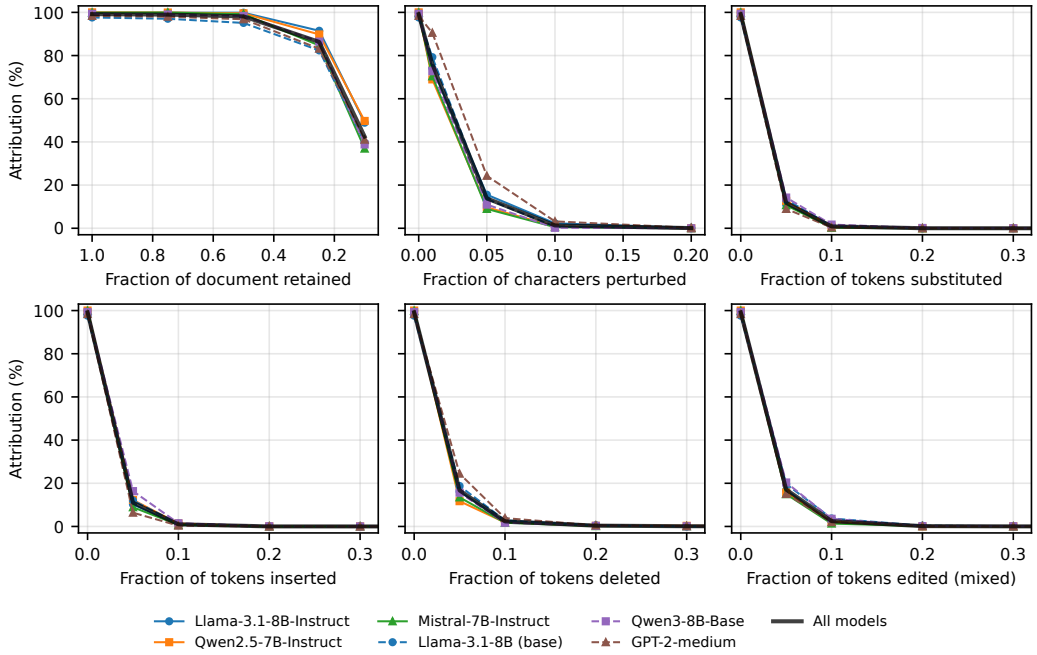

Figure 3: Robustness under signal removal, per model ($N = 1{,}350$ per model and condition). Solid lines: instruction-tuned models; dashed: base models; black: all $8{,}100$ documents. Token-edit axes are shown up to 30%; beyond it attribution is zero for every model.

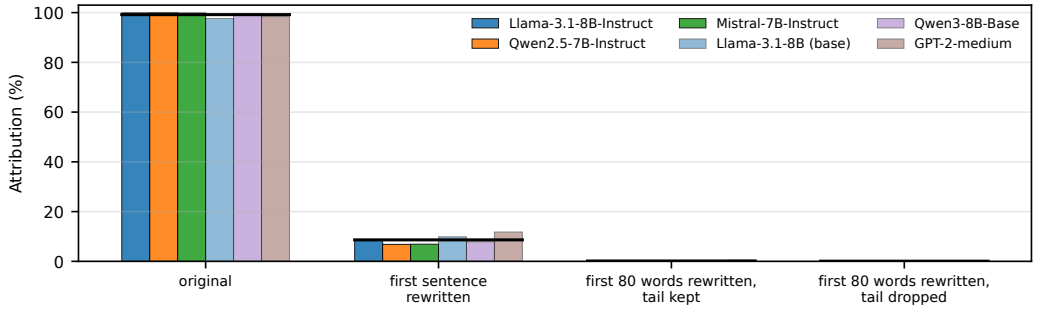

Figure 4: Paraphrase, per model: the four paired variants of Section 4.5. Black bars mark the six-model aggregate.

clean text and 0.10 on paraphrased text overall, so the fixed null $\rho = 0.25$ understated the collision probability by a factor of three. With $q = \max(\rho, \hat{d})$ the same trials yield 0 false triggers in both the calibration and the held-out split, and the end-to-end paraphrase misattribution rate falls to 12 of $32{,}400$ (0.037%), inside the bound; power on clean text is unchanged ($+0.00$ points overall and at 50 tokens).

F.6 Heterogeneous Channel Configuration

Table 9 gives the per-task figures for the asymmetric run of Section 4.4 (Llama-3.1-8B-Instruct, $N = 450$ per task, 150 per task and length). A null-check on the same corpus, with the $1{,}350$ documents each probed against 100 fresh keys, one key per trial at α/K, yields 76 accusations in

Table 9: Heterogeneous allocation: attribution by task and by length, with the zero-block fraction. All 405 failures (404 in `summary`, 1 in `news`) are abstentions.

Task	Channels	Accuracy	Zero-block	50 tok	100 tok	200 tok
`qa`	crypto+char	100.0%	30.7%	100.0%	100.0%	100.0%
`news`	crypto+char+syn	99.8%	21.6%	99.3%	100.0%	100.0%
`summary`	crypto+syn	10.2%	82.7%	2.7%	4.7%	23.3%

Table 10: Sensitivity of the detection-side parameters, one at a time, on the 8,100-document corpus. Failures are abstentions throughout, except one misattribution at $n_{\min} = 25$ on clean text. Defaults in bold.

	Clean		Truncated to 50%	
Value	Accuracy	Abstain	Accuracy	Abstain
$\delta = 0$	99.20%	65	98.15%	150
$\boldsymbol{\delta = 1}$	99.19%	66	98.15%	150
$\delta = 2$	99.19%	66	98.15%	150
$\delta = 3$	99.19%	66	98.15%	150
$\boldsymbol{n_{\min} = 5}$	99.19%	66	98.15%	150
$n_{\min} = 8$	98.78%	99	94.64%	434
$n_{\min} = 10$	98.35%	134	88.60%	923
$n_{\min} = 12$	97.89%	171	80.58%	1,573
$n_{\min} = 15$	96.41%	291	68.70%	2,535
$n_{\min} = 20$	90.74%	750	57.36%	3,454
$n_{\min} = 25$	82.00%	1,457	46.06%	4,369

135,000 trials (0.056%, Wilson $[0.045\%, 0.070\%]$): the per-key rate stays six times below its bound under an allocation that leaves one task without an entropy-independent channel.

F.7 Parameter Sensitivity

The two parameters that act purely at detection can be varied on the stored generations, with no model in the loop; Table 10 sweeps each in turn, the other held at its default, on the clean corpus and under truncation to 50%. Neither sweep produces a misattribution except at $n_{\min} = 25$, where one appears out of 8,100; soundness is insensitive to both, as Theorem 1 requires. The re-anchoring window is inert on this corpus: $\delta = 0$ and $\delta = 3$ give the same accuracy to within one document, and identical figures under truncation, because truncation displaces no compartment that survives the cut and the corpus contains no character-level drift. The default $\delta = 1$ therefore costs 0.01 accuracy points—the price of the $(2\delta{+}1)$ correction—and buys insurance against a perturbation these conditions do not exercise; the character-noise results of Section 4.5 are the regime where drift does occur, and there the loss is in re-deriving the PRF context, which a wider window does not recover. The orthographic quorum is the opposite case: accuracy falls monotonically as $n_{\min}$ rises, gently on clean text (99.2% to 82.0%) and sharply under truncation (98.1% to 46.1%), since a truncated document offers fewer predicted positions and a high quorum silences the channel exactly where it is the only carrier left. The default is the smallest value at which the binomial tail can clear α/K, and the sweep shows no reason to raise it.

F.8 Generation Quality

Undetectability implies that the cryptographic core preserves the text distribution exactly; we verify the consequence a practitioner cares about, namely that watermarked text is no worse than unwatermarked text under a third-party model. On a balanced subsample (10 documents per task and target length, 90 per model), we measure perplexity under OPT-1.3B (Zhang et al., 2022), a model outside the six under test, for three configurations generated from the same prompt and seed: no watermark, cryptographic core only, and the full system. The unwatermarked control is generated by the same bit-by-bit procedure over the binary encoding of the vocabulary, drawing from the full conditional distribution with a random source in place of the PRF, so that the comparison isolates the watermark

Table 11: Median perplexity under OPT-1.3B, interquartile range in brackets ($N = 90$ per model). "Equal ppl" counts documents on which core and unwatermarked configurations give exactly the same perplexity; "substituted" counts documents the lexical channel modified.

Model	No watermark	Core only	Full	Equal ppl	Substituted
Llama-3.1-8B-Instruct	30.3 [17.8, 73.3]	27.7 [16.1, 81.0]	27.7	36/90	9/90
Qwen2.5-7B-Instruct	16.2 [11.5, 22.5]	16.2 [12.1, 24.3]	16.2	75/90	0/90
Mistral-7B-Instruct	15.6 [12.8, 21.5]	16.0 [12.5, 22.7]	16.0	61/90	0/90
Llama-3.1-8B (base)	49.4 [30.1, 81.2]	45.8 [27.0, 80.4]	46.2	8/90	14/90
Qwen3-8B-Base	25.6 [17.0, 36.5]	25.6 [17.0, 35.2]	25.6	49/90	2/90
GPT-2-medium	94.3 [66.5, 181.2]	102.3 [65.6, 164.2]	103.3	3/90	25/90

rather than the sampling procedure. Perplexity is measured after normalisation, which inverts the orthographic map and restores the text a reader sees; we report medians with interquartile ranges, since the distribution is heavy-tailed and its mean is dominated by a few documents.

Table 11 gives the result. The core is indistinguishable from no watermark. Median perplexities differ by at most 9% and in both directions ($+8.4\%$ on GPT-2-medium, -8.7% on Llama-3.1-8B-Instruct, 0.0% on both Qwen models), a spread far narrower than the interquartile range of any single configuration; and the paired per-document difference is balanced in sign on every model—40 positive against 47 negative on GPT-2-medium, 25 against 29 on Llama-3.1-8B-Instruct, 15 against 14 on Mistral-7B-Instruct—so a two-sided sign test rejects on none of the six (p between 0.30 and 1.00). On a large share of documents the two configurations also yield exactly equal perplexity ($75/90$ on Qwen2.5-7B-Instruct, $61/90$ on Mistral-7B-Instruct, $49/90$ on Qwen3-8B-Base), which is what inversion sampling predicts where the conditional is near-deterministic and the sampled bit does not depend on the random value; we did not retain the generated strings, so whether those documents are textually identical is not verified here. The lexical channel is the only component that changes the words, and it rarely acts: it substitutes nothing at all on Mistral-7B-Instruct and Qwen2.5-7B-Instruct within this subsample, and on the remaining models it touches 2 to 25 documents out of 90. On the documents it does touch, the median paired cost is $+1.5$ to $+11.9$ perplexity points, with a single outlier of $+197$; at the level of corpus medians this is $+1.05\%$ on GPT-2-medium, $+0.76\%$ on Llama-3.1-8B (base), and zero on the other four. A channel that is inert in the ablation (0.21% attribution alone, Section 4.4), rarely fires, and is the only one to move perplexity when it does is a candidate for removal, which we note in Section 5.

Measuring the same texts *without* normalising first gives figures that move erratically and in both directions (from 60 against a canonical 103 on GPT-2-medium to 30 against 16 on Mistral-7B-Instruct), because homoglyphs fall outside the oracle's subword vocabulary and are tokenised as byte sequences. Those numbers describe the oracle's tokeniser, not the text, which the orthographic map leaves visually unchanged; they are a real cost for any downstream model consuming marked text, and no part of the quality claim.