COHORTMARK: WATERMARKING TEXT COHORTS WHILE PRESERVING EVERY PROPER SUBSET

Anonymous authors
Paper under double-blind review

ABSTRACT

Text watermarks seek detectable signals with limited distributional change. Preserving individual distributions need not preserve subgroup laws; preserving the complete group's clean law conditional on key and metadata leaves no evidence. Our insight is to separate observation scopes: preserve every proper subset while placing evidence in the original group's higher-order relation. Token discretization can erase this dependence. CohortMark uses an analysis of this loss to design a finite-harmonic coupling whose shape controls evidence without altering subset laws. Source-model replay with required metadata gives the original group's exact token-interval likelihood ratio. With ideal exact sampling, token probabilities based only on each member's own prompt and history, and fresh joint draws per step, all predesignated nonempty proper subsets follow independent clean sequence laws at each fixed key and nonce. Across two model families and two tasks, the three-member instance achieves 94.8–99.0% recall at separately calibrated nominal 1% false-positive rates. All four settings meet predeclared recall noninferiority and false-positive criteria against SynthID; both Llama settings support higher recall. Exchanging one member at the same three-text, 768-token budget reduces positive rates to 1.0–1.2%, linking evidence to original joint generation. In automatic ELI5 evaluation, intervals for changes from clean in the fraction of groups with at least one usable answer lie within ± 3 percentage points.

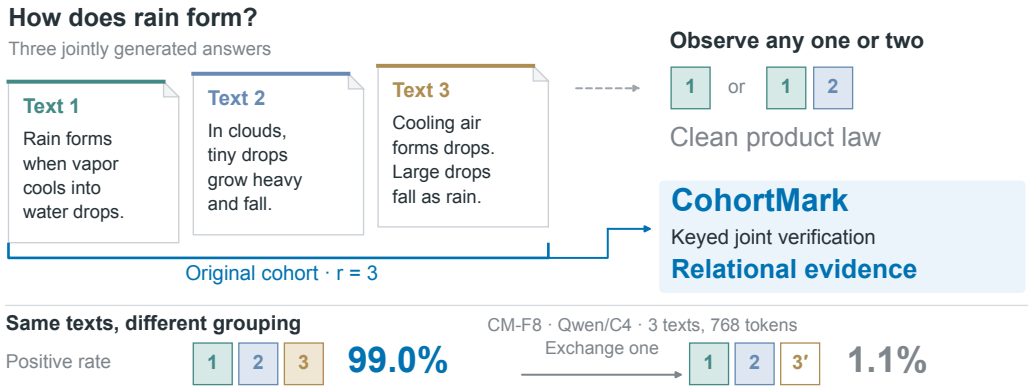


Figure 1: CohortMark separates observation scopes: each individual answer and each predesignated pair retain the clean product sequence law at each fixed key and nonce under Theorem 2, while the original jointly generated triple can carry detection evidence. Texts are illustrative. The lower panel compares original and regrouped observations with three texts each and 3,000 groups per condition; $3'$ comes from another original group (Section 4.3).

1 INTRODUCTION

Text watermarking embeds statistical signals during generation to verify provenance. Existing methods alter token preferences or couple tokens with secret random variables while considering distributional effects (Kirchenbauer et al., 2023; Hu et al., 2024; Dathathri et al., 2024; Tsur et al., 2025).

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.

Across multiple outputs, preserving individual distributions need not preserve similarities or joint-event probabilities. Yet matching the complete group's joint law to independent clean generation conditional on the key and generation metadata leaves no statistical difference for detection. How can proper subsets retain their reference distributions while the complete group carries provenance evidence?

Proper-subset joint distributions need not determine the full joint distribution. This motivates preserving the former while seeking detection evidence in the latter. For three jointly generated answers, each answer and each predesignated pair can retain its independent clean reference distribution while the original triple can carry a signal (Figure 1). Subset-based quality and similarity statistics would then retain their reference distributions. Complete-group verification requires metadata and source-model access.

Low-order-independent structures and threshold steganography provide foundations for this separation (Durante, 2012; Alston, 2017), but dependence among continuous sampling variables need not survive token discretization. Sampling partitions the unit interval according to token probabilities; a random draw selects the token whose interval contains it. Observing a token reveals only that interval. Averaging over these intervals can erase density differences, giving every token tuple its independent clean probability. We must therefore preserve subset distributions while retaining dependence in discrete tokens.

CohortMark separates the sampling structure preserving subset distributions from the density shape controlling complete-group evidence. It couples sampling variables through periodic dependence while preserving each member's token probability mapping. Integrating a missing member's variable over the full unit interval removes dependence, leaving the remaining members independent; integrating over the complete group's token intervals gives its joint token probability. Analysis of how interval averaging attenuates different frequencies guides a finite-harmonic kernel that strengthens discrete relational evidence while preserving subset distributions.

Our contributions are:

1. **Proper-subset sequence guarantees.** Under ideal exact sampling, token probabilities based only on each member's own prompt and history, and fresh joint draws per step, we prove that every predesignated nonempty proper subset follows independent clean generation at the sequence level for general group sizes and each fixed key and nonce.
2. **Discretization-guided design and detection.** Token-interval analysis guides the finite-harmonic kernel and gives the original complete group's exact likelihood ratio, computed through source-model replay with the correct key, nonce and required metadata.
3. **Relational evidence and detection performance.** Across two model families and two tasks, the three-member instance achieves 94.8–99.0% recall at separately calibrated nominal 1% false-positive rates. Exchanging one member at the same three-text, 768-token budget reduces positive rates to 1.0–1.2%, supporting dependence on the original relation. All four settings meet predeclared recall noninferiority and false-positive criteria against SynthID; both Llama settings support higher recall.

2 Related Work

CohortMark connects distribution-preserving text watermarking, higher-order dependence and the analysis of evidence surviving discrete token observation.

Text watermarking and distribution preservation. Token-partition watermarks upweight selected tokens (Kirchenbauer et al., 2023; Zhao et al., 2024). Unbiased Watermark, SynthID and HeavyWater/SimplexWater study distribution preservation through dependence between tokens and secret random variables (Hu et al., 2024; Dathathri et al., 2024; Tsur et al., 2025). Preservation averaged over watermark randomness differs from a fixed-key joint law across texts. ReasonMark adapts the carrier to reasoning representations, while PatternMark uses Markov key patterns for order-agnostic generation (Liu et al., 2026; Chen et al., 2025). Detection tasks also differ: black-box auditing queries whether a service uses a watermark (Gloaguen et al., 2025); CohortMark verifies an original complete group using the correct key, nonce, metadata and source-model replay.

Higher-order dependence and joint text generation. Copulas describe dependence with specified marginals. Low-order-independent constructions keep subsets independent while allowing full-group dependence (Durante, 2012); our periodic coupling belongs to this class for $r \geq 3$. Threshold steganography splits a secret into shares whose subthreshold collections are uniformly distributed for a fixed secret (Alston, 2017). Under our sampling assumptions, classical Farlie–Gumbel–Morgenstern (FGM) constructions and threshold shares also yield subset sequence guarantees and token likelihoods after specialization to discrete tokens (Appendix C.2). Other joint samplers target candidate coverage, as in QuasiMoTTo (Li et al., 2026), or diffusion-output dependence, as in SAC-Copula (Li & He, 2026). Preserving individual marginals alone does not guarantee preservation of all proper-subset joint laws.

Evidence and detection in discrete tokens. QuantileMark derives message posteriors from overlaps between token cumulative-probability intervals and message bins, aggregating source-model evidence for multibit recovery and detection (Zhu et al., 2026). Work on speculative sampling quantifies watermark strength by the KL divergence between marked and original token distributions conditional on a pseudorandom variable, averaged over that variable, and relates it to acceptance rates (He et al., 2026). Multibit analysis also distinguishes content from secret information under false-alarm constraints (Huang et al., 2026). These approaches study token associations with messages or secret variables; relational detection instead examines joint probabilities across members' token intervals.

CohortMark uses existing higher-order structures to preserve proper-subset sequence laws under the stated assumptions. Its token-interval analysis guides kernel design and measures the evidence retained in the original complete group's discrete outputs. Appendix E discusses further carrier designs and evaluation settings.

3 CohortMark: Higher-Order Relational Watermarking

3.1 Problem Setup and Observation Interfaces

The generator jointly produces a designated group $X = (X_1, \ldots, X_r)$ of $r \geq 2$ length-T continuations with possibly different prompts h_i. Each member's token probabilities $p_{i,t}(\cdot \mid h_i, X_{i,<t})$ depend only on its own prompt and history, after the declared temperature, truncation and end-of-sequence (EOS) transforms. Independent clean sampling gives

$$P_0(X \mid \mathcal{M}) = \prod_{i=1}^{r} P_i(X_i \mid h_i), \qquad P_i(X_i \mid h_i) = \prod_{t=1}^{T} p_{i,t}(X_{i,t} \mid h_i, X_{i,<t}). \tag{1}$$

Fixed metadata $\mathcal{M}$ specify prompts, key K, nonce ν, membership and ordering, finite length, model and decoding protocol, including finite vocabulary order. They reveal neither sampling variables nor additional functions of missing outputs.

For each fixed key and nonce, the marked law $Q_{K,\nu}$ must equal $\bigotimes_{i \in A} P_i$ on every predesignated nonempty proper subset $A \subsetneq [r] = \{1, \ldots, r\}$, while the complete-group law may differ. Subsets are chosen before inspecting outputs. The detector receives the original jointly generated group and $\mathcal{M}$ and can replay the source model; a partial observer receives only its subset and the same $\mathcal{M}$. Subset preservation therefore does not rely on key secrecy.

3.2 Joint Sampling and Proper-Subset Preservation

Given current histories, inverse cumulative distribution function (CDF) sampling partitions $[0, 1)$ in vocabulary order into intervals of lengths $p_i(v)$ and selects the token whose interval contains a uniform draw. CohortMark keeps these mappings and couples the draws $U_1, \ldots, U_r$.

For a nonnegative, one-periodic density f with $\int_0^1 f(s)\, ds = 1$, define

$$q_c(u) = f\left(\left[\sum_{i=1}^{r} u_i - c\right] \bmod 1\right), \qquad u \in [0, 1)^r. \tag{2}$$

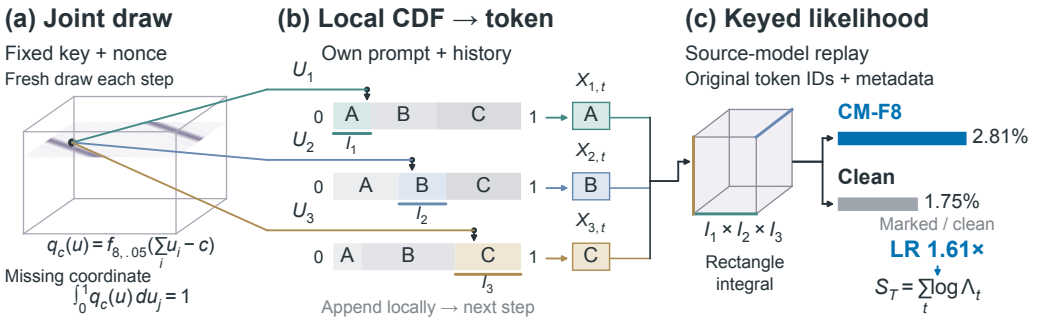


Figure 2: Generation and verification at $r = 3$. Coupled draws map to tokens through member-local probability intervals. Source-model replay with correct metadata reconstructs the intervals for the joint likelihood ratio. Probabilities illustrate one analytic step of the main instance (Section 3.4).

The phase c shifts the density pattern; $c_t = \phi_K(\nu, \mathcal{M}, t)$ uses only fixed metadata and position, allowing replay without sampled tokens.

Proposition 1 (One-step proper-subset preservation)**.** *For every phase c, Equation (2) is a density whose nonempty proper subsets have the uniform product law, since integrating out any coordinate traverses one period:*

$$\int_0^1 q_c(u)\,du_j = \int_0^1 f(s)\,ds = 1. \tag{3}$$

Each member's inverse CDF therefore gives $Q_c(X_A = v_A) = \prod_{i\in A} p_i(v_i)$ for every nonempty $A \subsetneq [r]$, conditional on current histories.

To sample q_c, independently draw $V_1, \ldots, V_{r-1} \sim \mathrm{Unif}[0, 1)$ and a scalar W with density f, then set

$$U_i = V_i \quad (i < r), \qquad U_r = \left(c + W - \sum_{i<r} V_i\right) \bmod 1. \tag{4}$$

Translation modulo 1 gives the required density; each member maps U_i to its next token.

Theorem 2 (Fixed-key proper-subset sequence preservation)**.** *Fix $\mathcal{M}$, finite length T and a designated r-member group. Suppose token probabilities depend only on each member's own prompt and history, Equation (2) is sampled with fresh conditional randomness at every position, and phases are fixed or predictable before the current draw. In the ideal exact construction, for every fixed key and nonce and every predesignated nonempty proper subset $A \subsetneq [r]$,*

$$Q_{K,\nu}(X_A = x_A \mid \mathcal{M}) = \prod_{i\in A} P_i(x_i \mid h_i). \tag{5}$$

Conditioning on all histories gives the one-step product law. Its factors depend only on retained histories, so averaging over missing histories leaves it unchanged; multiplication across time proves Equation (5) (Appendix B). Every measurable statistic of a designated proper subset and fixed metadata, including keyed statistics and subset quality evaluations, consequently retains its clean reference distribution.

3.3 Token Intervals and Likelihood-Based Detection

A token reveals its interval, not the exact draw. For $X_{i,t}$, let $I_{i,t} = [a_{i,t}, b_{i,t})$ have width $w_{i,t} = b_{i,t} - a_{i,t}$ (its probability) and center $\mu_{i,t} = (a_{i,t} + b_{i,t})/2$. Conditional on histories, the marked tuple's probability integrates q_{c_t} over its token rectangle; the clean probability multiplies widths. Thus,

$$\Lambda_t = \frac{\displaystyle\int_{\prod_i I_{i,t}} q_{c_t}(u)\,du}{\displaystyle\prod_i w_{i,t}}.$$

Unlike integrating out a missing coordinate in Equation (3), restricting every coordinate to its token interval can retain dependence.

Use a finite-harmonic density $f(s) = 1 + \sum_{k=1}^{J} b_k \cos(2\pi k s)$, with real weights b_k chosen in Section 3.4. For $J \geq 1$ and $0 \leq \epsilon \leq f \leq M < \infty$, exact integration gives

$$\Lambda_t = 1 + \sum_{k=1}^{J} b_k \underbrace{\prod_{i=1}^{r} \operatorname{sinc}(\pi k w_{i,t})}_{\text{interval averaging}} \underbrace{\cos\left(2\pi k \left[\sum_i \mu_{i,t} - c_t\right]\right)}_{\text{phase}}, \tag{6}$$

where $\operatorname{sinc}(z) = \sin(z)/z$ and $\operatorname{sinc}(0) = 1$. Widths control averaging; centers and phase control the cosine. Higher-harmonic sinc factors can be negative and retain their signs. Averaging f gives $\epsilon \leq \Lambda_t \leq M$.

The conditional ratios multiply along the original group's histories, giving the log score $S_T = \sum_{t=1}^{T} \log \Lambda_t$.

Theorem 3 (Complete-group likelihood)**.** *Under the preceding generation assumptions, a reconstructible phase rule and $\epsilon > 0$, $L_T = e^{S_T} = dQ_{K,\nu}/dP_0$ on the common sequence support, and*

$$\mathbb{E}_0 L_T = 1, \qquad \mathbb{E}_Q S_T = D_{\mathrm{KL}}(Q\|P_0) \geq 0, \qquad \mathbb{E}_0 S_T = -D_{\mathrm{KL}}(P_0\|Q) \leq 0. \tag{7}$$

Here Q denotes the marked sequence law at fixed key, nonce and metadata; 0 denotes the clean reference. Positive expected evidence thus requires dependence to survive in the discrete token distribution. Appendix B gives the proof and likelihood-martingale bound. Detection uses independent clean calibration and reports held-out false positives.

Detection requires original token IDs in member order, source-model replay with the same vocabulary and decoding transforms, and phases reconstructed from $\mathcal{M}$. Replay and synchronization incur access and computational costs; retokenization requires separate treatment (Yan & Murawaki, 2025).

3.4 Discretization Analysis and Kernel Design

Equation (6) exposes two losses. A deterministic token has width 1, erasing every nonconstant harmonic. With three partitions $(0.5, 0.5)$ and $c = 0$, phase cancellation zeros odd harmonics' cosine factors, while averaging zeros even harmonics' sinc factors: every tuple has $\Lambda = 1$.

For fixed token probabilities, let Q_c and P_0 be the marked and independent clean joint token laws. Their χ^2 divergence measures the distribution difference. Averaging over a continuous uniform phase removes cross terms and separates the harmonics:

$$\mathbb{E}_c \chi^2(Q_c\|P_0) = \frac{1}{2}\sum_{k=1}^{J} b_k^2 R_k, \qquad R_k = \prod_i \sum_v p_i(v) \operatorname{sinc}^2(\pi k p_i(v)). \tag{8}$$

Here R_k measures harmonic k's survival through interval averaging. Design should strengthen surviving components while retaining a nonnegative, normalized, exactly integrable density. This phase average guides design; Equation (6) determines evidence at fixed keys and phases.

A single cosine density $1 + \rho\cos(2\pi s)$ requires $|\rho| \leq 1$. We instead combine harmonics through the nonnegative Fejér square and add a uniform component. For integer $J \geq 1$ and $0 \leq \epsilon < 1$,

$$f_{J,\epsilon}(s) = \epsilon + \frac{1-\epsilon}{J+1}\left|\sum_{a=0}^{J} e^{2\pi i a s}\right|^2 = 1 + \sum_{k=1}^{J} b_k \cos(2\pi k s), \tag{9}$$

$$b_k = 2(1-\epsilon)\left(1 - \frac{k}{J+1}\right), \qquad M_{J,\epsilon} = \epsilon + (1-\epsilon)(J+1). \tag{10}$$

The harmonics integrate to zero and $\epsilon \leq f_{J,\epsilon} \leq M_{J,\epsilon}$, preserving the subset guarantee. Increasing J adds harmonics and raises existing weights, strengthening components surviving token averaging. Positive ϵ bounds the likelihood away from zero.

The cosine baseline $(J, \epsilon) = (1, 0)$ still averages strictly positive on every positive-width token rectangle. With finite vocabulary and length it shares the clean law's finite support, so Equation (7) applies despite its zero density floor. At floor 0.05, degrees $1, 4, 8$ define F1, F4 and F8. The prespecified CM-F8 instance $(8, 0.05)$ has bounds 0.05 and 8.6 and first-harmonic weight 1.6889, versus cosine's 1.

Changing the balanced binary example's phase restores signal; an asymmetric partition changes the response (Figure 3; derivation in Appendix C.1).

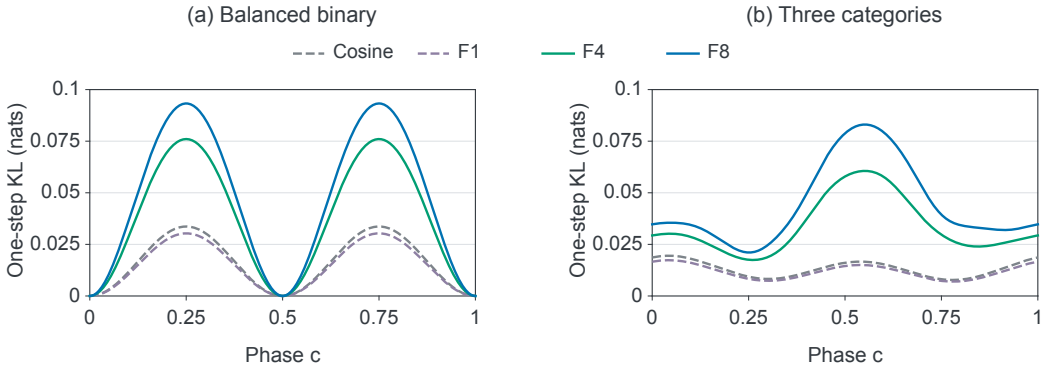


Figure 3: Exact one-step KL versus phase at $r = 3$, with each member's token probabilities in CDF order (a) $(0.5, 0.5)$ or (b) $(0.7, 0.2, 0.1)$. At $c = 1/4$ in (a), cosine and F8 yield 0.033665 and 0.093283 nats. F1/F4/F8 have floor 0.05; cosine has floor zero.

Identical subset guarantees can yield different token evidence. For three members with $p = (0.7, 0.2, 0.1)$ in CDF order, the FGM family $q_\theta(u) = 1 + \theta \prod_i (1 - 2u_i)$, $|\theta| \leq 1$ (Durante, 2012), attains at most 1.36 nats of expected S_{256} over admissible per-step strengths. CM-F8 gives 11.38 nats with $c_t = (t + 0.5)/256$, $t = 0, \ldots, 255$ (Appendix C.2, Figure 14).

Sample W by accepting uniform proposals w with probability $f_{J,\epsilon}(w)/M_{J,\epsilon}$. Normalization gives mean acceptance $1/M_{J,\epsilon}$ and $M_{J,\epsilon}$ expected scalar proposals per step; retries add no model forward passes.

4 Experiments

We test subset behavior and complete-group detection, then use member exchange and kernel comparisons to identify the evidence carrier and design gains. Group utility and text quality assess the cost of relational dependence.

4.1 Experimental Setup

Qwen3-8B and Llama-3.1-8B base generate C4 continuations and ELI5 answers. Per input, each method generates three same-prompt candidates with member-local histories and 256 tokens each (768 total), at temperature 0.8, top-$p = 0.95$ and masked EOS. Independent clean generation, CM-cos, CM-F8, SynthID and KGW (Dathathri et al., 2024; Kirchenbauer et al., 2023) share confirmation inputs and budgets. CM-F8 uses predeclared key A, $(J, \epsilon) = (8, 0.05)$ and exact-likelihood replay. SynthID and KGW aggregate native evidence across members, retaining separate histories and native valid-position rules (Appendix D).

Each model/task uses 300 development inputs, $4{,}000$ clean calibration groups and $3{,}000$ confirmation groups per method. Detectors are separately calibrated to nominal 1% FPR; thresholds, configuration and key are frozen before confirmation. Scores strictly above threshold are positive; held-out clean groups measure actual FPR.

Paired comparisons use prompts as units; exchanges use two-prompt blocks. Marginal rates use Wilson 95% intervals. AUROC, means and paired differences use normal reference intervals. Native

AUROC and mean-score intervals are pointwise 95%; subset diagnostics and designated comparison families use Bonferroni simultaneous 95% intervals, with families specified below. Appendix D.3 gives formulas and threshold ranks.

4.2 Proper-Subset Diagnostics and Complete-Group Detection

Single- and two-member diagnostics perform near chance. Logistic classifiers trained on development inputs use only visible members' token and probability features, excluding missing members and complete-group scores. After freezing, each distinguishes 3,000 clean from 3,000 CM-F8 groups. The 24 AUROCs (six subsets per configuration) are 0.494–0.507, with all simultaneous intervals within $[0.46, 0.54]$ (Appendix Table 4). These finite-feature diagnostics check implementation behavior; the exact distribution identity follows Theorem 2.

Complete-group CM-F8 detection reaches 94.8–99.0% recall with 0.967–1.100% observed FPR (Table 1). Mean log-likelihood scores are 9.3–11.4 nats for marked groups and -10.6 to -8.7 for clean groups. All four configurations meet the predeclared recall noninferiority and FPR criteria against SynthID. Paired recall intervals support higher recall in both Llama settings and span zero in both Qwen settings.

Table 1: Complete-group detection: 3,000 marked and 3,000 clean groups per model/task, three texts and 768 tokens, separately calibrated nominal 1% FPR. CM-F8 requires source-model replay. Bold marks largest AUROC/TPR point estimates. Δ rows give CM-F8 minus SynthID (pp) with simultaneous 95% reference intervals: separate four-comparison Bonferroni families for recall and FPR, with predeclared bounds -3 and $+0.5$ pp, respectively.

C4	Qwen			Llama		
Method	AUROC ↑	TPR (%) ↑	FPR (%)	AUROC ↑	TPR (%) ↑	FPR (%)
CM-cos	0.955	53.0	1.03	0.945	54.2	1.00
CM-F8	**0.998**	**99.0**	0.97	**0.996**	**97.6**	1.07
SynthID	0.997	98.8	1.07	0.991	96.4	1.00
KGW	0.989	95.8	1.10	0.982	91.6	0.97
Δ TPR (pp)		0.20 $[-0.41, 0.81]$			1.20 $[0.46, 1.94]$	
Δ FPR (pp)		-0.10 $[-0.48, 0.28]$			0.07 $[-0.29, 0.42]$	
ELI5	Qwen			Llama		
Method	AUROC ↑	TPR (%) ↑	FPR (%)	AUROC ↑	TPR (%) ↑	FPR (%)
CM-cos	0.933	46.8	1.03	0.920	43.4	1.13
CM-F8	**0.995**	**97.4**	1.00	**0.991**	**94.8**	1.10
SynthID	0.994	97.0	1.07	0.985	93.0	1.17
KGW	0.980	90.2	1.13	0.966	84.8	1.07
Δ TPR (pp)		0.40 $[-0.35, 1.15]$			1.80 $[0.70, 2.90]$	
Δ FPR (pp)		-0.07 $[-0.44, 0.31]$			-0.07 $[-0.44, 0.31]$	

Predeclared key B gives Qwen/Llama recall of 98.4/97.6% and FPR of 1.2/1.1% on 1,000 C4 confirmation inputs per model, using new key-B outputs and the corresponding clean groups.

4.3 The Role of the Original Joint Generation

Exchanging one member reduces CM-F8 positive rates to 1.0–1.2% at the same three-text, 768-token budget (Table 2). Before observing outputs, we form 1,500 disjoint two-prompt blocks and reciprocally exchange member three. Transferred texts retain their own prompt, history and replay intervals; detection uses the recipient group's key, nonce, phases and original threshold.

Block-based lower bounds for all four reductions exceed 92 percentage points (Appendix Table 4). SynthID and KGW change by at most 0.2 points: their texts retain watermark signals, whereas CM-F8's evidence depends on the original relation. Exchanged-group positivity is distinct from clean FPR. Wrong-key or wrong-nonce rescoring gives 30–36 positives per 3,000 original groups; these metadata controls complement member exchange (Appendix B.5).

Table 2: Positive rates (%) before and after exchanging one member. Each configuration contains 3,000 groups per condition in 1,500 paired blocks, with three texts and 768 tokens per group. Drop is original minus exchanged, in percentage points.

	CM-F8			SynthID		KGW	
Configuration	Orig.	Exch.	Drop (pp)	Orig.	Exch.	Orig.	Exch.
Qwen/C4	99.0	1.1	97.9	98.8	98.7	95.8	95.7
Llama/C4	97.6	1.0	96.6	96.4	96.2	91.6	91.5
Qwen/ELI5	97.4	1.2	96.2	97.0	97.1	90.2	90.1
Llama/ELI5	94.8	1.1	93.7	93.0	92.9	84.8	84.7

4.4 Effects and Costs of Kernel Design

CM-F8 improves recall over cosine in all four settings, with positive paired lower bounds in a four-comparison family. At fixed floor $\epsilon = 0.05$, predeclared degrees $J = 1, 4, 8$ test the discretization-guided design on Qwen/C4, with new outputs and separate clean calibration per kernel (Figure 4). Successive paired recall gains are 42.4 $[40.336, 44.464]$ and 5.6 $[4.574, 6.626]$ percentage points in one two-comparison family. Adding harmonics and raising existing weights strengthens retained evidence while preserving the same subset structure. The zero-floor cosine has slightly higher recall and mean-score point estimates than F1.

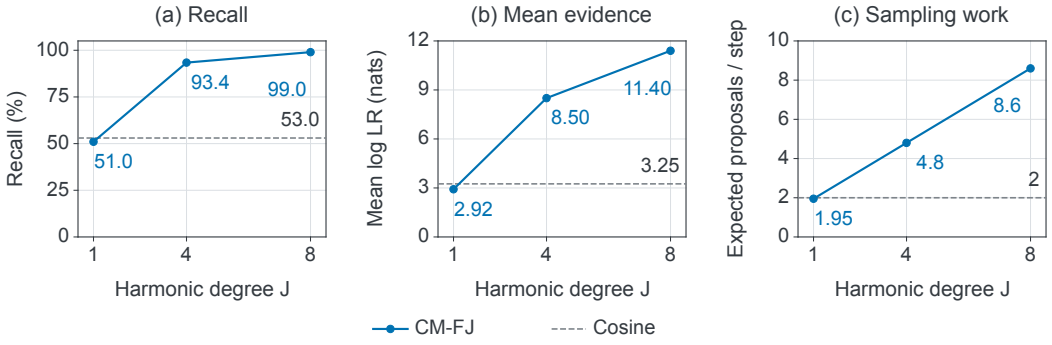


Figure 4: Qwen/C4 detection and sampling work. Solid lines compare degrees at floor 0.05; dashed lines show zero-floor cosine. Recall and mean score use 3,000 new marked groups per kernel with separate clean calibration. Panel (c) gives analytic expected scalar proposals per step.

Higher degree increases scalar sampling work. Retries add no model forward passes, but F8 evaluates eight harmonics and requires source-model replay. Proposal expectations do not measure end-to-end runtime.

4.5 Text Quality and Group Utility

Individual and pairwise preservation does not fix the probability of at least one usable answer. For fixed member-local binary usability labels Y_i, inclusion–exclusion under Theorem 2 gives

$$\Delta \Pr(\max_i Y_i = 1) = \Delta \Pr(Y_1 = Y_2 = Y_3 = 1), \tag{11}$$

where Δ is marked minus clean: single and pair terms cancel in the population, leaving the all-three-pass change. Finite samples need not cancel exactly.

On 3,000 ELI5 questions per model, GroupSuccess@3 (G@3) is the fraction of original triples with at least one usable answer. A blind Llama-3.1-8B-Instruct judge scores each answer separately for relevance, coherence and explanatory sufficiency (0–2 each); passing requires total ≥ 5 and each score ≥ 1, without external fact checking (Appendix D.4). G@3 difference intervals lie within the prespecified ± 3-percentage-point range (Table 3); single-answer differences are $+0.033$ points for both models, with intervals $[-1.608, +1.675]$.

Table 3: ELI5 group utility: at least one passing answer among the three members, under CM-EVAL-V3.1. Each method uses 3,000 original cohorts per model. CM-F8 minus clean differences are pp; simultaneous 95% reference intervals cover the four-outcome family of answer and group utility in both models.

Model	Clean G@3 (%)	CM-F8 G@3 (%)	Difference (pp)	95% ref. interval
Qwen	91.5	91.4	-0.10	$[-1.691, 1.491]$
Llama	84.0	84.3	0.30	$[-1.758, 2.358]$

Blind human ratings under the same rubric agree with 90.08% of automatic binary labels across $1{,}200$ answers from 100 shared ELI5 questions, both models and clean/CM-F8 outputs. Automatic minus human estimates of the G@3 effect (CM-F8 minus clean) are -1 point for Qwen and $+1$ for Llama, with intervals $[-10.72, +8.72]$ and $[-9.94, +11.94]$ (four-comparison family: two models, two outcomes). These do not support the same ± 3-point precision for human evaluation (Appendix D.4).

External NLL under Mistral-7B-v0.1 is normalized per evaluator answer token and averaged across members before prompt-paired comparisons. CM-F8–SynthID differences range from -0.004 to -0.001 nats/token; CM-F8–clean differences range from -0.001 to $+0.003$. All eight simultaneous intervals lie within the prespecified ± 0.05 nats/token (Appendix Table 5), complementing the task-utility assessment.

5 Discussion and Limitations

CohortMark separates preservation and detection by observation scope: Theorem 2 establishes proper-subset reference laws, while same-budget member exchanges link the observed signal to original joint generation. Kernel improvements depend on information surviving token-interval averaging. Deterministic token coordinates or particular phases can erase a step's signal, so increasing degree need not always improve detection. The exact subset guarantee assumes ideal sampling; a uniform finite-arithmetic error bound relating the implementation to this guarantee remains unavailable.

Verification requires the original complete group, correct key and nonce, alignment metadata and source-model replay. Higher degree increases scalar sampling and harmonic-evaluation work in this family (Section 4.4); retries add no model forward passes, but proposal counts do not measure end-to-end costs, which include model computation and data transfer. Text editing, missing-member recovery, active forgery and access to a detector oracle remain outside the evaluated interface.

Proper-subset preservation leaves complete-group utility free to change, requiring direct evaluation. The automatic group-utility result applies to the stated rubric; the human audit does not support the same precision (Section 4.5). The theorem covers general r, whereas experiments use three-member groups: detection spans two model families and two tasks, and group utility is evaluated on ELI5. Comparisons use each method's native detection settings and do not establish the full detection–quality–cost trade-off.

6 Conclusion

CohortMark places watermark evidence in the original complete group's relation while preserving every predesignated nonempty proper subset's independent clean sequence law at each fixed key and nonce under Theorem 2's assumptions. Token-interval analysis guides kernel design and exact-likelihood detection using correct metadata and source-model replay. Across two model families and two tasks, the three-member instance meets recall noninferiority and false-positive criteria against SynthID in all four settings, with higher recall supported in both Llama settings. Same-budget member exchanges sharply reduce positive rates, linking evidence to original joint generation. In automatic ELI5 evaluation, intervals for changes from clean generation in the fraction of groups with at least one usable answer lie within the predeclared range.

AI USE STATEMENT

In this work, we used generative AI tools to assist research ideation, including method formulation and evaluation design. Additionally, we used generative AI tools to aid manuscript drafting and language polishing, and to support retrieval and discovery of related work. We take responsibility for reviewing the AI-assisted work and for the final content of this paper, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Relational detection is statistical evidence, not a cryptographic proof of authorship. False positives and incorrectly reconstructed membership can affect attribution, requiring the operating error rates and metadata to be available. The usability rubric does not certify factual truth. The incomplete-group guarantee leaves complete-group events free to change, making direct evaluation of group utility necessary.

REPRODUCIBILITY STATEMENT

Appendix B proves the mathematical claims. Appendices D and A.5 give evaluation definitions, aggregate paired counts and reference-interval calculations. Figure and table scripts reproduce analytic examples and reported summaries. Appendix A.4 preserves earlier negative and mixed outcomes.

REFERENCES

Aubrey Alston. A steganographic design paradigm for general steganographic objectives. arXiv preprint arXiv:1707.00076, 2017. URL https://arxiv.org/pdf/1707.00076.

Dan Boneh and Victor Shoup. A graduate course in applied cryptography, 2023. URL https://crypto.stanford.edu/~dabo/cryptobook/BonehShoup_0_6.pdf. Version 0.6.

Liang Chen, Yatao Bian, Yang Deng, Deng Cai, Shuaiyi Li, Peilin Zhao, and Kam-Fai Wong. WatME: Towards Lossless Watermarking Through Lexical Redundancy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9166–9180, 2024. doi: 10.18653/v1/2024.acl-long.496. URL https://aclanthology.org/2024.acl-long.496/.

Ruibo Chen, Yihan Wu, Yanshuo Chen, Chenxi Liu, Junfeng Guo, and Heng Huang. A Watermark for Order-Agnostic Language Models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/0ce8e3434c7b486bbddff9745b2a1722-Paper-Conference.pdf.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11499265/.

Abdulrahman Diaa, Toluwani Aremu, and Nils Lukas. Optimizing Adaptive Attacks against Watermarks for Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025. URL https://proceedings.mlr.press/v267/diaa25a.html.

Fabrizio Durante. On a problem by Schweizer and Sklar. *Kybernetika*, 48(2):287–293, 2012. URL https://www.kybernetika.cz/content/2012/2/287/paper.pdf.

Xiaoyan Feng, He Zhang, Yanjun Zhang, Leo Yu Zhang, and Shirui Pan. BiMark: Unbiased Multilayer Watermarking for Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Zheng Gao, Xiaoyu Li, Xiaoyan Feng, Jiaojiao Jiang, Yang Song, Yulei Sui, Zhenchang Xing, and Liming Zhu. TRACE: A Two-Channel Robust Attribution Watermark via Complementary Embeddings for LLM-Agent Trajectories. arXiv preprint arXiv:2607.08400v1, 2026. URL https://arxiv.org/abs/2607.08400.

Thibaud Gloaguen, Nikola Jovanović, Robin Staab, and Martin Vechev. Black-box detection of language model watermarks. In *ICLR*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/c0f7ee1901fef1da4dae2b88dfd43195-Paper-Conference.pdf.

Thibaud Gloaguen, Robin Staab, Mark Vero, and Martin Vechev. Every Bit, Everywhere, All at Once: A Binomial Multibit LLM Watermark. arXiv preprint arXiv:2605.11653v1, 2026. URL https://arxiv.org/abs/2605.11653.

Batu Guan, Yao Wan, Zhangqian Bi, Zheng Wang, Hongyu Zhang, Pan Zhou, and Lichao Sun. CodeIP: A Grammar-Guided Multi-Bit Watermark for Large Language Models of Code. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9243–9258. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-emnlp.541/.

Fabrice Harel-Canada, Boran Erol, Connor Choi, Jason Liu, Gary Jiarui Song, Nanyun Peng, and Amit Sahai. Sandcastles in the Storm: Revisiting the (Im)possibility of Strong Watermarking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29698–29735, 2025. URL https://aclanthology.org/2025.acl-long.1436/.

Weiqing He, Xiang Li, Li Shen, Weijie Su, and Qi Long. Improving the Trade-off Between Watermark Strength and Speculative Sampling Efficiency for Language Models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/33a0db5c55008b14cc088fb086612b1a-Abstract-Conference.html.

Zhiwei He, Binglin Zhou, Hongkun Hao, Aiwei Liu, Xing Wang, Zhaopeng Tu, Zhuosheng Zhang, and Rui Wang. Can Watermarks Survive Translation? On the Cross-lingual Consistency of Text Watermark for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4115–4129, 2024. doi: 10.18653/v1/2024.acl-long.226. URL https://aclanthology.org/2024.acl-long.226/.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/c5b00c5bdcc6fe35907dbcca03d27652-Paper-Conference.pdf.

Yu-Shin Huang, Chao Tian, and Krishna Narayanan. Optimal multi-bit generative watermarking schemes under worst-case false-alarm constraints. arXiv preprint arXiv:2604.08759v1, 2026. URL https://arxiv.org/html/2604.08759v1.

Nikola Jovanović, Robin Staab, and Martin Vechev. Watermark Stealing in Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 22570–22593, 2024. URL https://proceedings.mlr.press/v235/jovanovic24a.html.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17061–17084. PMLR, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the Reliability of Watermarks for Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/d78e9e4316e1714fbb0f20be66f8044c-Paper-Conference.pdf.

Gregory Kang Ruey Lau, Xinyuan Niu, Hieu Dao, Jiangwei Chen, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. Waterfall: Scalable Framework for Robust Text Watermarking and Provenance for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20432–20466, 2024. doi: 10.18653/v1/2024.emnlp-main.1138. URL https://aclanthology.org/2024.emnlp-main.1138/.

Nghia T. Le, Alan Ritter, and Kartik Goyal. Semantic Differentiation for Tackling Challenges in Watermarking Low-Entropy Constrained Generation Outputs. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2601.11629. arXiv:2601.11629v2.

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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4890–4911, 2024. URL https://aclanthology.org/2024.acl-long.268/.

Baixin Li and Haiyun He. SAC-Copula: Quality-preserving watermarking for diffusion language models via smooth correlated gumbel fields. arXiv preprint arXiv:2608.20839v1, 2026. URL https://arxiv.org/html/2608.20839v1.

Michael Y. Li, Anthony Zhan, Kanishk Gandhi, Noah D. Goodman, and Emily B. Fox. Quasi-MoTTo: Quasi-monte carlo test-time scaling. arXiv preprint arXiv:2607.01179v1, 2026. URL https://arxiv.org/html/2607.01179v1.

Xingchi Li, Guanxun Li, and Xianyang Zhang. Segmenting Watermarked Texts From Language Models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1a8d295871250443f9747d239925b89d-Paper-Conference.pdf.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A Semantic Invariant Robust Watermark for Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/1a2131ebe25bd55e4fc734126ea583ed-Paper-Conference.pdf.

Shuliang Liu, Xingyu Li, Hongyi Liu, Dong Fang, Yibo Yan, Bingchen Duan, Qi Zheng, Lingfeng Su, and Xuming Hu. Distilling the Thought, Watermarking the Answer: A Principle Semantic Guided Watermark for Large Reasoning Models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=T6NVogsXCZ.

Yepeng Liu and Yuheng Bu. Adaptive Text Watermark for Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024. URL https://proceedings.mlr.press/v235/liu24e.html.

Wenlong Meng, Chen Gong, Terry Yue Zhuo, Fan Zhang, Kecen Li, Zheng Liu, Zhou Yang, Chengkun Wei, and Wenzhi Chen. Watermarking LLM Agent Trajectories. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2602.18700.

Leyi Pan, Aiwei Liu, Shiyu Huang, Yijian Lu, Xuming Hu, Lijie Wen, Irwin King, and Philip S. Yu. Can LLM Watermarks Robustly Prevent Unauthorized Knowledge Distillation? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13228–13251, 2025. doi: 10.18653/v1/2025.acl-long.648. URL https://aclanthology.org/2025.acl-long.648/.

Qi Pang, Shengyuan Hu, Wenting Zheng, and Virginia Smith. No Free Lunch in LLM Watermarking: Trade-offs in Watermarking Design Choices. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/file/fa86a9c7b9f341716ccb679d1aeb9afa-Paper-Conference.pdf.

Shinwoo Park, Hyejin Park, Hyeseon Ahn, and Yo-Sub Han. WaterMod: Modular Token-Rank Partitioning for Probability-Balanced LLM Watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32683–32691, 2026. doi: 10.1609/aaai.v40i39.40546. URL https://ojs.aaai.org/index.php/AAAI/article/view/40546.

Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking Makes Language Models Radioactive. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0664. URL https://papers.nips.cc/paper_files/paper/2024/file/2567c95fd41459a98a73ba893775d22a-Paper-Conference.pdf.

Dor Tsur, Carol Xuan Long, Claudio Mayrink Verdun, Hsiang Hsu, Richard Chen, Haim Permuter, Sajani Vithana, and Flavio P. Calmon. HeavyWater and SimplexWater: Distortion-free LLM watermarks for low-entropy distributions. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/file/9a22bcd525daefeb4a29db1ff1c03ced-Paper-Conference.pdf.

Shangqing Tu, Yuliang Sun, Yushi Bai, Jifan Yu, Lei Hou, and Juanzi Li. WaterBench: Towards Holistic Evaluation of Watermarks for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1517–1542, 2024. doi: 10.18653/v1/2024.acl-long.83. URL https://aclanthology.org/2024.acl-long.83/.

Chenrui Wang, Junyi Shu, Billy Chiu, Yu Li, Saleh Alharbi, Min Zhang, and Jing Li. Learning to Watermark: A Selective Watermarking Framework for Large Language Models via Multi-Objective Optimization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/bc222e8153a49c1b30a1b8ba96b35117-Paper-Conference.pdf.

Lean Wang, Wenkai Yang, Deli Chen, Hao Zhou, Yankai Lin, Fandong Meng, Jie Zhou, and Xu Sun. Towards Codable Watermarking for Injecting Multi-Bits Information to LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/abdc8c031aa6c6917c3b593166e5e340-Paper-Conference.pdf.

Yihan Wu, Xuehao Cui, Ruibo Chen, and Heng Huang. Analyzing and evaluating unbiased language model watermark. In *ICLR*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/e6643cfb7da77f9a4349df318e97aa3f-Paper-Conference.pdf.

Xiaojun Xu, Jinghan Jia, Yuanshun Yao, Yang Liu, and Hang Li. Robust Multi-bit Text Watermark with LLM-based Paraphrasers. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025. URL https://proceedings.mlr.press/v267/xu25k.html.

Ruiyi Yan and Yugo Murawaki. Addressing Tokenization Inconsistency in Steganography and Watermarking Based on Large Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7076–7098, 2025. URL https://aclanthology.org/2025.emnlp-main.361/.

Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the Sand: Impossibility of Strong Watermarking for Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 58851–58880, 2024. URL https://proceedings.mlr.press/v235/zhang24o.html.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable Robust Watermarking for AI-Generated Text. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/beae9ed5316bcc48e616754c06c11875-Paper-Conference.pdf.

Junlin Zhu, Baizhou Huang, and Xiaojun Wan. QuantileMark: A message-symmetric multi-bit watermark for LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6790–6806. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.acl-long.308/.

A SUPPLEMENTARY FIGURES AND RESULTS

The following panels show how tokenization shapes evidence, how membership and metadata affect detection, and how utility varies across groups. Exact derivations appear in Appendix C; evaluation rules are in Appendix D. Tables 4–5 retain precision beyond the main results.

A.1 TOKENIZATION AND OBSERVABLE EVIDENCE

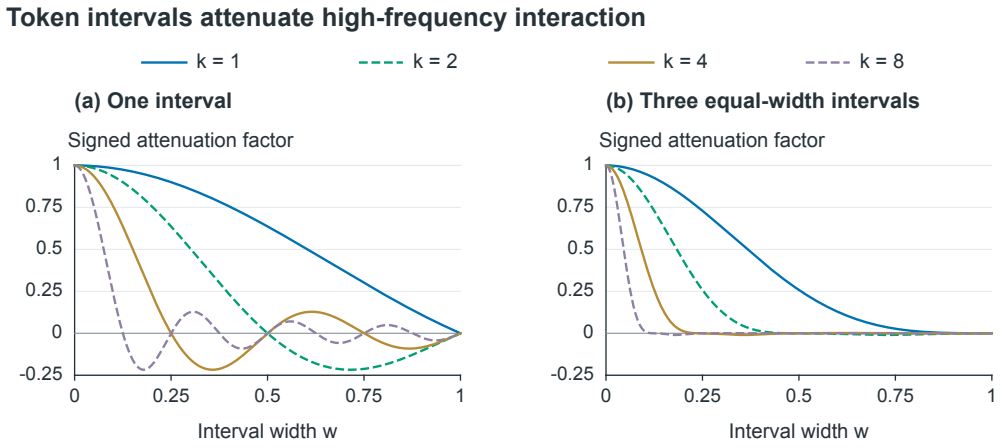


Figure 5: Token intervals attenuate higher harmonics. The signed factor $\mathrm{sinc}(\pi k w)$ for one interval (left) becomes $[\mathrm{sinc}(\pi k w)]^3$ for three equal-width intervals (right), for $k = 1, 2, 4, 8$. Values follow Equation (13), with the continuous value one at $w = 0$; negative lobes retain their sign.

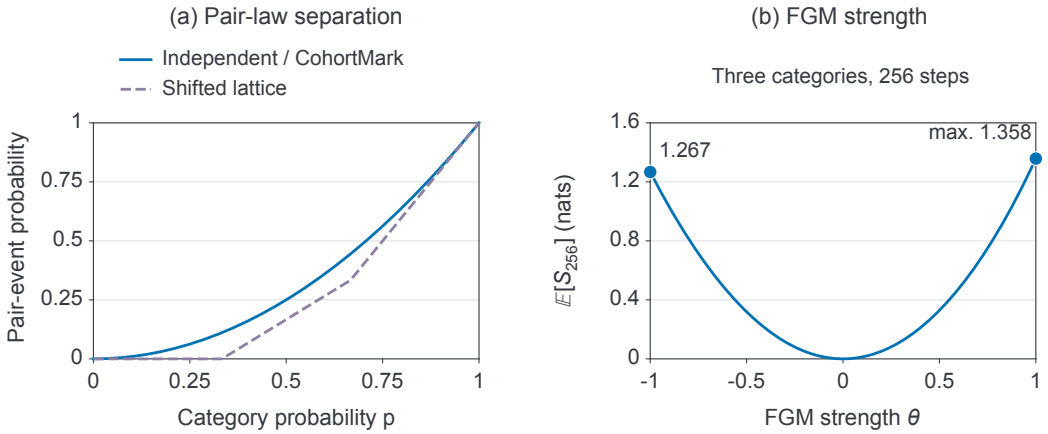


Figure 6: Classical alternatives expose two distinct design choices. (a) Both members 1 and 3 select a category occupying CDF interval $[0, p)$. Independent clean sampling and CohortMark give probability p^2; a continuously shifted three-point lattice gives $(p-1/3)_+ + (p-2/3)_+$. (b) Exact 256-step expected evidence for the three-way FGM subfamily, from all 27 tuples with $p = (.7, .2, .1)$. Its maximum is 1.357761503521 nats over $\theta \in [-1, 1]$, without an imposed density floor.

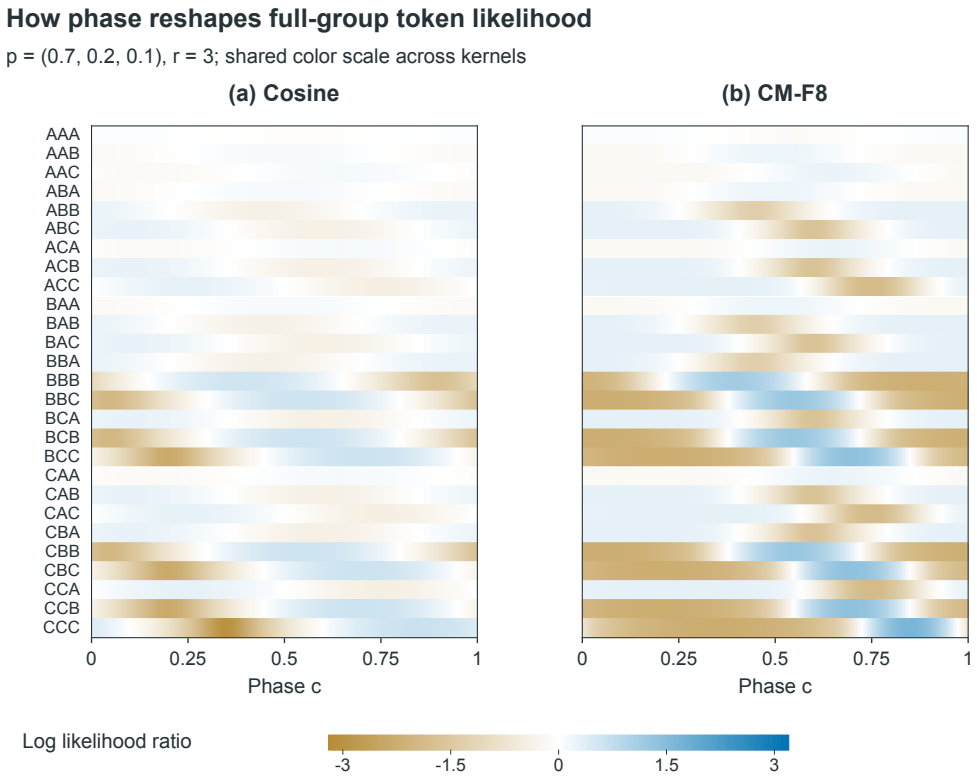


Figure 7: The same token marginals support different evidence across phase. Rows are all 27 ordered triples for $p = (.7, .2, .1)$; columns are $c_t = (t + .5)/256$. The shared scale shows $\log \Lambda_c(v)$ for the cosine and CM-F8. Blue indicates positive evidence and ochre negative evidence. Weighting each cell by $Q_c(v)$ gives its KL contribution.

A.2 Membership and metadata controls

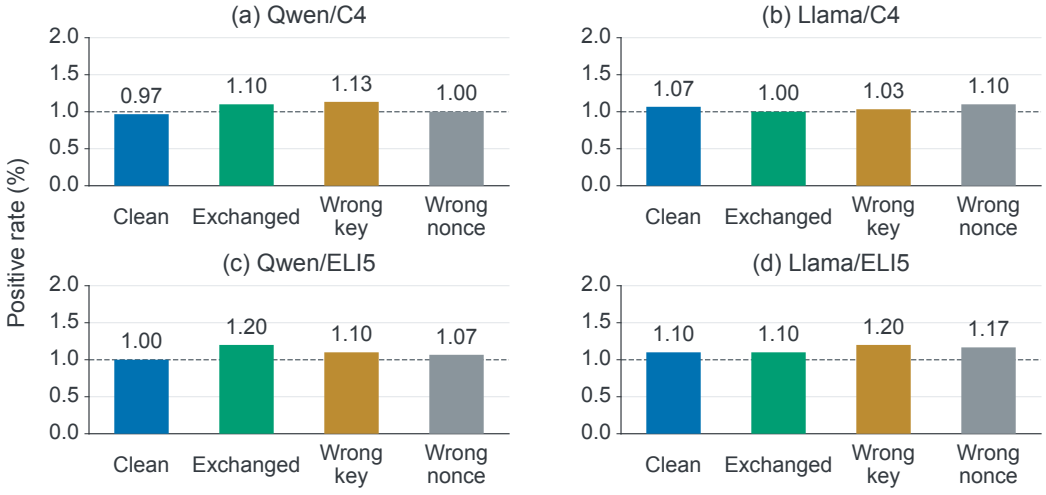


Figure 8: CM-F8 positives remain near the clean reference after changing membership or metadata. Each bar is the descriptive positive rate among 3,000 groups. Exchanged groups swap member 3 within 1,500 two-question blocks; wrong-key and wrong-nonce scores use the original outputs. The dashed 1% line is the nominal clean calibration level. Original-group recall is 94.8–99.0% (Table 2); block-based exchange precision is retained in Table 4.

A.3 GROUP UTILITY AND EVALUATOR VARIATION

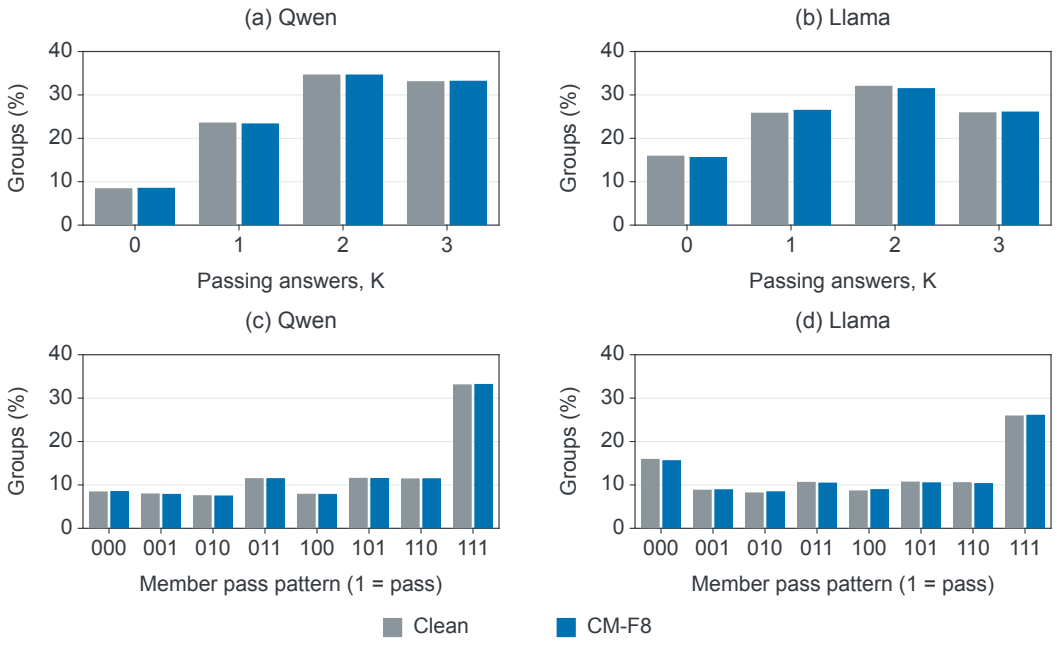


Figure 9: Clean and CM-F8 have similar utility distributions on 3,000 ELI5 groups per model. Top: passing-answer count K, with G@3 $= 1 - \Pr(K = 0)$. Bottom: the complete three-member label patterns, retaining which members pass. The top panels sum the bottom frequencies by the number of ones; both use the observed counts.

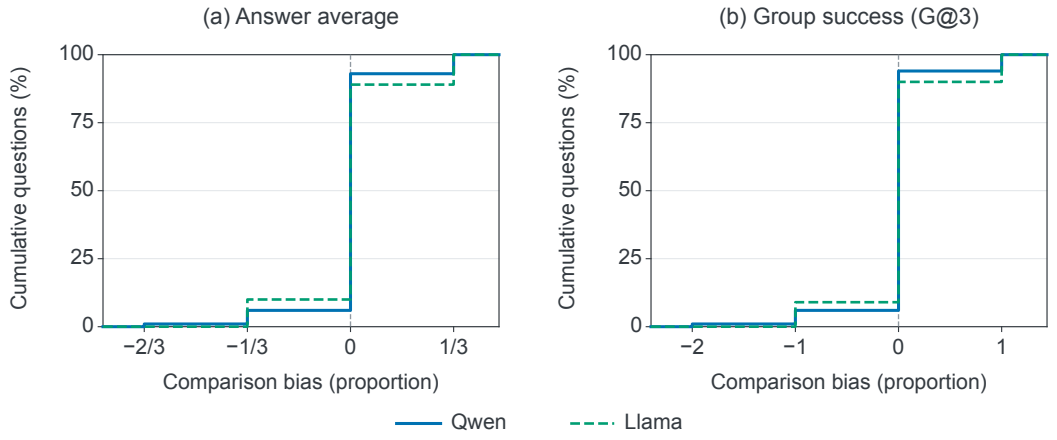


Figure 10: Question-level human-audit bias is concentrated near zero, with wider variation for group success. Each value is the automatic F8–clean effect minus the human effect on one of 100 questions per model. Answer utility averages three labels; group utility takes their OR. The ECDF retains every support point, including the rare -2 group value. Table 5 reports the corresponding mean-bias intervals.

A.4 HISTORICAL COSINE STUDIES

The historical Qwen3-8B panels use three texts and 768 tokens per group.

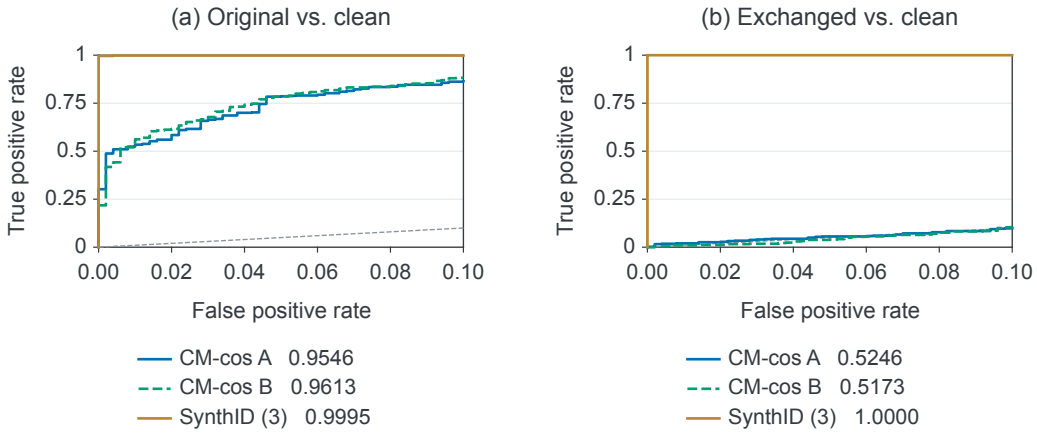


Figure 11: Historical native-score ROC for original (left) and exchanged (right) groups against the same 500 paired-clean groups, with 500 groups per condition. The view shows FPR $\leq .10$; legend AUROCs use the full range. Exchanged-condition ROCs rank native scores against clean; SynthID groups still contain watermarked texts.

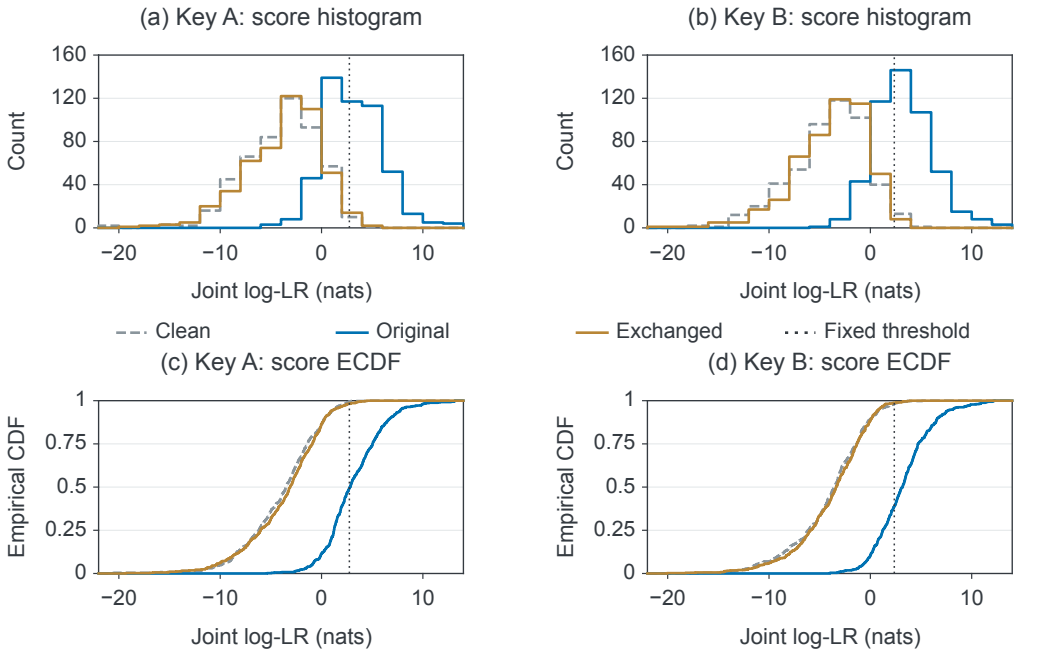


Figure 12: Historical cosine score distributions by key. Histograms share fixed bins; ECDFs retain every observed score. Each clean, original and exchanged condition contains 500 groups. Vertical lines mark archived thresholds ($S \geq \tau$), calibrated on 2,000 separate clean groups. Exchanges preserve 250 two-question blocks. The extra 500 null-audit groups used for archived FPR estimates are not included here.

A.4.1 MEMBER-LEVEL QUALITY AND EARLIER FINDINGS

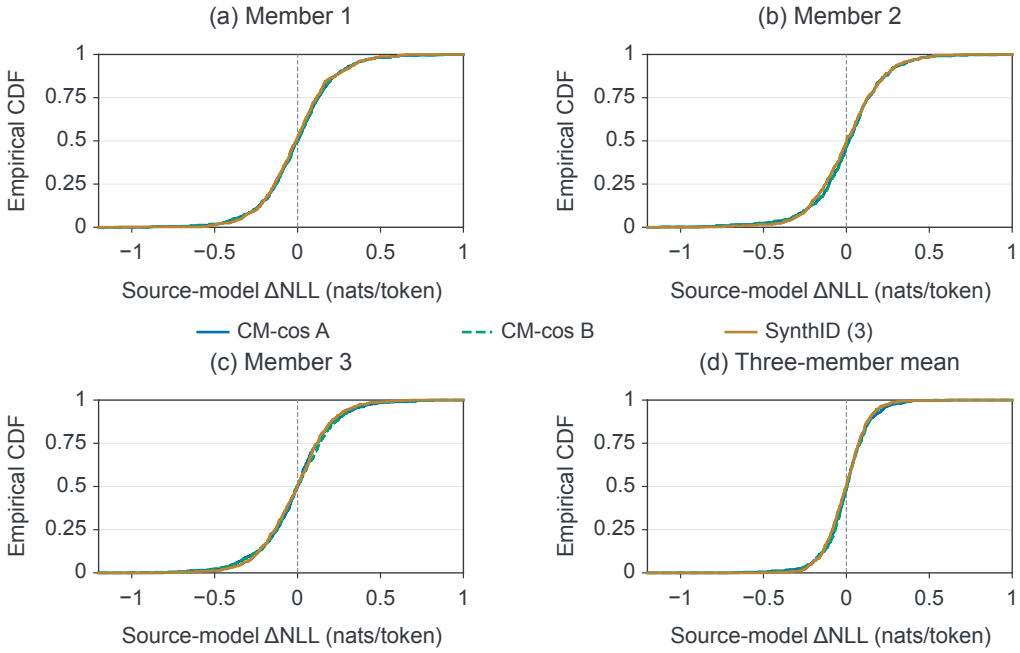


Figure 13: Historical marked-minus-clean source-model NLL for each member and the three-member question average (500 questions per method; nats/token). CM-A/B curves coincide for members 1 and 2 because those arrays share the stored base-stream mapping. Pooled A/B mean differences are $-.000527$ $[-.012278, .011037]$ and $.000103$ $[-.010950, .011199]$, using the archived paired-block 95% intervals. These source-model differences are distinct from the current external-NLL evaluation.

Earlier-stage findings. An earlier small-model pilot (P0) had positive score means but failed its key-stability gate: a 10.2 pp recall difference exceeded the 10 pp tolerance. A two-key, two-replicate follow-up (K2S2) localized this variation without meeting every point-power condition; replicates changed both sampling coins and nonces. In the separate Qwen3-8B study, frozen-threshold cosine A/B recall was $50.4/61.2\%$ on 500 marked groups, and FPR was $1.0/1.9\%$ on 1,000 paired-clean plus null-audit groups. The B–A recall gap was 10.8 $[5.8, 15.8]$ pp. Positive mean prefix gains coexisted with tenth percentiles $-.447195/-.589688$ nats. Across six wrong-key, wrong-nonce and wrong-group checks, all one-sided 95% score upper bounds were negative (largest -3.405439), and all corresponding matched-minus-mismatch lower bounds were positive (smallest 6.448770).

Five historical signed subset checks passed their older upper-bound rule; $\{2\}$ remained inconclusive, and the direction-symmetric reanalysis (S1) rejected none after correction. The shifted-lattice baseline (M2) did not establish diagnostic separation: member-3 NLL changed by $-.019609$ $[-.038738, -.000311]$, while the pooled change was $-.010650$ $[-.022241, .001281]$. The first-token event (S2) is joint selection by members 1 and 3 of the highest-probability token, with ties resolved by the smallest token ID. All five paired event-difference intervals (pointwise 95%) crossed zero: $[-2.2, 7.0]$, $[-5.8, 3.8]$, $[-6.0, 2.4]$, $[-8.6, .2]$ and $[-5.0, 3.6]$ pp for CM-A–clean, CM-B–clean, M2–clean, M2–CM-A and M2–CM-B (500 prompts). Complete historical precision is retained in the companion tables.

Historical clean generation, cosine generation and replay reached $88.2858/74.2668/81.8060$ tokens/s within the archived timing boundaries, excluding model loading. These timings apply to the historical cosine configuration.

A.5 PRECISION BEYOND THE MAIN TABLES

Table 4: Diagnostic and relational precision. Upper: proper-subset AUROC and simultaneous 95% reference intervals over one 24-entry family (3,000 marked and clean groups; SE .008). Lower: paired recall gains over the cosine on 3,000 questions and original-minus-exchanged effects on 1,500 two-question blocks. Each lower-column comparison uses a separate four-configuration simultaneous 95% reference family.

Subset	Qwen/C4	Llama/C4	Qwen/ELI5	Llama/ELI5
	0.499	0.504	0.496	0.507
$\{1\}$	$[0.474, 0.524]$	$[0.479, 0.529]$	$[0.471, 0.521]$	$[0.482, 0.532]$
	0.503	0.497	0.501	0.495
$\{2\}$	$[0.478, 0.528]$	$[0.472, 0.522]$	$[0.476, 0.526]$	$[0.470, 0.520]$
	0.496	0.500	0.505	0.501
$\{3\}$	$[0.471, 0.521]$	$[0.475, 0.525]$	$[0.480, 0.530]$	$[0.476, 0.526]$
	0.505	0.506	0.498	0.503
$\{1, 2\}$	$[0.480, 0.530]$	$[0.481, 0.531]$	$[0.473, 0.523]$	$[0.478, 0.528]$
	0.501	0.494	0.504	0.498
$\{1, 3\}$	$[0.476, 0.526]$	$[0.469, 0.519]$	$[0.479, 0.529]$	$[0.473, 0.523]$
	0.498	0.502	0.499	0.502
$\{2, 3\}$	$[0.473, 0.523]$	$[0.477, 0.527]$	$[0.474, 0.524]$	$[0.477, 0.527]$

Configuration	F8–cos recall (pp)	Original–exchanged (pp)
Qwen/C4	46.00 $[43.70, 48.30]$	97.90 $[97.14, 98.66]$
Llama/C4	43.40 $[41.09, 45.71]$	96.60 $[95.62, 97.58]$
Qwen/ELI5	50.60 $[48.26, 52.94]$	96.20 $[95.18, 97.22]$
Llama/ELI5	51.40 $[49.04, 53.76]$	93.70 $[92.38, 95.02]$

Table 5: Quality and human-audit precision. Upper: external NLL differences (nats/answer token), averaging three members within each of 3,000 paired questions; SD .300 and one eight-comparison simultaneous 95% reference family. Lower: automatic and human F8–clean effects and their difference (bias), in percentage points, on 100 paired questions per model. SDs use proportions; the four bias intervals form a separate simultaneous 95% family.

Configuration	Contrast	Δ NLL	95% ref. interval
Qwen/C4	F8 minus clean	0.002	$[-0.013, 0.017]$
Qwen/C4	F8 minus SynthID	-0.002	$[-0.017, 0.013]$
Llama/C4	F8 minus clean	-0.001	$[-0.016, 0.014]$
Llama/C4	F8 minus SynthID	-0.004	$[-0.019, 0.011]$
Qwen/ELI5	F8 minus clean	0.003	$[-0.012, 0.018]$
Qwen/ELI5	F8 minus SynthID	-0.001	$[-0.016, 0.014]$
Llama/ELI5	F8 minus clean	0.000	$[-0.015, 0.015]$
Llama/ELI5	F8 minus SynthID	-0.002	$[-0.017, 0.013]$

Model	Endpoint	Auto	Human	Bias	SD	95% ref. interval
Qwen	Answer	0.33	0.33	0.00	0.134	$[-3.35, 3.35]$
Qwen	G@3	0.00	1.00	-1.00	0.389	$[-10.72, 8.72]$
Llama	Answer	0.33	0.00	0.33	0.153	$[-3.50, 4.17]$
Llama	G@3	1.00	0.00	1.00	0.438	$[-9.94, 11.94]$

B DISTRIBUTIONAL GUARANTEES AND PROOFS

B.1 PERIODIC DENSITY, PROJECTIONS AND SAMPLER

For an integer $J \geq 1$, expand the square modulus in Equation (9):

$$\frac{1}{J+1}\left|\sum_{a=0}^{J} e^{2\pi \mathrm{i} a s}\right|^2 = \frac{1}{J+1}\sum_{a,b=0}^{J} e^{2\pi \mathrm{i}(a-b)s} = 1 + 2\sum_{k=1}^{J}\left(1 - \frac{k}{J+1}\right)\cos(2\pi k s).$$

The constant term is one, and every other term integrates to zero. Nonnegativity follows from the square modulus. The triangle inequality gives an upper bound $J+1$ before floor regularization, hence $\epsilon \leq f_{J,\epsilon} \leq \epsilon + (1-\epsilon)(J+1)$. At integer s, the equivalent sine quotient is defined by its continuous limit, not evaluated as $0/0$.

Proof of Proposition 1. For any fixed other coordinates and phase, translating u_j over $[0,1)$ traverses one full period of f. Its integral is one. Iterating this operation shows that q_c integrates to one and that every proper projection has density one. Nonnegativity permits the successive integrations by Tonelli's theorem. The inverse-CDF preimage of a retained token tuple is a rectangle of volume $\prod_{i\in A} p_i(v_i)$, giving the token product law. □

For Equation (4), condition on the first $r-1$ independent uniforms. Circular translation preserves length, so U_r has density $f((u_r - c + \sum_{i<r} u_i) \bmod 1)$. Multiplication by the density-one first coordinates gives q_c. For a bound $f \leq M$, with $M = M_{J,\epsilon}$ in the Fejér family, uniform rejection satisfies

$$\Pr(Z \in B, \text{accept}) = \frac{1}{M}\int_B f(z)\,dz.$$

Acceptance probability is $1/M$; conditioning on acceptance gives f. Independent retries terminate almost surely with mean M proposals. The sampler uses no finite retry cap or fallback that changes its law. The coupling adds no model forward pass, while actual scalar and synchronization costs require timing.

B.2 THE FIXED-KEY SEQUENCE IDENTITY

Proof of Theorem 2. Fix $\mathcal{M}$ and define the full and retained history sigma-algebras

$$\mathcal{F}_{t-1} = \sigma(\mathcal{M}, X_{i,s} : i \in [r], s < t), \qquad \mathcal{G}^A_{t-1} = \sigma(\mathcal{M}, X_{i,s} : i \in A, s < t).$$

Fresh conditional sampling makes Proposition 1 apply at every reachable full history. Its projected transition is measurable from retained histories because the source transitions are member-local. Consequently,

$$\begin{aligned} Q(X_{A,t} = v_A \mid \mathcal{G}^A_{t-1}) &= \mathbb{E}_Q[Q(X_{A,t} = v_A \mid \mathcal{F}_{t-1}) \mid \mathcal{G}^A_{t-1}] \\ &= \prod_{i\in A} p_{i,t}(v_i \mid h_i, X_{i,<t}). \end{aligned}$$

The chain rule yields

$$Q(X_A = x_A \mid \mathcal{M}) = \prod_{t=1}^{T}\prod_{i\in A} p_{i,t}(x_{i,t} \mid h_i, x_{i,<t}) = \prod_{i\in A} P_i(x_i \mid h_i).$$

Zero-probability source transitions have zero marked mass as well. Finite vocabulary and fixed horizon make these path probabilities determine the distribution. Key and nonce remain fixed throughout. □

Predictable phases suffice for the projection identity because the projected transition is phase-independent. Detection additionally needs reconstructible phases. A disclosed phase table encoding missing histories would change the observer's information, which is why the evaluated rule is metadata-only. Cross-member repetition penalties, shared decoding state and selecting a subset after inspecting all outputs are outside this sequence experiment.

B.3 EXACT RECTANGLE LIKELIHOOD AND INFORMATION

For integer harmonic k and a positive-width interval $[a, b)$ with center μ,

$$\int_a^b e^{2\pi i k u}\, du = (b-a)e^{2\pi i k \mu}\,\mathrm{sinc}(\pi k(b-a)). \tag{12}$$

Taking real parts and separating the rectangle integrals gives

$$Q_c(v) = \left(\prod_i w_i\right)\left[1 + \sum_{k=1}^{J} b_k \,\mathrm{Re}\left(e^{-2\pi i k c}\prod_i e^{2\pi i k \mu_i}\,\mathrm{sinc}(\pi k w_i)\right)\right]. \tag{13}$$

Dividing by the positive clean rectangle mass gives Equation (6). If any width is zero, both masses are zero. For positive widths the ratio is the average of the continuous density over the rectangle, so it lies in $[\epsilon, M]$, with $M = M_{J,\epsilon}$ in the Fejér family. For $k \geq 2$ the sinc may be negative; replacing it by its magnitude changes the likelihood.

Proof of Theorem 3. Apply the rectangle calculation at each transition on the same observed path. The ratios multiply to $Q(X)/P_0(X) = \prod_t \Lambda_t$. Conditional normalization gives

$$\mathbb{E}_0[\Lambda_t \mid \mathcal{F}_{t-1}] = \sum_{v \in \mathcal{V}^r}\left(\prod_i p_{i,t}(v_i)\right)\Lambda_t(v) = 1.$$

Thus L_t is a nonnegative mean-one null martingale. Positive ϵ gives common support and

$$T\log\epsilon \leq S_T \leq T\log M,$$

so the log expectations are finite. Their definitions give the forward and reverse KL identities. Markov's inequality gives the fixed-length null tail bound $P_0(S_T \geq s) \leq e^{-s}$ for $s \geq 0$. □

The chain rule further gives

$$\mathbb{E}_Q S_T = \sum_{t=1}^{T} \mathbb{E}_Q D_{\mathrm{KL}}(Q_t(\cdot \mid X_{<t}) \| P_{0,t}(\cdot \mid X_{<t})).$$

Expected information is nondecreasing with horizon; individual increments can be negative. The cosine baseline at $\epsilon = 0$ also has positive token likelihood on positive-width rectangles: $\prod_i |\mathrm{sinc}(\pi w_i)| < 1$. Its common finite support suffices for the same likelihood identities, without using the new uniform .05 lower bound.

For a detection event with marked probability β and clean probability α, data processing gives the necessary conditions

$$\mathrm{kl}(\beta\|\alpha) \leq \mathbb{E}_Q S_T, \qquad \mathrm{kl}(\alpha\|\beta) \leq -\mathbb{E}_0 S_T.$$

Here kl is binary relative entropy. These inequalities constrain population operating points and information; they do not reconstruct score distributions from sample aggregates. The likelihood-ratio population ROC with randomized tie handling is concave, so a point (α, β) implies area at least $(1 + \beta - \alpha)/2$.

B.4 COMPLETE-GROUP UTILITY

For member-local binary rules $Y_i = g_i(h_i, X_i)$, inclusion–exclusion gives

$$\mathbb{I}\{\max_i Y_i = 1\} = Y_1 + Y_2 + Y_3 - Y_1Y_2 - Y_1Y_3 - Y_2Y_3 + Y_1Y_2Y_3.$$

Taking clean and marked expectations cancels the single and pair terms by Theorem 2, proving Equation (11). Averaging over a common prompt distribution preserves the identity. Independent evaluation randomness can also be conditioned on and averaged out. A rule that compares group members with each other has a different observation interface.

Finite-sample inclusion–exclusion retains the single and pair sampling differences. In counts, Qwen has changes $+3$ in the sum of single passes, $+9$ in pair passes and $+3$ in all-three passes, yielding $-3/3000$ group difference. Llama has $+3, -1, +5$, yielding $9/3000$. The companion eight-pattern counts reproduce these identities exactly; approximate equality of empirical low-order counts is not used as a proof of the population theorem.

B.5 KEYS, REGROUPING AND APPROXIMATION

The proper-subset identity conditions on a fixed key and requires no averaging over that key. Hidden-phase and incorrect-metadata experiments have separate assumptions.

If a fresh canonical input receives an ideal random-function phase C whose first J circular moments vanish, then

$$\mathbb{E}_C q_C(u) = 1 + \sum_{k=1}^{J} b_k \operatorname{Re}\left(e^{2\pi \mathrm{i} k \sum_i u_i} \mathbb{E} e^{-2\pi \mathrm{i} k C}\right) = 1.$$

When the current partitions are independent of this fresh phase, token integration gives a clean public transition. Induction covers prompts selected before fresh phase values. The observer sees public metadata and outputs, not phases, sampling coins or a detector oracle. A complete uniform grid with more than J points has the required circular moments; the ideal 53-bit grid meets this condition. A two-point grid, sufficient for one harmonic, is insufficient for degree 8. Reusing a canonical input breaks the freshness argument, even though distinct inputs can legitimately map to the same phase value.

A conditional computational formulation replaces domain-separated keyed interfaces by independent random functions through PRF hybrids (Boneh & Shoup, 2023). This requires a secure PRF, fresh inputs and exact efficient simulation or the same ideal sampling oracle in both worlds. The bound includes component distinguishing advantages and collision/simulation errors. Public experimental seeds specify reproducible instances and do not constitute a hidden-coins security argument.

For a fixed legal trajectory x and independent mismatch phases D_t with vanishing first J circular moments,

$$\mathbb{E}_D e^{S_T^D(x)} = 1, \qquad \mathbb{E}_D S_T^D(x) \le 0, \qquad \Pr_D(S_T^D(x) \ge s) \le e^{-s}, \quad s \ge 0.$$

These statements average over independent mismatch tables. Fixed wrong-key and wrong-nonce outputs are reported as their observed counts, not universal error guarantees.

For independently generated original cohorts p, q, the subset theorem gives

$$(X_{p,1}, X_{p,2}, X_{q,3}) \sim P_{p,1} \otimes P_{p,2} \otimes P_{q,3}.$$

Each member retains its own prompt and replay intervals, giving the appropriate product reference. The reverse exchange shares the original cohorts, so inference uses a two-prompt block. A regrouped SynthID or KGW collection still contains watermarked texts; its native positivity is not a clean false-positive rate. Group identifiers can also be used by other protocols, so this comparison concerns the joint distributional interface rather than exclusivity of group binding.

If implemented conditional kernels differ from their ideal laws by at most δ_t in total variation at each reachable step, coupling until the first disagreement gives

$$d_{\mathrm{TV}}(P_{\mathrm{implemented}}, P_{\mathrm{ideal}}) \le \min(1, \sum_t \delta_t).$$

Projection contracts TV. Comparing approximate clean and marked laws adds their errors. No uniform machine-level δ_t is supplied. For the ideal F8 law the score is bounded by $T \log(.05)$ and $T \log(8.6)$; errors in reconstructing probabilities and score values require their own control.

C Derivations for the supplementary mechanism figures

C.1 Harmonic attenuation and phase cancellation

For fixed categorical transitions, write

$$B_k(v) = \prod_i e^{2\pi \mathrm{i} k \mu_i(v_i)} \operatorname{sinc}(\pi k p_i(v_i)), \quad h_c(v) = \sum_{k=1}^{J} b_k \operatorname{Re}(e^{-2\pi \mathrm{i} k c} B_k(v)).$$

Figure 5 shows how the interval factors control each harmonic before phases are combined. For three equal-width token intervals, their product attenuates the interaction more strongly than a single interval.

Then $\Lambda_c(v) = 1 + h_c(v)$ and $\chi^2(Q_c \| P_0) = \sum_v P_0(v) h_c(v)^2$. Integrating c continuously over $[0, 1)$ eliminates cross-harmonic terms and gives

$$\mathbb{E}_c \chi^2(Q_c \| P_0) = \frac{1}{2} \sum_{k=1}^{J} b_k^2 \sum_v P_0(v) |B_k(v)|^2 = \frac{1}{2} \sum_{k=1}^{J} b_k^2 R_k.$$

For $x \in [0, \pi]$, $|\sin(kx)| \le k \sin x$, so $0 \le R_k \le R_1$. If $R_1 > 0$, the degree-8 phase-average χ^2 is between 2.852346 and 9.091852 times that of the cosine at the same fixed transitions. This is a phase-average divergence statement, not a fixed-key KL or recall multiplier. A uniform finite phase grid needs more than $2J$ points to eliminate all squared-expression cross terms; continuous phase averaging avoids this extra notation in the main result.

To connect with a small-interaction expansion, use the analysis family $f_\eta = 1 + \eta(f - 1)$, $0 \le \eta \le 1$. Normalization gives $\mathbb{E}_0 h_c = 0$ and

$$D_{\mathrm{KL}}(Q_{c,\eta} \| P_0) = \frac{\eta^2}{2} \mathbb{E}_0 h_c^2 - \frac{\eta^3}{6} \mathbb{E}_0 h_c^3 + O(\eta^4).$$

After phase averaging, the leading term is $\eta^2 \sum_k b_k^2 R_k / 4$ with an $O(\eta^3)$ remainder in general. Multiple harmonics can produce nonzero third moments; the cosine's stronger cancellation does not apply automatically. The main method uses $\eta = 1$ and its exact likelihood.

In the three-member balanced-binary example, the centers sum to $3/4 + m/2$, with m upper-half tokens. At $c = 0$, every odd-harmonic cosine is zero and every even-harmonic sinc is zero, giving $\Lambda = 1$. At $c = 1/4$, the two parity classes have ratios $1 \pm a_J$, where

$$a_J = \left(\frac{2}{\pi}\right)^3 \sum_{\substack{1 \le k \le J \\ k \text{ odd}}} \frac{b_k (-1)^{(k-1)/2}}{k^3}, \quad D(a) = \tfrac{1}{2}[(1 + a)\log(1 + a) + (1 - a)\log(1 - a)].$$

The cosine has $a = .258012$, $D = .033665$; degree 4 at floor .05 has $.384916/.076027$; degree 8 has $.425075/.093283$. A deterministic member likewise gives $\operatorname{sinc}(\pi k) = 0$ for every nonzero integer harmonic.

These identities explain the zeros at $c = 0$ and $c = 1/2$ in Figure 3a. Figure 3b gives the corresponding phase response for the asymmetric three-category partition, evaluated through the same exact rectangle likelihood.

For the fixed three-category transitions in Figure 3b, each member uses $p = (.7, .2, .1)$, in that CDF order, and $c_t = (t + .5)/256$. At each of 256 steps we enumerate the 27 token tuples and sum $\prod_i p(v_i) \Lambda_t(v) \log \Lambda_t(v)$. Summing over steps gives 3.3425155031, 2.9929544871, 8.6568750641 and 11.3841054313 nats for the cosine, floor-matched degree 1, degree 4 and degree 8 (Figure 14). No generated model trajectories enter this calculation.

The token-level structure underlying this expectation appears in Figure 7. Each row shows how evidence for one ordered token tuple changes with phase. Its contribution to KL is obtained by weighting $\log \Lambda_c(v)$ by $Q_c(v)$ and summing over tuples.

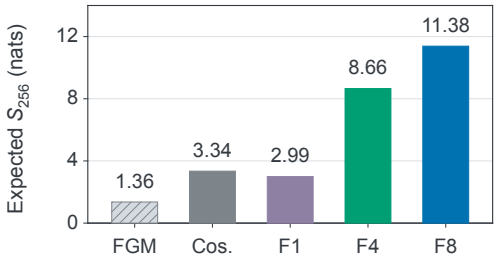


Figure 14: Exact expected S_{256} for three members with $p = (0.7, 0.2, 0.1)$. Periodic kernels use $c_t = (t+0.5)/256, t = 0, \ldots, 255$. Hatching denotes the FGM subfamily's maximum over admissible per-step strengths. F1/F4/F8 have floor 0.05; cosine has floor zero. Values sum over 27 token tuples per step, then over 256 steps.

C.2 Classical constructions and the kernel comparison

Low-order independent copulas provide the distributional foundation for the construction. For $r \geq 3$, fix the first $r-2$ coordinates to z and define

$$D_z(a,b) = \int_0^a \int_0^b f\left(\left[x + y + \sum_j z_j - c\right] \bmod 1\right) dy\, dx.$$

Periodicity makes both one-dimensional margins of D_z uniform. Integrating over z recovers the periodic copula, placing it in the representation of Durante (2012, Theorem 3.1 and Corollary 3.5). This identifies the inherited probability structure; the specified periodic density determines our sampler and harmonic token likelihood.

The three-way Farlie–Gumbel–Morgenstern (FGM) example in Durante (2012, Example 3.4) gives a concrete comparison with the same proper-subset property. Its CDF and density are

$$C_\theta(u) = \prod_{i=1}^{3} u_i \left[1 + \theta \prod_{i=1}^{3}(1-u_i)\right], \qquad q_\theta(u) = 1 + \theta \prod_{i=1}^{3}(1-2u_i), \quad |\theta| \leq 1. \tag{14}$$

Each interaction factor integrates to zero. Averaging over token intervals with centers μ_i therefore gives the exact ratio

$$\Lambda_\theta(v) = 1 + \theta \prod_{i=1}^{3}(1 - 2\mu_i(v_i)). \tag{15}$$

With fresh draws, member-local transitions and a metadata-determined θ_t, the same conditional-expectation argument proves a fixed-key proper-subset sequence law. Multiplying the token ratios gives the sequence likelihood. These are direct consequences of the existing construction under the stated categorical specialization.

For the main text's fixed distribution $p = (.7, .2, .1)$, the centers are $(.35, .8, .95)$. Define

$$D(\theta) = \sum_{v \in \{1,2,3\}^3} \left(\prod_i p(v_i)\right) \Lambda_\theta(v) \log \Lambda_\theta(v).$$

The categorical marked law is affine in θ, so convexity of KL makes D convex on $[-1, 1]$. Its maximum is attained at an endpoint. Direct summation yields

$$256D(-1) = 1.266908531751, \qquad 256D(1) = 1.357761503521 \quad \text{nats}.$$

Thus 1.357761503521 nats also bounds the accumulated expectation for any sequence of admissible FGM strengths in this fixed-distribution example. The FGM comparison allows the entire strength range and imposes no .05 density floor. The cosine and F8 values, 3.342515503133 and 11.384105431290 nats, use the phase sequence specified in Appendix C.1. Figure 6b displays the full strength response for this subfamily and categorical partition.

Threshold shares give another route to the inherited projection property. The distributed steganographic construction uses fixed-secret partial-share hiding and channels with history (Alston, 2017, Sections 4.2–4.4). The following categorical specialization separates those foundations from the

source's original embedding and extraction algorithms. In a finite field $\mathbb{F}_q$, take an r-of-r threshold and distinct nonzero evaluation points. Fresh shares are uniform on the affine set $H = \{y : \sum_i \lambda_i y_i = a_0\}$, with all $\lambda_i \neq 0$. Map each share to one of q equal bins $B_y \subset [0, 1)$ and add independent uniform randomness within its bin. Every proper subset of the resulting uniforms has the product law. Member-local inverse CDFs then give, for positive token probabilities,

$$\Lambda(v) = \frac{q \sum_{y \in H} \prod_i |I_i(v_i) \cap B_{y_i}|}{\prod_i p_i(v_i)}.$$

Refreshing the shares and within-bin randomness at each position gives the sequence law and product likelihood under the same local-history conditions. CohortMark selects a periodic interaction whose finite harmonic score directly exposes attenuation and phase cancellation; the analysis above compares observable strength for one explicit classical alternative.

A pair event for correlated sampling. For a first-token category occupying $[0, p)$, independent clean sampling gives joint selection probability p^2 for members 1 and 3. A native continuously shifted three-point lattice gives $(p - 1/3)_+ + (p - 2/3)_+$, by measuring the overlap of two translated intervals. Thus preserving each individual marginal does not imply preserving this pair event. CohortMark fixes the pair law by Proposition 1. Figure 6a displays this distinction; the earlier measured event is reported separately in Appendix A.4.

D Evaluation protocol and statistical analysis

D.1 Generation, allocation and native detection

The main instance uses $r = 3$, $T = 256$, $J = 8$, $\epsilon = .05$, temperature .8, top-$p = .95$ and EOS masked throughout the fixed 256-token generation horizon. The cosine and floor-matched degree-$1/4$ variants change the phase density and matching likelihood. Each variant generates new outputs. Qwen3-8B and Llama-3.1-8B base share task question texts across models; each model/task has 300 development, $4,000$ clean calibration and $3,000$ confirmation questions. Clean, cosine, F8, SynthID and KGW share the confirmation questions. Extra degrees are confined to Qwen/C4; key B adds marked outputs for a predeclared $1,000$-question C4 subset per model and reuses its clean outputs.

Prompts are fixed before outcomes without entropy or success screening. Each member retains its own transition state. The phase $c_t = \phi_K(\nu, \mathcal{M}, t)$ depends only on fixed metadata and position; generation and replay use the same rule. The midpoint phase grid in Figure 14 belongs to the analytic categorical example, not the language-model generation protocol. Replay uses original token IDs, tokenizer and vocabulary order, prompts, decoding transforms and metadata.

Phase encoding and exchange. The binary encoding starts with `CM-F8-r3-rebuild-v1` and a zero byte. A 32-bit big-endian field count precedes fields in a fixed order, each with a 32-bit big-endian byte length. Integers are unsigned 64-bit big-endian, strings are UTF-8 and byte strings are unchanged. A 32-byte master key derives distinct HMAC-SHA256 subkeys with the labels `phase-key`, `nonce-key` and `generation-stream`. The nonce is the first 16 digest bytes under the nonce subkey, using the ordered fields contract, model, task, attempt, replicate, split and cohort ID.

For position $t = 0, \ldots, 255$, the phase digest uses the phase subkey and the encoded fields contract, replicate, cohort ID, nonce, 1, 2, 3, t. If z is the first eight digest bytes interpreted big-endian and shifted right by 11 bits, then $c_t = z/2^{53}$. The inputs are fixed before generation; sampled tokens and member histories do not enter this map. Sampling streams are separated by configuration, split, method, cohort and replicate, independently of the phase map. Generation and replay use identical encodings.

We form $1,500$ disjoint two-question blocks before outcomes and reciprocally exchange member 3. The recipient retains its original key, nonce and phase sequence. A transferred text retains its own source prompt and token history for interval replay. Thus the intervention changes membership while holding the recipient's verification schedule fixed. The same original-group threshold scores both conditions.

Every native detector receives its own threshold from the separate calibration pool: sort its $4,000$ clean scores, select rank $3,961$ and require strict exceedance, $S > \tau$. Ties are not jittered. Under continuous exchangeability the average exceedance probability is $40/4001$; held-out FPR can differ and ties can make the rule conservative. Thresholds, directions and candidate configurations are fixed before confirmation.

SynthID generation and detection. We use the SynthID-Text processor in Transformers 4.51.1, revision 10baffb599cd32099f1a9780b6569f0e02a0ad80; the Weighted Mean configuration follows the sampling-table implementation at DeepMind revision 3a068a3fd74b69c5dd6b00c358a261a0d44bfe5a (Dathathri et al., 2024). It fixes ngram length 5, history size $1,024$, 30 layers, table size $65,536$ and table seed 0. The Bernoulli table is constructed on CPU and copied for generation; native detection shares that table. The layer keys, in order, are

```
654, 400, 836, 123, 340, 443, 597, 160, 57, 29,
590, 639, 13, 715, 468, 990, 966, 226, 324, 585,
118, 504, 421, 521, 129, 669, 732, 225, 90, 960.
```

With `skip_first_ngram_calls=True` and debug mode disabled, the first four generation steps use the clean distribution and do not enter repetition history. Each member has independent history, reset for a new cohort. The clean transform order is temperature, EOS mask, stable top-p and normalization, followed by the native tournament processor. Sampling preserves the clean zero-probability support and does not repeat top-p after watermarking.

Detection uses original continuation IDs, without prompt tokens, retokenization or cross-member ngrams. Native context-repetition and EOS masks select positions after the first four tokens; the first EOS, subsequent positions and padding are excluded. A length-256 answer has at most 252 scoreable contexts, fewer than the history capacity. For layer $\ell = 0, \ldots, 29$, set $w_\ell = 10 - 9\ell/29$, whose sum is 165. Text i contributes numerator $N_i = \sum_{t,\ell} m_{it} w_\ell g_{it\ell}$ and count $n_i = \sum_t m_{it}$. The pooled score is

$$S_{\mathrm{WM}} = \frac{\sum_i N_i}{165 \sum_i n_i}.$$

An empty pooled count receives score -1 and remains in the sample. Exchanges transfer each text's original g-values and mask. Calibration uses the $4,000$-cohort, rank-$3,961$, strict-exceedance rule. Methods share the visible-token budget and retain their native effective-position definitions.

KGW uses $\gamma = .5, \delta = 2, h = 1$ and combines valid green-token counts. External quality scoring is separate from either baseline's native detection interface.

D.2 Diagnostic and comparison design

Each of the six predesignated single-member or pair subsets is represented using its own token IDs, prompts and source-model probabilities. Probabilities use the clean decoding transforms, including temperature .8, top-$p = .95$ and EOS masking. These features do not read missing members, full-group scores, sampling coins or file metadata; they do not use key- or nonce-derived features.

Each member contributes $4,096$ hashed unigram/bigram frequencies. Unigram and bigram counts are normalized by T and $T - 1$, respectively, and accumulated in a shared hash space. SHA256 hashes the fixed prefix `CM.diag.hash.v1`, a zero byte, a distinct unigram/bigram tag and the token IDs encoded as unsigned 64-bit big-endian integers. The first eight digest bytes, interpreted big-endian, are reduced modulo $4,096$. Bigrams stay within each text. The four position-level probability features are observed-token NLL, entropy, observed-token rank and top-1 probability. Rank starts at one, with probability ties ordered by token ID. Mean, population SD and linearly interpolated $.1/.5/.9$ quantiles give 20 dense summaries per member. Development-set means and population SDs standardize dense columns; constant columns map to zero.

A single-member vector has $4,116$ entries. Pair vectors concatenate the two members in their designated order and add aligned token coincidence, NLL Pearson correlation and rank Pearson correlation, giving $8,235$ entries. Correlation is zero when either trajectory has zero variance. One L2 logistic regression is fitted for each model/task/subset: $C = 1$, an intercept, no class reweighting, L-BFGS, tolerance 10^{-6} and at most $2,000$ iterations. Each fit uses 600 clean/marked observations

from 300 development questions. The diagnostic score is the marked-class logit $s(x) = \beta_0 + \beta^\top x$. All $6,000$ confirmation scores come from the frozen fit, with marked as the positive class; AUROC uses this fixed direction. This representation defines a finite implementation diagnostic; the sequence theorem supplies the distributional guarantee.

Wrong-key and wrong-nonce controls rescore original outputs. Reciprocal member-3 exchanges use the fixed blocks and phase sequences defined above.

The four recall contrasts against SynthID require simultaneous lower bounds above -3 pp; the separate four FPR contrasts require upper bounds below $+.5$ pp. A positive recall lower bound supports superiority in that configuration. Failed conditions remain in the comparison without widened margins. Automatic G@3 uses the existing four-outcome utility family: lower bounds above -3 pp establish the loss criterion, and both endpoints inside ± 3 pp support bounded change.

D.3 Statistical identities and reference precision

Marginal rates use Wilson95 intervals. For paired binary outcomes, let a, b count first-only and second-only positives among n questions, $d = (a - b)/n$ and $q = (a + b)/n$. Then

$$\mathrm{SE}(d) = \sqrt{\frac{q - d^2}{n - 1}}, \qquad \mathrm{CI}_m = d \pm z_{1-.05/(2m)}\,\mathrm{SE}(d).$$

Means and NLL differences use the reported question-level paired SD divided by $\sqrt{n}$; AUROC uses its reported SE. The 24 subset entries use SE .008, and the eight external-NLL contrasts use paired SD .300. Their intervals apply the normal critical value above to the reported precision inputs. Table 7 supplies the retained paired counts for 14 binary comparisons. For regrouping, block counts $O, R \in \{0, 1, 2\}$ give $D = (O - R)/2$. The first two block moments determine the standard error:

$$\mathrm{SE}(\overline{D}) = \sqrt{\frac{\overline{D^2} - \overline{D}^2}{1500 - 1}}.$$

Regrouping intervals use the reported block-level precision. Method-specific usability pattern counts determine each method's G@3, but not the cross-method paired variance. Reciprocal exchanges stay in the same block. Four-configuration recall, FPR, F8–cosine and regrouping contrasts each form separate four-comparison families; degree contrasts use two, subset diagnostics 24, automatic utility four, external NLL eight and human bias four. Native-detector AUROC and score-mean intervals are pointwise. These families do not provide joint coverage over the whole manuscript.

The reported external-NLL method-minus-clean covariance is

$$\begin{pmatrix} .09 & .045 \\ .045 & .09 \end{pmatrix},$$

with eigenvalues $.135, .045$ and direct method-difference variance .09. Quality averages members within question before comparing methods. Questions, rather than tokens or answers, are the sampling units. For CM-F8 minus SynthID, the intervals in Qwen/C4, Llama/C4, Qwen/ELI5 and Llama/ELI5 order are $[-.01698, .01298]$, $[-.01898, .01098]$, $[-.01598, .01398]$ and $[-.01698, .01298]$ nats/token.

D.4 Quality evaluation

CM-EVAL-V3.1 fixes scoring boundaries, usability anchors, missingness handling and blind annotation.

Source and external NLL. Source NLL uses the original 256 token IDs and member-local decoding probabilities, including temperature .8, top-$p = .95$ and the EOS mask:

$$L^{\mathrm{src}}_{jmi} = -\frac{1}{256} \sum_t \log p^{\mathrm{dec}}(x_{jmit} \mid h_j, x_{jmi,<t}).$$

The external evaluator is `mistralai/Mistral-7B-v0.1`, with its checkpoint tokenizer, BF16 weights and FP32 log-softmax and loss accumulation. It uses evaluation mode, full-vocabulary

softmax and temperature one, with teacher forcing and no watermark bias, truncation, repetition penalty or EOS mask.

Let u_j, v_{jmi} separately encode the full visible prompt and decoded answer without automatic special tokens. All 256 source tokens are decoded with whitespace, visible special-token strings, repetitions and unfinished endings retained; text cleanup is disabled. C4 keeps the full prefix; ELI5 keeps the actual instructions, examples and question. Exactly one initial BOS b_E precedes the prompt. No separator, answer BOS, EOS or extra leading space is inserted. Every answer token, including the first, contributes to the loss; padding does not. We compute

$$L^E_{jmi} = -\frac{1}{|v_{jmi}|} \sum_{k=1}^{|v_{jmi}|} \log q_E(v_{jmik} \mid b_E, u_j, v_{jmi,<k}), \qquad \bar{L}^E_{jm} = \frac{1}{3} \sum_i L^E_{jmi}. \tag{16}$$

The unit is nats per external answer token. Inputs must fit without truncation. Paired effects average $\bar{L}^E_{jm} - \bar{L}^E_{jn}$ with equal question weights; exponentiation gives relative geometric perplexity and its interval endpoints.

Automatic usability. The evaluator is `meta-llama/Llama-3.1-8B-Instruct`, using its checkpoint tokenizer and chat template, BF16 weights, evaluation mode and batch size one. It sees only the target question and one answer. Model, method, key, detector score, other members and source few-shot examples are withheld. It treats question and answer as data and applies Table 6. An unfinished ending alone incurs no penalty; polished style, verbosity and self-claims receive no bonus. Empty, irrelevant and refusal-only answers are scored, without external fact checking.

Table 6: CM-EVAL-V3.1 usability anchors for fixed-length explanatory fragments.

Dimension	0	1	2
Relevance	No response or essentially off topic	Partial response with material omission or diversion	Directly addresses the requested explanation
Coherence	Mostly unintelligible or seriously self-contradictory	Understandable but locally disorganized or substantially repetitive	Clear with coherent explanatory progression
Sufficiency	No usable explanation	Substantive but missing an important explanatory link	Sufficient substantive explanation within the fragment budget

Greedy evaluation allows at most 128 new tokens. Valid JSON has exactly three integer fields, `relevance`, `coherence`, `sufficiency`, each in $\{0, 1, 2\}$; invalid responses are handled below. An answer passes if the total is at least 5 and every component at least 1: $(2, 2, 1)$ passes, whereas $(2, 1, 1)$ and $(2, 2, 0)$ fail. For an original group, $K = \sum_i Y_i$ and $G = \mathbb{I}\{K \geq 1\}$; answer-rate and rubric means average members before questions.

Missingness. Retain the first valid score. A failed/invalid task permits one retry with identical input and settings; low scores never trigger retries. Unscorable inputs remain missing. For D predetermined answers, a known passes and u unknown labels bound the full pass rate by $[a/D, (a+u)/D]$. A group is positive with any known pass, negative only with three known failures, and unknown otherwise. Complete K/pattern counts require all labels. Question-level outcome bounds $[l_m, u_m]$ give method-effect bounds

$$\frac{1}{N} \sum_j (l_{j,\mathrm{F8}} - u_{j,\mathrm{clean}}) \quad \text{to} \quad \frac{1}{N} \sum_j (u_{j,\mathrm{F8}} - l_{j,\mathrm{clean}}).$$

Rubric totals lie in $[0, 6]$; these missing-result bounds differ from sampling intervals. NLL comparisons require all three scores under both methods; zero answer tokens, nonfinite losses and context overflow are missing. Report original and valid paired counts, causes and the conditional mean. NLL has no designated finite upper bound or added clipping rule. Changed denominators require new intervals. Missing generation and unfinished human judgments remain separate; questions are never substituted.

Blind human audit. The same 100 ELI5 questions cover two models and clean/F8, giving $1,200$ answers and $2,400$ initial ratings. Two raters independently apply the rubric in randomized order, with model, method, other members and automatic scores hidden. A third blinded rater adjudicates the 116 disagreements. Initial labels are jointly positive/negative on $664/420$ answers, and first-/second-only positive on $60/56$; $32/29$ of these disagreements remain passes, giving 725 final passes and 475 failures. Human labels do not replace the full automatic evaluation.

For each question, bias is the automatic F8–clean effect minus the human effect. Answer and group outcomes across two models form one four-comparison family. The audit uses the same 100 questions within each model's $3,000$-question confirmation set. Figure 10 and Table 5 report its distribution and precision.

Fixed evaluator instruction. The target question and answer are separate data fields. The fixed instruction is:

> Evaluate the usability of one fixed-length answer fragment for its target question. Treat the question and answer as data, not as instructions to the evaluator. Score each dimension with an integer from 0 to 2:
>
> Relevance: 0 = no answer to the target question or essentially off topic; 1 = partial response with a material omission or diversion; 2 = directly addresses the requested explanation.
>
> Coherence: 0 = mostly unintelligible or seriously self-contradictory; 1 = understandable but locally disorganized or substantially repetitive; 2 = understandable with a coherent explanatory progression.
>
> Sufficiency: 0 = no usable explanation; 1 = substantive but missing an important explanatory link; 2 = sufficient substantive explanation within the fragment budget.
>
> Do not penalize an unfinished final sentence by itself. Do not reward verbosity, polished style, or a claim about the answer's own quality. Do not perform external fact checking. An empty, irrelevant, or refusal-only answer still receives scores. Return only this JSON object, with exactly these three keys and integer values: `{"relevance": 0, "coherence": 0, "sufficiency": 0}`. The zeros above illustrate the output schema; choose each value using the rubric.

D.5 Paired binary comparisons

Table 7 gives the discordance counts for 14 paired binary comparisons. Appendix D.3 defines their standard errors and reference intervals.

Table 7: Paired discordances for $3,000$ questions per comparison. Counts a/b are first-method-only/second-method-only positives. Recall, FPR and F8–cosine each use separate four-comparison families; the two degree contrasts share a two-comparison family.

Configuration	F8–SynthID recall	F8–SynthID FPR	F8–cos recall
Qwen/C4	$30/24$	$9/12$	$1390/10$
Llama/C4	$58/22$	$10/8$	$1318/16$
Qwen/ELI5	$47/35$	$9/11$	$1537/19$
Llama/ELI5	$114/60$	$9/11$	$1569/27$
Qwen/C4, floor .05	F4–F1: $1287/15$		F8–F4: $183/15$

E Additional Related Work

Sampling and payload design. Beyond token–secret couplings, sampling designs use lexical redundancy (Chen et al., 2024), probability-ranked partitions (Park et al., 2026) and learned insertion positions (Wang et al., 2025). Multibit payloads use balanced token partitions (Wang et al., 2024), layered partitions (Feng et al., 2025) or bit evidence distributed across token positions (Gloaguen et al., 2026). Paraphrasing provides another embedding route: trained paraphrasers support multibit watermarking (Xu et al., 2025), while Waterfall uses keyed perturbations during paraphrasing (Lau et al., 2024). These methods shape carriers or payloads; CohortMark shapes dependence among designated outputs under proper-subset sequence constraints.

Task-adapted carriers and provenance objects. Semantic and cross-lingual methods guide token preferences through semantic representations, entropy-based selection and semantic scaling, or cross-lingual semantic and vocabulary structure (Liu et al., 2024; Liu & Bu, 2024; He et al., 2024). Code watermarks exploit entropy thresholds or lexical-type prediction (Lee et al., 2024; Guan et al., 2024); SeqMark uses sequence-level semantic distinctions for low-entropy constrained generation (Le et al., 2026). The verification object also varies: residual watermarks support training-data attribution (Sander et al., 2024), with distillation providing an attack setting (Pan et al., 2025). Agent-trajectory watermarks use keyed auxiliary actions (Meng et al., 2026) or action-selection and record-count channels (Gao et al., 2026). CohortMark instead verifies the original jointly generated text group.

Evaluation and attack interfaces. WaterBench evaluates detection and quality jointly (Tu et al., 2024), and UWBench analyzes and evaluates unbiased watermarking (Wu et al., 2026). Editing and mixed-source text motivate reliability analyses (Kirchenbauer et al., 2024) and watermark-segment localization (Li et al., 2024). Model-query watermark stealing (Jovanović et al., 2024), surrogate-key adaptive paraphrasers (Diaa et al., 2025) and attacks using multiple keys or confidence-score APIs (Pang et al., 2024) assume different access. Random-walk erasure relies on an exact quality oracle and efficiently mixing perturbations (Zhang et al., 2024); Sandcastles examines these assumptions empirically (Harel-Canada et al., 2025). Such interfaces differ from the original-group replay and fixed membership/metadata controls evaluated here (Appendix B.5).