WHAT SURVIVES A REWRITE? INFORMATION BUDGETS FOR INHERITED WATERMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Watermarked text often reaches a verifier after another model has rewritten it. That model may add a new watermark, but the upstream key reaches it only through the inherited document. We ask what survives such a rewrite. The coupling of old and new text has three information budgets: erased, transmitted, and fresh. A key enters the new text only through the fresh budget and leaves only through the erased one; the inherited key, which cannot enter, is capped by the transmitted budget. A task bounds the fresh budget through the outputs it allows and the erased budget through the inputs each output tolerates, and what every admissible output reveals fixes a retention floor. When fidelity means preserving a statistic T, the floor is exactly $I(K_1;T)$, attained by conditional resampling that preserves the text distribution. For binary symmetric rewriting, we derive the exact injection optimum for uniform keys of every bit length. Failures cost at most the key alphabet, and a sharp loss bound for nearly exact copying is independent of document length. Finite-alphabet experiments exhibit the optimal operations and strict gaps below the budgets. Controlled experiments on two language models separate task execution from detector survival: Qwen3-4B rewrites all four target facts correctly in $64/64$ reports, while inherited-key detection falls from $14/64$ upstream to $1/64$ in each sampling arm. Paired 200-document comparisons isolate re-watermarking, and complete three-hop trajectories show that early stopping leaves hop-3 detection counts unchanged.

> *"You cannot step twice into the same rivers; for fresh waters are ever flowing in upon you."* (explained in Appendix I) — Heraclitus, tr. John Burnet (1920)

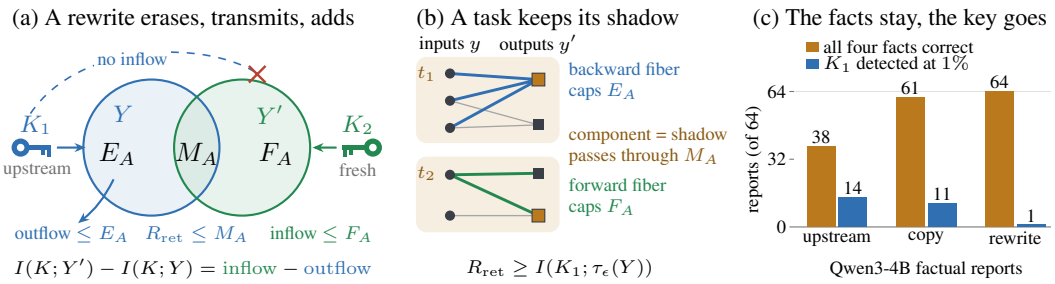

Figure 1: What survives a rewrite. (a) The coupling of the inherited text Y with the new text Y' has three cells. A key enters Y' only through the fresh cell and leaves only through the erased one; the inherited key K_1, which the rewriter never sees, cannot enter and is capped by the transmitted cell (Theorem 1). (b) A task joins each input to its admissible outputs. Backward fibers cap the erased cell, forward fibers cap the fresh cell, and every admissible output determines the component of its input, the shadow τ_ϵ, whose key information therefore survives (Theorem 2). When a task preserves facts t, the components are their classes; an output common to a component (amber) attains the floor. (c) Qwen3-4B rewrites all four facts in $64/64$ reports, while inherited-key detections fall from $14/64$ to $1/64$ (Section 6).

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.

1 INTRODUCTION

Watermarking couples generated text to a secret key. A verifier who holds the key can test for this dependence even when the text otherwise resembles an ordinary model sample (Kirchenbauer et al., 2023a; Aaronson, 2023; Kuditipudi et al., 2023; Christ et al., 2023). In a provenance chain, however, the verifier may not see the original text. The document may have been translated, lightly edited, paraphrased, or summarized by another model. Evidence of the first generator must then survive an operator that never possessed its key.

The same operator may also embed a new watermark. Encoder-controlled analyses connect embedding to the randomness available during generation (He et al., 2025; Huang et al., 2024; He et al., 2026). When one encoder holds both secrets, it can allocate a shared budget between them (Li et al., 2026, Prop. 3.13). The inherited setting starts after the first text has been written: the new key is available at generation time, while the old key reaches the operator only through that text. A task such as copying or rewriting further restricts what the operator may change.

What survives a rewrite? Before any watermark is considered, every rewrite produces a coupling of the old text with the new one. Its information diagram has three cells: part of the input is erased, part is transmitted, and part of the output is fresh. A key can enter the new text only through the fresh cell and leave only through the erased one; the inherited key, which cannot enter, is capped by the transmitted cell. A task constrains the coupling from both ends: the outputs it allows bound the fresh cell, and the inputs each output tolerates bound the erased cell. What every admissible output reveals is a partition of the texts; its key information is always protected, and exactly so when each block has a common admissible output, as for tasks that preserve a statistic. Evidence encoded only in freely changeable style may therefore disappear under a perfectly faithful transformation.

Present work. We describe each downstream transformation by its three budgets and each task by its fibers and shadow (Figure 1). The results then follow from short arguments:

1. **How keys flow.** Injection is capped by the fresh budget, retention by the transmitted budget, and retention loss by the erased budget, each with an exact gap (Section 3). For binary symmetric rewriting we derive the exact injection optimum under uniform b-bit keys and its sharp gap to the optimized-key value (Section 4).
2. **What a task protects.** Forward fibers cap injection, backward fibers cap retention loss, and the shadow fixes a retention floor. When fidelity means preserving a statistic T, the floor is exactly $I(K_1;T)$, attained by conditional resampling that preserves the text distribution. Failures cost at most the key alphabet, and conditional-entropy continuity gives a sharp bound for nearly exact copying (Section 5).
3. **Direct information tests and controlled task comparisons.** Finite-alphabet experiments compute the budgets, gaps, and equality cases from complete joint laws. On two language models, 200-document paired comparisons isolate re-watermarking, factual reports separate four-field success from detector survival, and complete three-hop trajectories test an early-stopping rule; an operation study and a six-model copy ladder add context (Section 6).

2 RELATED WORK

Information-theoretic watermarking studies distributional embedding, entropy budgets, and statistical detection (He et al., 2025; 2026; Huang et al., 2024; 2023; Li et al., 2024; Cai et al., 2024). The two-secret construction of Li et al. (2026, Prop. 3.13) gives an encoder-controlled shared-budget region; in the inherited interface, independent keys and key blindness instead separate the budgets.

The optimization questions connect to established information theory: fixed-channel injection is a perfect-privacy problem (Section 4), and task-preserving removal is related to the privacy funnel (Makhdoumi et al., 2014); our task-partition result identifies an exact optimizer under a hard preservation constraint. The sharp copy guarantee applies the conditional-entropy continuity bound of Alhejji & Smith (2020).

Empirical survival after paraphrase (Kirchenbauer et al., 2023b), bounded-edit guarantees (Zhao et al., 2023; Kuditipudi et al., 2023), watermark collisions (Luo et al., 2024), quality-preserving

removal (Zhang et al., 2023), and multi-step rewriting (Ameen et al., 2026) examine particular channels or schemes. Task-dependent detection is studied by Bahri & Wieting (2025); Le et al. (2026); He et al. (2024); Tu et al. (2023). Our operation-level analysis separates the averaged channel, its keyed realization, the task constraint, and the eventual verifier. Section F.2 gives further context.

3 The Three Budgets of a Rewrite

A downstream model sees the inherited document, but not the key that created it. We model this interface as an operation that receives an instruction X, a document Y, and an independent fresh key K_2, and then emits Y'. The upstream key K_1 remains hidden from the operator and can influence Y' only through Y.

Formally, let $\mathcal{V}$ be a finite vocabulary and $\mathcal{Y} \subseteq \mathcal{V}^{\le n}$ a finite text space. We fix an instruction $X = x$; all laws and the task relation are those given x, and we suppress this conditioning when it is clear. All entropies and mutual informations are in nats; h denotes binary entropy and $d(a\|b)$ binary relative entropy.

Definition 1 (Operation with an inherited key)**.** An *operation* consists of (i) an inherited document $Y \sim Q^{\mathrm{up}}_{K_1}$ produced upstream under a key $K_1 \sim \mu_1$ that the operator does not hold; (ii) a fresh key $K_2 \sim \mu_2$ with $K_2 \perp (K_1, Y) \mid X$; and (iii) a generator producing $Y' \sim G_{K_2}(\cdot \mid x, Y)$, so that

$$Y' \perp K_1 \;\big|\; (Y, X). \tag{1}$$

The operation is ϵ-*faithful* if $Y' \in \mathcal{Y}_\epsilon(Y)$ almost surely, where $\mathcal{Y}_\epsilon(y)$ is the set of outputs counting as a correct execution of the task on input y. Here ϵ indexes a specified task relation; it is not a numerical distortion level unless the task supplies a metric and a threshold.

By (1), Y is the only route by which K_1 can reach Y'; auxiliary randomness is part of the generator, which may otherwise transform Y arbitrarily. Re-using a downstream key in later hops breaks this structure (Example 1). Every operation couples the old text with the new one, and the information diagram of this coupling has three cells.

Definition 2 (The three budgets)**.** The *erased*, *transmitted*, and *fresh* budgets of an operation are

$$E_A = H(Y \mid Y', X), \qquad M_A = I(Y; Y' \mid X), \qquad F_A = H(Y' \mid Y, X), \tag{2}$$

so that $H(Y \mid X) = E_A + M_A$ and $H(Y' \mid X) = M_A + F_A$. The injected and retained information are $C_{\mathrm{inj}} = I(Y'; K_2 \mid X)$ and $R_{\mathrm{ret}} = I(Y'; K_1 \mid X)$.

The budgets belong to the operation, not to any watermark: exact copying has $E_A = F_A = 0$, and input-independent generation has $M_A = 0$. The key-averaged channel $A(y' \mid y) = \mathbb{E}_{K_2} G_{K_2}(y' \mid y)$ and the input law determine all three budgets and, with the upstream law, retention; the keyed realization of A also determines injection.

Theorem 1 (Key flow)**.** *For every operation in the sense of Definition 1 and every random variable K defined with it,*

$$I(K; Y') - I(K; Y) \;=\; \underbrace{I(K; Y' \mid Y)}_{\text{inflow} \le F_A} \;-\; \underbrace{I(K; Y \mid Y')}_{\text{outflow} \le E_A}. \tag{3}$$

The fresh key starts from $I(K_2; Y) = 0$, and key blindness gives the inherited key zero inflow. Hence

$$C_{\mathrm{inj}} \;\le\; F_A, \qquad I(Y; K_1) - E_A \;\le\; R_{\mathrm{ret}} \;\le\; \min\big\{M_A,\; I(Y; K_1)\big\}, \tag{4}$$

with exact gaps

$$C_{\mathrm{inj}} = F_A - H(Y' \mid Y, K_2) - I(Y; K_2 \mid Y'), \tag{5}$$

$$R_{\mathrm{ret}} = M_A - I(Y'; Y \mid K_1), \qquad I(Y; K_1) - R_{\mathrm{ret}} = E_A - H(Y \mid Y', K_1)\,. \tag{6}$$

Thus $C_{\mathrm{inj}} = F_A$ iff the operation is deterministic given (Y, K_2) and $Y \perp K_2 \mid Y'$; $R_{\mathrm{ret}} = M_A$ iff $Y \perp Y' \mid K_1$; the loss equals E_A iff Y is a function of (Y', K_1); the last two hold whenever $H(Y \mid K_1) = 0$.

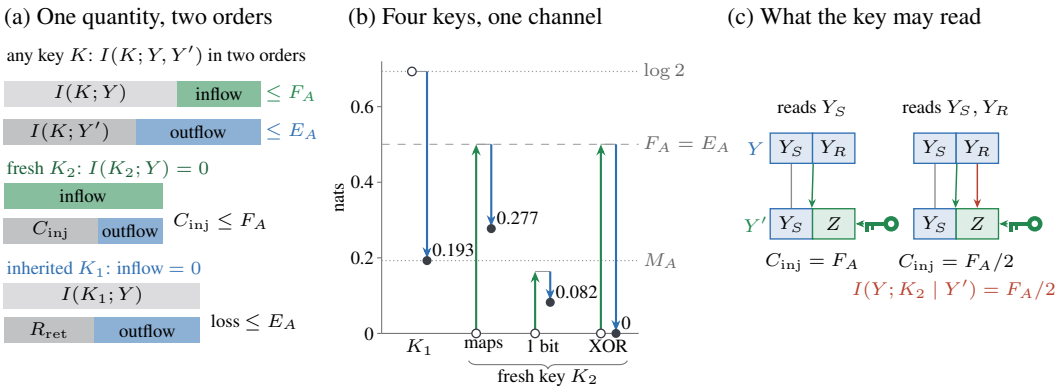

Figure 2: Every key obeys one balance. (a) The proof of Theorem 1 expands $I(K;Y,Y')$ in its two orders. The two bars have the same length, so $I(K;Y') - I(K;Y)$ is inflow minus outflow, and the conditional terms are bounded by F_A and E_A. Independence removes the first term for the fresh key; key blindness removes the inflow of the inherited key. (b) The binary symmetric channel of Section 4 with $Y = K_1$ uniform and $\delta = 0.2$. Open circles mark $I(K;Y)$, green arrows the inflow, blue arrows the outflow, and dots $I(K;Y')$. The inherited key loses exactly E_A and keeps exactly M_A. The optimized maps and the XOR realization both take in the whole fresh budget, but XOR returns all of it. (c) The corner of Corollary 1 with $Y = (Y_S, Y_R)$ two uniform bits: a continuation Z keyed on the kept part fills F_A; one keyed on the whole input spends half the fresh budget on the discarded bit.

Every key obeys the same balance: it enters the text only through what is fresh and leaves only through what is erased; the two keys differ only in where they start (Figure 2a). Independence starts the fresh key at zero, key blindness closes the inherited key's inlet, and data processing along $K_1 \to Y \to Y'$ pins its retention to M_A up to the upstream sampling entropy, since $I(Y;K_1) - E_A = M_A - H(Y \mid K_1)$ (Section A.1).

Each gap has an operational meaning: unkeyed randomness and the fresh key's outflow, key effort spent on discarded input, waste fresh budget; transmitted information unrelated to K_1 cannot carry it; and erased input that K_1 does not determine costs nothing. Both caps are attained at once when the operation keeps a prefix of a key-determined document and continues it with K_2, reading only that prefix (Corollary 1); reading the discarded suffix can halve injection (Figure 2c; Section A.2).

A verifier sees these quantities only through detectors. By data processing, a detector applied to (Y', K_1) gives $R_{\text{ret}} \ge d(\text{TPR}\|\text{FPR})$, with the false-positive probability evaluated under the product of the output and key marginals. A realized score such as $-\ln p$ describes one verification outcome, not the information itself; we keep the two levels separate in the experiments.

4 How Much of the Fresh Budget a Key Can Use

The fresh budget is a property of the channel, not of the key. Keyed realizations of one averaged channel retain the same upstream information but can carry different amounts of new-key information. Optimizing over all finite independent key laws is the full-data perfect-privacy problem of Rassouli & Gündüz (2023), solved by a deterministic-map linear program (Section A.6). By (5), its gap below F_A is the excess functional information of Li & El Gamal (2018), so even an optimal key need not fill the fresh budget.

A binary channel makes the resource distinction explicit. Let $Y = K_1$ be uniform and let A flip its input with probability $\delta \in [0, 1/2]$. The largest injection is

$$J_{\text{free}}(\delta) = 2\delta \log 2, \qquad J_{\text{bit}}(\delta) = \log 2 - h(1/2 + \delta), \tag{7}$$

when the key law is optimized or fixed to one fair bit, respectively. Both retain $\log 2 - h(\delta)$, and both lie strictly below $F_A = h(\delta)$ in the interior. For a prescribed b-bit key, we can characterize the optimum exactly.

Proposition 1 (Uniform finite-key injection). *Let $Y = K_1$ be a fair bit, let K_2 be an independent uniform key on $m = 2^b$ values with $b \in \mathbb{N}$, $b \ge 1$, and fix the averaged channel to a binary symmetric*

channel of crossover $\delta \in [0, 1/2]$. Among stochastic keyed realizations of this channel, the largest injection is

$$J_m(\delta) = \frac{2}{m}\left[j \log 2 + \log 2 - h\left(\frac{1+r}{2}\right)\right], \qquad j = \lfloor m\delta \rfloor, \quad r = m\delta - j. \tag{8}$$

Moreover, $0 \le J_{\text{free}}(\delta) - J_m(\delta) \le 2\log(5/4)/m$, and the last constant is sharp.

The proof reduces the channel constraints to balanced output biases, then uses convexity to concentrate them (Figure 8). An optimal realization assigns j keys to each constant output, two keys to fractional channels when $r > 0$, and the remaining keys to copying. This construction proves achievability under the prescribed key law (Section A.7). Section 6.1 computes these realizations directly.

5 What a Task Protects

A task limits the outputs an input may receive, and hence the inputs an output may have come from. These two directions, and one partition they generate, give the three bounds below.

Definition 3 (Fibers and shadow of a task)**.** Let S be the support of Y. The *forward fiber* of an input y is $\mathcal{Y}_\epsilon(y)$; the *backward fiber* of an output y' is $B_\epsilon(y') = \{y \in S : y' \in \mathcal{Y}_\epsilon(y)\}$, and

$$N_\epsilon = \max_{y'} |B_\epsilon(y')|. \tag{9}$$

The *shadow* τ_ϵ is the finest function of the input that every admissible output determines: $\tau_\epsilon(y) = \tau_\epsilon(\tilde{y})$ whenever y and $\tilde{y}$ share an admissible output. Equivalently, $\tau_\epsilon(y)$ is the connected component of y in the bipartite graph that joins each $y \in S$ to every $y' \in \mathcal{Y}_\epsilon(y)$, the Gács–Körner common part (Gács & Körner, 1973) of the task relation.

Exact copying has singleton fibers and $\tau_\epsilon(y) = y$. Lower-casing has singleton forward fibers, backward fibers of all case variants, and shadow equal to the lower-cased text. A task whose entire requirement is to preserve a statistic T, with $\mathcal{Y}_\epsilon(y) = \{z : T(z) = T(y)\}$, has shadow T.

Theorem 2 (What a task protects)**.** *For every almost surely ϵ-faithful operation and every upstream scheme with support S:*

(i) forward fibers cap the fresh budget: $F_A \le \mathbb{E}\log|\mathcal{Y}_\epsilon(Y)|$, hence $C_{\text{inj}} \le \mathbb{E}\log|\mathcal{Y}_\epsilon(Y)|$;

(ii) backward fibers cap the erased budget: $E_A \le \log N_\epsilon$, hence

$$R_{\text{ret}} \ge I(K_1; Y) - \log N_\epsilon; \tag{10}$$

(iii) the shadow passes: $\tau_\epsilon(Y)$ is a function of Y', hence $H(\tau_\epsilon(Y)) \le M_A$ and $R_{\text{ret}} \ge I(K_1; \tau_\epsilon(Y))$.

If every component C contains some $z_C \in \bigcap_{y \in C \cap S} \mathcal{Y}_\epsilon(y)$, the minimum retention over faithful operations is $I(K_1; \tau_\epsilon(Y))$, attained by $y \mapsto z_{\tau_\epsilon(y)}$. For a task that preserves T,

$$\min_{A:\, T(Y')=T(Y) \text{ a.s.}} I(K_1; Y') = I(K_1; T(Y)), \tag{11}$$

attained by resampling $Y' \sim P_{Y|T}$ given $T(Y)$, which also preserves $P_{Y'} = P_Y$.

Parts (i) and (ii) are one line each plus Theorem 1: given Y, the output lies in its forward fiber; given Y', the input lies in its backward fiber. Part (iii) is data processing, since every realized pair is an edge of the graph. Neither (ii) nor (iii) uses key blindness: fidelity alone forces these floors, even on an operator that knows K_1 (Sections A.3 and A.4).

The two directions of the task belong to the two keys: forward fibers limit what an output may add, and so what a fresh key can carry; backward fibers limit what it may erase, and so what the inherited key can lose. A forward ball and a pre-image are not rival measures of a task's size (Section A.11).

Counting and structure give complementary floors. Bound (ii) is sharp for copying, where $N_\epsilon = 1$; bound (iii) is sharp for tasks that preserve a statistic, and there dominates (ii) because $I(K_1; Y \mid$

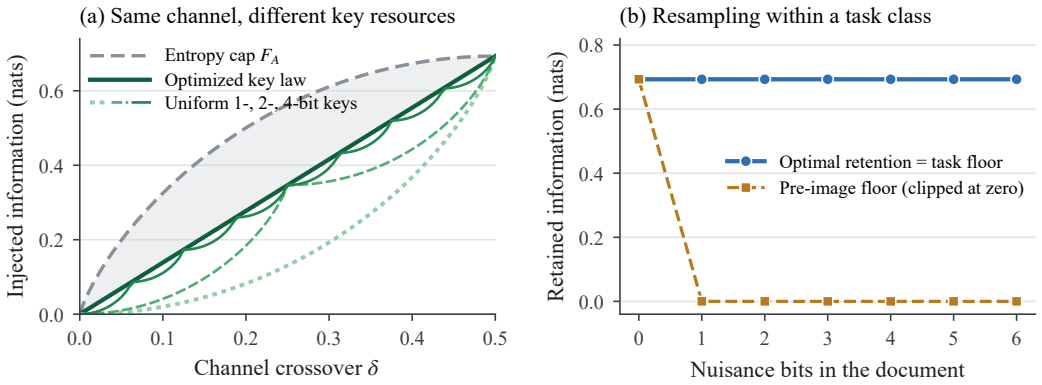

Figure 3: Exact mutual information from finite joint laws. (a) A fixed binary symmetric channel has different attainable injection under optimized and prescribed uniform key laws; increasing key size approaches the optimized value, not the entropy cap (shaded gap). (b) Task-conditioned resampling attains the task floor even as nuisance information makes the pre-image bound vacuous. Lines are analytic values, checked by finite enumeration and, in (a), a deterministic-map linear program.

$T) \le H(Y \mid T) \le \log N_\epsilon$. For lower-casing, (ii) degrades with the number of case variants, while (iii) keeps exactly the key information in the lower-cased text; without a common output per component, (iii) can be strict (Section A.3). A faithful operation can remove at most $I(K_1; Y \mid T)$, so a key independent of T can be erased completely without changing the task statistic or the output distribution.

5.1 Imperfect execution and finite keys

A task may fail, and the inherited key has its own finite alphabet; failures are charged to the key.

Proposition 2 (Failures cost at most the key alphabet)**.** *Assume $N_\epsilon \ge 1$, let p be the actual failure probability, let K_1 have an alphabet of size $M \ge 2$, and put $a = \min\{\log N_\epsilon, \log M\}$. Then*

$$I(K_1; Y) - R_{\mathrm{ret}} \le h(p) + (1-p)a + p \log M. \tag{12}$$

If only $p \le \eta \le 1/2$ is known, replacing p by η is valid.

On successes the output leaves at most $\log N_\epsilon$ nats of input ambiguity; on failures the lost *key* information is at most $\log M$, irrespective of document length (Section A.3, which also gives the Fano correction for the shadow).

For copying, $N_\epsilon = 1$ and (12) reads $h(p) + p \log M$. Conditional-entropy continuity (Alhejji & Smith, 2020) sharpens $\log M$ to $\log(M-1)$ and shows the result is tight: if $\Pr[Y' \ne Y] \le \eta \le 1 - 1/M$, then

$$I(K_1; Y) - R_{\mathrm{ret}} \le g_M(\eta) := h(\eta) + \eta \log(M-1), \tag{13}$$

with $g_M(\eta) = \log M$ for larger η, and uniform $K_1 = Y$ passed through an M-ary symmetric channel attains equality. With a 63-bit key and $\eta = 0.02$, the loss budget is below 0.972 nats, rather than a penalty proportional to the text space; an absolute guarantee still needs a lower bound on $I(K_1; Y)$. Along fresh-randomness pipelines the losses add hop by hop, each at most its hop's erased budget, while transmitted budgets only shrink (Corollary 4); substitutions are priced by pre-image degree (Corollary 3).

6 Experiments

We first compute the information quantities exactly in finite models where the full joint law is known, then study how inherited detector evidence behaves in language-model rewriting.

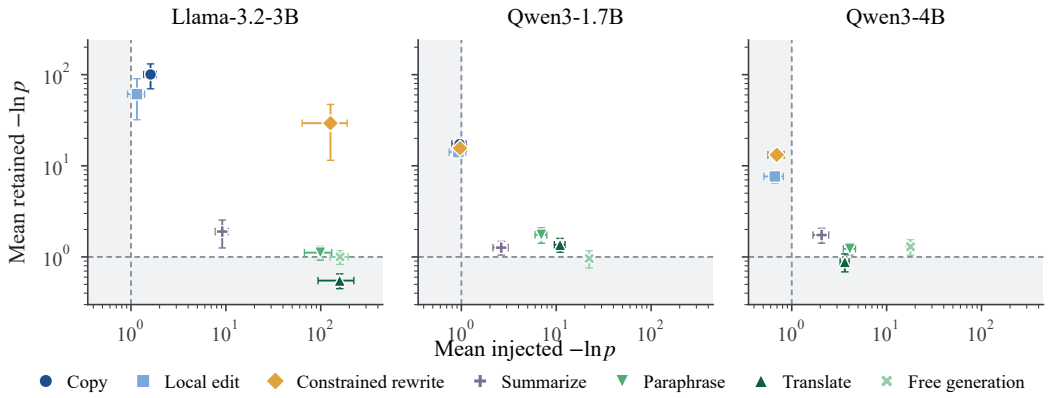

Figure 4: Injected and retained detector scores at hop 1 in the fixed-key operation study, with a second race watermark on every output. Points and standard-error bars show the three models; dashed lines mark the null reference 1; scores near it carry no detectable evidence. The axes measure mean $-\ln p$, not mutual information.

6.1 Direct information measurements

Same channel, different injection. At crossover $\delta = 0.2$, the fresh budget is 0.5004 nats, but maximal injection is 0.2773 with an optimized key law and 0.0823 with a fair-bit key. The XOR realization $Y' = Y \oplus K_2$, $K_2 \sim \text{Bernoulli}(\delta)$, injects zero. All three induce the same channel and retain 0.1927 nats (Figure 2b). Over 101 crossover values and uniform keys of $b \in \{1, 2, 4, 6\}$ bits, the linear-program optimum agrees with the optimized-key closed form, and each prescribed-key construction attains (8) (Figure 3a).

Same task, removable and unavoidable information. Task-conditioned resampling attains the shadow floor of Theorem 2. To make the two parts explicit, let $Y = (C, W, U)$, where C, W are independent fair bits and U contains r independent nuisance bits. Set $K_1 = (C, W)$ and require $T(Y) = C$ to remain unchanged. Resampling (W, U) retains exactly $\log 2$ nats for every $r = 0, \ldots, 6$, while preserving P_Y. The counting floor (10) is already zero at $r = 1$ (Figure 3b). With a style-only key $K_1 = W$, the same operation retains zero. Finally, symmetric channels with key alphabets of size 2, 3, and 8 attain the sharp copy-loss bound (13) with equality.

6.2 Operations on the injection–retention plane

Seven operations, two coordinates. We first place seven operation prompts on detector-score analogues of the two coordinates of Definition 2. In a fixed-key operation study, three models each write 20 documents under an exponential-race watermark (Aaronson, 2023). Each prompt then transforms these documents under a second race key, and every output is scored under both keys (Figure 4). Copying and local editing, whose target output is unique, keep the inherited evidence but inject little, as singleton forward fibers leave no fresh budget: on Llama-3.2-3B, copying retains a mean score of 100.54 and injects 1.61. Translation, paraphrase, and free generation reverse the roles; Llama translation retains 0.55 and injects 158.42. Llama's constrained rewrite lies between them, with both scores well above the null. This is consistent with the caps of Theorem 1. The Qwen models inject less, consistent with their smaller fresh-entropy estimates (Table 7), and Qwen3-4B largely copies under the constrained-rewrite prompt. Sections F and G give the complete table and further diagnostics.

6.3 Task execution and inherited evidence

A controlled comparison. We evaluate Qwen3-4B and Llama-3.2-3B-Instruct in three prespecified studies: 200 upstream documents per model for paired copying, paraphrasing, and summarization; 64 for complete three-hop trajectories; and 64 factual reports. Each exponential-race watermarked upstream document feeds ordinary sampling (plain) and a fresh race-watermarked operation (race), with document-specific keys. We report all 5,360 generations, including failed tasks and initially undetected documents. We measure inherited evidence by $-\ln p$ and key detection at 1%. Confi-

Table 1: Task execution and inherited-key detection on all 64 factual reports per model and arm. Fields counts strict JSON outputs with all four target values; key counts whole-output detections at 1%. The first row describes the upstream reports before any operation. Failed outputs remain in every denominator.

		Qwen3-4B		Llama-3.2-3B	
Operation	Arm	Fields	Key	Fields	Key
Upstream	–	38/64	14/64	49/64	61/64
Copy	Plain	61/64	11/64	51/64	61/64
Copy	Race	61/64	11/64	51/64	61/64
Rewrite	Plain	64/64	1/64	40/64	1/64
Rewrite	Race	64/64	1/64	38/64	3/64
Task-only	Plain	27/64	0/64	35/64	2/64
Task-only	Race	24/64	0/64	38/64	0/64

dence intervals use a document-level bootstrap that keeps paired outputs together. Section H gives the prompts, length budgets, and complete diagnostics.

Specified facts and detector survival separate. The factual task makes the preservation requirement explicit. Each report contains an organizer, city, year, and attendance count, followed by a requested 60–90-word narrative. We compare copying, full-context rewriting, and task-only regeneration from the parsed four fields. Table 1 puts strict four-field success beside inherited-key detection; its first row gives the upstream reports. On Qwen, both rewrite arms return all four target values in $64/64$ reports, with key detection on $1/64$; copying gives $61/64$ and $11/64$, respectively. Rewriting can also repair malformed inputs: only $38/64$ upstream reports pass the field test. Within that upstream-success subset, both rewrite arms retain the fields on $38/38$ reports with $0/38$ key detections. Field success does not imply narrative compliance: only $10/64$ plain and $11/64$ race rewrites meet the requested word range.

Llama starts with stronger detectable evidence: $61/64$ upstream reports are detected, compared with $14/64$ for Qwen. Copying retains Llama's $61/64$ detections; rewriting reduces them to $1/64$ (plain) or $3/64$ (race), while four-field success falls from copying's $51/64$ to $40/64$ or $38/64$. Thus the same operation can reduce detector response with different consequences for the task.

Task-only regeneration tests a concrete input restriction motivated by the partition result: it receives the parsed facts, not the source narrative. It produces few detections but also fewer correct reports than full-context rewriting on Qwen. Its input depends on parsing the source, so the baseline reflects both input availability and task execution; it is not the ideal conditional-resampling kernel. The eight paired comparisons with copying, narrative diagnostics, and upstream-success sensitivity are in Section H.3.

6.4 Adding a fresh key and continuing the chain

Matched re-watermarking. Under the ideal-randomness model, race sampling has the same key-averaged channel as plain sampling. Every realization of that channel induces the same joint law of (K_1, Y') (Section A.6), so a second watermark can change injection but not R_{ret} or the distribution of any inherited-key score. Removal by re-watermarking, reported for images (Bulychev et al., 2026), therefore needs a distorting scheme or a change in the rewrite. Table 2 tests this prediction at the detector level on the same 200 upstream documents per model. Its first four numerical columns give each arm's inherited-key detections and mean score; the last two give the paired race-minus-plain difference and the number of untied pairs. Copying keeps almost all inherited detections, whereas paraphrase and summary leave 7–28 of 200. Within each operation we detect no arm difference: all six mean-difference intervals contain zero, and no exact sign test on the nonzero differences rejects after Holm correction. This does not establish equivalence: copying leaves only 6 untied Qwen pairs and 61 untied Llama pairs, and the Llama copy interval spans $[-12.0, 7.0]$.

Complete trajectories. The fixed-key study froze a trajectory at its first exact reproduction or very short output. We now run three actual operations on every document in the 64-document cohorts, then simulate that rule on the same trajectories. At hop 3, inherited-key detection counts agree between the two policies in all eight model/operation/arm cells. Continuous scores do change: Qwen

Table 2: Inherited evidence and the effect of a second race watermark. Each row uses the same 200 upstream documents; 192 (Qwen) and 200 (Llama) are detected before the operation. Key counts are whole-output detections at 1%. Race − plain is the mean difference in $-\ln p$ with a pointwise 95% document-bootstrap interval including ties; $N_{\neq 0}$ counts untied pairs. All six two-sided sign tests have Holm-adjusted $p = 1$.

		Key (of 200)		Mean $-\ln p$		Race − plain	
Model	Operation	Plain	Race	Plain	Race	Mean [95% CI]	$N_{\neq 0}$
Qwen	Copy	191	191	14.97	14.98	$0.009\ [-0.039, 0.066]$	6
Qwen	Paraphrase	25	28	2.33	2.28	$-0.055\ [-0.275, 0.175]$	200
Qwen	Summary	10	7	1.38	1.38	$-0.002\ [-0.194, 0.195]$	200
Llama	Copy	189	193	109.21	106.52	$-2.689\ [-12.035, 6.975]$	61
Llama	Paraphrase	9	10	1.83	1.69	$-0.145\ [-0.737, 0.290]$	200
Llama	Summary	14	8	1.45	1.56	$0.117\ [-0.157, 0.395]$	200

translation would carry $22/64$ plain and $14/64$ race outputs, with mean score changes of -0.027 and -0.003 under continued generation (Section H.4).

6.5 Finite-key fidelity budgets

The fixed-key copy ladder on six models (Table 3) shows the scale of the finite-key loss budget in (13). Qwen3-4B copies $150/150$ documents exactly, giving $\eta = 0.01977$ and a loss expression of 0.961 nats. This quantity bounds loss from the initial key information; Section B.3 gives the sampling interpretation.

Table 3: Copying after boundary-whitespace trimming and finite-key loss-budget sensitivity. η is the nominal one-sided 95% Clopper–Pearson upper failure rate, recomputed from the displayed counts. The last column assumes a 2^{63}-element master-key alphabet and evaluates $g_{2^{63}}(\eta)$ using unrounded rates; it is a loss budget, not an absolute retention estimate. The sampling scope is described in Section B.3.

Model	Exact copies / N	η	Loss budget (nats)
Llama-3.2-1B	87/132	0.41478	18.791
Llama-3.2-3B	113/149	0.30626	13.990
Qwen3-0.6B	132/148	0.15956	7.407
Qwen3-1.7B	139/150	0.11846	5.537
Qwen3-8B	149/150	0.03123	1.503
Qwen3-4B	150/150	0.01977	0.961

7 Discussion

A rewrite erases, transmits, and adds. Keys enter only through what is added and leave only through what is erased; the inherited key, which cannot enter, is capped by what is transmitted. A task caps the two outer budgets by its fibers and always protects the key information in its shadow; when fidelity means preserving a statistic, it protects exactly the key information in that statistic, and the rest can be erased without changing the text distribution.

7.1 Limitations

The attainment statements describe an ideal optimum for a completely specified task relation, not an implemented semantic rewriter; natural-language fidelity needs such a relation before its fibers can be read.

The language-model evaluation uses short documents, small open-weight models, and a fixed source collection. Its detector and replay scores do not estimate R_{ret} or M_A. Four-field success checks a specified task, not every claim in the narrative; outputs can preserve the fields yet add unsupported details, and many miss the requested narrative length. The copy-ladder failure rates would cover a random-key population only under independent sampling from it, which the fixed-key runs do not provide.

AI USE STATEMENT

The research question, the framing of inherited watermarks as a key-blind operation with separate injection and retention budgets, the choice of experiments and the decisions taken at every stage of this work are the authors'. We used LLM-based assistants throughout as tools: for coding assistance in implementing the experiment and analysis scripts and the figures, for feedback on experimental design and the interpretation of results, for generating the schematic in Figure 5, for drafting and editing text to improve readability, for literature search and formatting references, for suggesting titles and keywords, and for assistance in developing and checking the theoretical arguments, including drafting proof steps that the authors then verified. Language models are also the objects of study in the experiments of Section 6.

We have reviewed all AI-assisted work: every theoretical claim was checked by hand and, where it admits a numerical test, against the exact finite-alphabet computations in Section 6.1; every reported experimental number is reproduced from the stored experiment records by our analysis code; and the literature discussion was checked against the cited sources. We take responsibility for the final content of this work, including text, claims, code and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The operation assumptions and main identities are stated in Sections 3 to 5; proofs and boundary cases appear in Sections A and E. Experimental protocols, detector calibration, and measurement definitions are described in Sections B and H. Additional replay and paired-prefix experiments are detailed in Sections C and D. Throughout, theoretical quantities, detector measurements, and replay statistics are reported separately.

ETHICS STATEMENT

Watermark evidence can inform provenance but should not be treated as proof of authorship or misconduct. False positives, transformation sensitivity, and the distinction between retained information and a detector's score are relevant to deployment. The removal analysis can also inform watermark attacks; we present it to clarify the scope of provenance claims under task constraints.

REFERENCES

Scott Aaronson. Watermarking of large language models, 2023. Talk, Simons Institute workshop on Large Language Models and Transformers.

Mohammad A. Alhejji and Graeme Smith. A tight uniform continuity bound for equivocation. In *IEEE International Symposium on Information Theory*, 2020. URL https://arxiv.org/abs/1909.00787.

Mohd Ruhul Ameen, Akif Islam, Nadim Mahmud, and Md. Ekramul Hamid. Chainwash: Multi-step rewriting attacks on diffusion language model watermarks. *arXiv preprint arXiv:2605.05503*, 2026. URL https://arxiv.org/abs/2605.05503.

Dara Bahri and John Wieting. Improving detection of watermarked language models. *arXiv preprint arXiv:2508.13131*, 2025. URL https://arxiv.org/abs/2508.13131.

Matthieu R. Bloch. Covert communication over noisy channels: A resolvability perspective. *arXiv preprint arXiv:1503.08778*, 2015. URL https://arxiv.org/abs/1503.08778.

Maria Bulychev, Neil G. Marchant, and Benjamin I. P. Rubinstein. Watermarks attack watermarks: Re-watermarking as a generic removal strategy. *arXiv preprint arXiv:2605.16796*, 2026. URL https://arxiv.org/abs/2605.16796.

John Burnet. *Early Greek Philosophy*. A & C Black, London, third edition, 1920.

Zhongze Cai, Shang Liu, Hanzhao Wang, Huaiyang Zhong, and Xiaocheng Li. Towards better statistical understanding of watermarking LLMs. *arXiv preprint arXiv:2403.13027*, 2024. URL https://arxiv.org/abs/2403.13027.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. *arXiv preprint arXiv:2306.09194*, 2023. URL https://arxiv.org/abs/2306.09194.

Max H. M. Costa. Writing on dirty paper. *IEEE Transactions on Information Theory*, 29(3):439–441, 1983.

Peter Gács and János Körner. Common information is far less than mutual information. *Problems of Control and Information Theory*, 2(2):149–162, 1973.

S. I. Gel'fand and M. S. Pinsker. Coding for channel with random parameters. *Problems of Control and Information Theory*, 9(1):19–31, 1980.

Atefeh Gilani, Sajani Vithana, Carol Xuan Long, Oliver Kosut, Lalitha Sankar, and Flavio P. Calmon. ArcMark: Distortion-free multi-byte LLM watermark via optimal transport. *arXiv preprint arXiv:2602.07235*, 2026. URL https://arxiv.org/abs/2602.07235.

Sidong Guo, Tyler Kann, Teodora Baluta, and Matthieu R. Bloch. Covert multi-bit LLM watermarking: An information theory and coding approach. *arXiv preprint arXiv:2605.16709*, 2026. URL https://arxiv.org/abs/2605.16709.

Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. Theoretically grounded framework for LLM watermarking: A distribution-adaptive approach. *arXiv preprint arXiv:2410.02890*, 2024. URL https://arxiv.org/abs/2410.02890.

Haiyun He, Yepeng Liu, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. Distributional information embedding: A framework for multi-bit watermarking. *arXiv preprint arXiv:2501.16558*, 2025. URL https://arxiv.org/abs/2501.16558.

Haiyun He, Yepeng Liu, Zhuoer Shen, Ziqiao Wang, Yongyi Mao, and Yuheng Bu. Fundamental trade-offs in multi-bit watermarking of stochastic processes. *arXiv preprint arXiv:2605.08826*, 2026. URL https://arxiv.org/abs/2605.08826.

Baihe Huang, Hanlin Zhu, Banghua Zhu, Kannan Ramchandran, Michael I. Jordan, and Jason D. Lee. Towards optimal statistical watermarking. *arXiv preprint arXiv:2312.07930*, 2023. URL https://arxiv.org/abs/2312.07930.

Yu-Shin Huang, Peter Just, Hanyun Yin, Krishna Narayanan, Ruihong Huang, and Chao Tian. OD-Stega: LLM-based relatively secure steganography via optimized distributions. *arXiv preprint arXiv:2410.04328*, 2024. URL https://arxiv.org/abs/2410.04328.

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

Jeongyeon Hwang, Sangdon Park, and Jungseul Ok. LLM watermark evasion via bias inversion. *arXiv preprint arXiv:2509.23019*, 2025. URL https://arxiv.org/abs/2509.23019.

Young-Han Kim, Arak Sutivong, and Thomas M. Cover. State amplification. *IEEE Transactions on Information Theory*, 54(5):1850–1859, 2008.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023a. URL https://arxiv.org/abs/2301.10226.

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. *arXiv preprint arXiv:2306.04634*, 2023b. URL https://arxiv.org/abs/2306.04634.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023. URL https://arxiv.org/abs/2307.15593.

Nghia T. Le, Alan Ritter, and Kartik Goyal. Semantic differentiation for tackling challenges in watermarking low-entropy constrained generation outputs. *arXiv preprint arXiv:2601.11629*, 2026. URL https://arxiv.org/abs/2601.11629.

Cheuk Ting Li and Abbas El Gamal. Strong functional representation lemma and applications to coding theorems. *IEEE Transactions on Information Theory*, 64(11):6967–6978, 2018. doi: 10.1109/TIT.2018.2865570.

Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J. Su. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *arXiv preprint arXiv:2404.01245*, 2024. URL https://arxiv.org/abs/2404.01245.

Xiaoyu Li, Zheng Gao, Xiaoyan Feng, Jiaojiao Jiang, Yulei Sui, and Jiankun Hu. Watermark forensics for generative models: An information-theoretic perspective. *arXiv preprint arXiv:2607.13003*, 2026. URL https://arxiv.org/abs/2607.13003.

Yiyang Luo, Ke Lin, Chao Gu, Jiahui Hou, Lijie Wen, and Ping Luo. Lost in overlap: Exploring logit-based watermark collision in LLMs. *arXiv preprint arXiv:2403.10020*, 2024. URL https://arxiv.org/abs/2403.10020.

Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Médard. From the information bottleneck to the privacy funnel. In *IEEE Information Theory Workshop*, 2014. URL https://arxiv.org/abs/1402.1774.

Noam Mazor, Andrew Morgan, and Rafael Pass. Can we watermark low-entropy LLM outputs? *arXiv preprint arXiv:2604.12051*, 2026. URL https://arxiv.org/abs/2604.12051.

Neri Merhav and Shlomo Shamai. Information rates subject to state masking. *IEEE Transactions on Information Theory*, 53(6):2254–2261, 2007.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. arXiv:1609.07843; source of the WikiText-103 corpus.

Pierre Moulin and Joseph A. O'Sullivan. Information-theoretic analysis of information hiding. *IEEE Transactions on Information Theory*, 49(3):563–593, 2003.

Borzoo Rassouli and Deniz Gündüz. Information-theoretic privacy-preserving schemes based on perfect privacy. *arXiv preprint arXiv:2301.11754*, 2023. URL https://arxiv.org/abs/2301.11754.

Dor Tsur, Carol Xuan Long, Claudio Mayrink Verdun, Hsiang Hsu, Haim Permuter, and Flavio P. Calmon. Optimized couplings for watermarking large language models. *arXiv preprint arXiv:2505.08878*, 2025. URL https://arxiv.org/abs/2505.08878.

Shangqing Tu, Yuliang Sun, Yushi Bai, Jifan Yu, Lei Hou, and Juanzi Li. WaterBench: Towards holistic evaluation of watermarks for large language models. *arXiv preprint arXiv:2311.07138*, 2023. URL https://arxiv.org/abs/2311.07138.

Ying Wang and Pierre Moulin. Perfectly secure steganography: Capacity, error exponents, and code constructions. *arXiv preprint cs/0702161*, 2007. URL https://arxiv.org/abs/cs/0702161.

Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. *arXiv preprint arXiv:2311.04378*, 2023. URL https://arxiv.org/abs/2311.04378.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. *arXiv preprint arXiv:2306.17439*, 2023. URL https://arxiv.org/abs/2306.17439.

Appendix roadmap. Section A gives the complete proofs; Section B specifies the protocol and estimands; Section C reports detector replay studies; Section D details the paired-prefix experiment; Section E collects boundary cases and supporting results; and Section F provides the full operation table and further literature context. Section G collects the matched, multi-hop, green-list, and verifier-control measurements. Section H gives the larger paired, factual-report, and fully generated three-hop studies, including quality diagnostics and uncertainty estimates. Section I explains the epigraph.

A PROOFS

Throughout, all statements are conditional on $X = x$ and we suppress it. We use the chain rule for mutual information, $I(A; B, C) = I(A; B) + I(A; C \mid B)$, and the data processing inequality in the form: if $A \to B \to C$ is a Markov chain then $I(A; C) \le \min\{I(A; B), I(B; C)\}$. Figure 5 depicts the interface of Definition 1: the operation sees K_1 only through Y, may use an independent fresh key K_2, and is constrained by task fidelity.

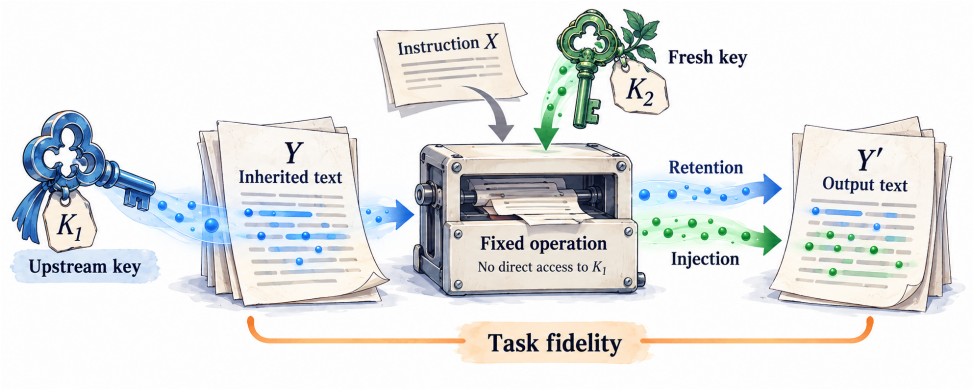

Figure 5: Inherited-key interface. A fixed operation maps Y to Y' under X: it sees K_1 only through Y and may use an independent fresh key K_2. Blue and green paths represent retained and injected information. Task fidelity constrains which inputs can lead to the same output.

A.1 PROOF OF THEOREM 1: THE CAPS AND THE LOSS

Injection. Since $K_2 \perp (K_1, Y)$ we have $I(Y; K_2) = 0$. Expanding $I(Y, Y'; K_2)$ by the chain rule in the two possible orders,

$$I(Y; K_2) + I(Y'; K_2 \mid Y) \;=\; I(Y, Y'; K_2) \;=\; I(Y'; K_2) + I(Y; K_2 \mid Y'),$$

and using $I(Y; K_2) = 0$,

$$I(Y'; K_2 \mid Y) - I(Y'; K_2) \;=\; I(Y; K_2 \mid Y') \;\ge\; 0. \tag{14}$$

Hence $C_{\text{inj}} = I(Y'; K_2) \le I(Y'; K_2 \mid Y) \le H(Y' \mid Y) = F_A$, the last step because conditional mutual information is at most the conditional entropy of either argument.

Retention. By (1), $Y' \perp K_1 \mid Y$, so $K_1 \to Y \to Y'$ is a Markov chain. Data processing gives $I(Y'; K_1) \le I(Y; K_1)$ and $I(Y'; K_1) \le I(Y; Y') = M_A$ simultaneously, which is the stated minimum.

Flux. For any K, expanding $I(K; Y, Y')$ in the two orders gives $I(K; Y) + I(K; Y' \mid Y) = I(K; Y') + I(K; Y \mid Y')$, which is (3). The inflow is at most $H(Y' \mid Y) = F_A$ and the outflow at most $H(Y \mid Y') = E_A$. For $K = K_2$ this is (14); for $K = K_1$ the inflow vanishes by (1), which gives the loss below. Finally $I(Y; K_1) - E_A = H(Y) - H(Y \mid K_1) - H(Y) + M_A = M_A - H(Y \mid K_1)$. Figure 2 draws the two orders and the resulting flows of four keys.

Loss. By (1), $I(K_1; Y' \mid Y) = 0$, so expanding $I(K_1; Y, Y')$ in the two orders gives $I(K_1; Y) = I(K_1; Y') + I(K_1; Y \mid Y')$. Hence

$$I(K_1; Y) - R_{\text{ret}} = I(K_1; Y \mid Y') = H(Y \mid Y') - H(Y \mid Y', K_1) = E_A - H(Y \mid Y', K_1) \le E_A,$$

which is (6) and the lower bound in (4). If $H(Y \mid K_1) = 0$, then $H(Y \mid Y', K_1) = 0$ and the loss equals E_A.

Sums and separation. $E_A + M_A = H(Y)$ and $F_A + M_A = H(Y')$ by the definition of mutual information. A coordinate pair with $C_{\text{inj}} > F_A$ or $R_{\text{ret}} > M_A$ is impossible even when its sum is below $H(Y')$: the budgets are not interchangeable. □

Remark 1 (Which hypothesis does which work). Only (14) uses $K_2 \perp Y$. The retention cap, the exact retention and loss identities, and the nonnegativity of the loss use (1); the lower bound $I(Y; K_1) - E_A \le R_{\text{ret}}$ holds for every joint law, since the loss equals $I(K_1; Y \mid Y') - I(K_1; Y' \mid Y) \le E_A$. Dropping $K_2 \perp Y$ — an operator who chooses its fresh key as a function of the input — can break the injection cap; dropping (1) — an operator given external access to K_1 — can break the retention cap and returns us to the encoder-chosen simplex of (Li et al., 2026, Prop. 3.13). Thus independence controls injection, while key blindness controls retention.

A.2 Proof of Theorem 1: the exact gaps and the corner

Injection. From (14) in Section A.1, $C_{\text{inj}} = I(Y'; K_2) = I(Y'; K_2 \mid Y) - I(Y; K_2 \mid Y')$. Expanding the first term,

$$I(Y'; K_2 \mid Y) = H(Y' \mid Y) - H(Y' \mid Y, K_2) = F_A - H(Y' \mid Y, K_2)\,,$$

which gives (5). Both subtracted terms are nonnegative, and each vanishes exactly under the stated condition: $H(Y' \mid Y, K_2) = 0$ iff Y' is a deterministic function of (Y, K_2), and $I(Y; K_2 \mid Y') = 0$ iff $Y \perp K_2 \mid Y'$.

Retention. By (1), $I(Y'; K_1 \mid Y) = 0$, so expanding $I(Y'; Y, K_1)$ in the two orders,

$$\begin{aligned} M_A &= I(Y'; Y) + I(Y'; K_1 \mid Y) = I(Y'; Y, K_1) \\ &= I(Y'; K_1) + I(Y'; Y \mid K_1) = R_{\text{ret}} + I(Y'; Y \mid K_1). \end{aligned}$$

which is (6). The deficit $I(Y'; Y \mid K_1)$ vanishes iff Y' and Y are conditionally independent given K_1, which holds in particular when $H(Y \mid K_1) = 0$ (then Y is a function of K_1, and conditioning on K_1 fixes Y). When $H(Y \mid K_1) > 0$ the term can be strictly positive. □

Proof of Corollary 1. Under (i), $H(Y \mid K_1) = 0$, so $I(Y'; Y \mid K_1) = 0$ and $R_{\text{ret}} = M_A$ by (6). Under (ii), $H(Y' \mid Y, K_2) = 0$. For (iii): write $Y' = (Y_S, Z)$ with $Z = f_{K_2}(Y_S)$. Then the event $\{Y' = (y_S, z)\}$ equals $\{Y_S = y_S\} \cap \{k_2 : f_{k_2}(y_S) = z\}$, a product event in (Y, K_2) given Y_S, so $Y \perp K_2 \mid Y'$ and $I(Y; K_2 \mid Y') = 0$. Then (5) gives $C_{\text{inj}} = F_A$. □

The counterexample of Corollary 1. Let $Y = (Y_S, Y_R) \in \{0,1\}^2$ with Y determined by a uniform K_1 on four values, let $S = \{1\}$, and let K_2 be uniform over all 16 functions $\phi : \{0,1\}^2 \to \{0,1\}$, with $Z = \phi_{K_2}(Y)$. This operation is deterministic given (Y, K_2) and distortion free, so it satisfies (i) and (ii) but not (iii). Direct calculation gives $H(Y') = 1.386$, $F_A = M_A = 0.693$, $H(Y' \mid K_2) = 1.040$ and $C_{\text{inj}} = 0.347 = F_A/2$, with $I(Y; K_2 \mid Y') = 0.347$ accounting for the entire shortfall. □

A.3 Proof of Theorem 2 and Proposition 2

By (3) with zero inflow for K_1,

$$R_{\text{ret}} = I(K_1; Y) - I(K_1; Y \mid Y') \;\ge\; I(K_1; Y) - H(Y \mid Y') = I(K_1; Y) - E_A. \tag{15}$$

Forward fibers. Conditionally on $Y = y$, the output is supported on $\mathcal{Y}_\epsilon(y)$, so $H(Y' \mid Y = y) \le \log |\mathcal{Y}_\epsilon(y)|$. Averaging gives $F_A \le \mathbb{E} \log |\mathcal{Y}_\epsilon(Y)|$, and Theorem 1 gives the injection cap.

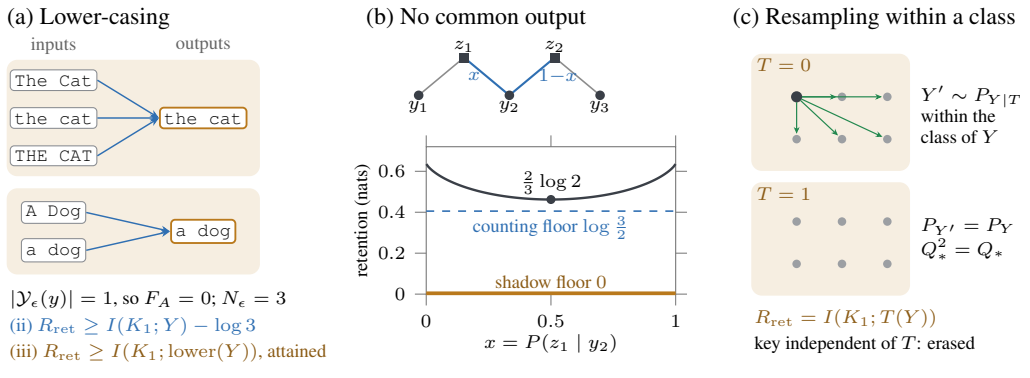

Figure 6: Fibers count, components pass. (a) Lower-casing has singleton forward fibers, so it cannot inject, and its backward fibers are the case variants. The counting floor of Theorem 2(ii) degrades with their number; the components are the lower-cased texts, each is its own common output, and lower-casing attains the shadow floor (iii). (b) The relation of Section A.3 without a common output, with $K_1 = Y$ uniform on $\{y_1, y_2, y_3\}$ and y_2 sent to z_1 with probability x. The graph is connected, so the shadow floor is 0; the counting floor is $\log 3 - \log 2$; the true minimum over faithful channels is $\frac{2}{3}\log 2$ at $x = \frac{1}{2}$, strictly above both. (c) For a task that preserves T, the components are the classes of T. Conditional resampling draws the output from the input's class: it keeps exactly $I(K_1;T(Y))$, leaves the text distribution unchanged, and is idempotent (Section A.4).

Backward fibers. Conditionally on $Y' = y'$, the variable Y is supported on $B_\epsilon(y')$, whose cardinality is at most N_ϵ by (9). Hence $E_A \leq \mathbb{E}\log|B_\epsilon(Y')| \leq \log N_\epsilon$, and (15) gives (10). Only outputs that occur enter the expectation, so each of their backward fibers is nonempty.

The shadow. Let $c(y')$ be the component containing an output y'. If $(Y, Y') = (y, y')$ has positive probability, then $y \in S$ and $y' \in \mathcal{Y}_\epsilon(y)$, so y and y' are joined and $c(y') = \tau_\epsilon(y)$. Thus $\tau_\epsilon(Y) = c(Y')$ almost surely, and data processing gives $I(K_1;Y') \geq I(K_1;c(Y')) = I(K_1;\tau_\epsilon(Y))$. If each component C contains an output $z_C \in \bigcap_{y\in C}\mathcal{Y}_\epsilon(y)$, the channel $y \mapsto z_{\tau_\epsilon(y)}$ is faithful and key-blind. Distinct components contain distinct outputs, so Y' and $\tau_\epsilon(Y)$ determine each other and $I(K_1;Y') = I(K_1;\tau_\epsilon(Y))$.

The common-output condition cannot be dropped. Let $\mathcal{Y}_\epsilon(y_1) = \{z_1\}$, $\mathcal{Y}_\epsilon(y_2) = \{z_1, z_2\}$, and $\mathcal{Y}_\epsilon(y_3) = \{z_2\}$, with $K_1 = Y$ uniform on $\{y_1, y_2, y_3\}$. The graph is connected, so the shadow bound is zero. A faithful channel sends y_2 to z_1 with some probability x, and its retention $h\big((1+x)/3\big) - h(x)/3$ is minimized at $x = 1/2$, where it equals $\frac{2}{3}\log 2$. Figure 6 places this relation between two tasks that have common outputs.

Tasks that preserve a statistic. If $\mathcal{Y}_\epsilon(y) = \{z : T(z) = T(y)\}$, an input and an output are joined exactly when they have the same value of T, so the components are the classes of T met by S, and every class member is admissible for all inputs of its class. The attainment condition holds, and resampling within classes attains it while preserving P_Y (Section A.4). The counting floor never beats the shadow here: $I(K_1;Y) = I(K_1;T) + I(K_1;Y \mid T)$ and $I(K_1;Y \mid T) \leq H(Y \mid T) \leq \log N_\epsilon$.

Failures charged to the key. Let $E = \mathbf{1}\{Y' \notin \mathcal{Y}_\epsilon(Y)\}$, $p = \Pr[E = 1]$, and $n(y') = |B_\epsilon(y')|$. Since E is determined by (Y, Y'),

$$\begin{aligned} I(K_1;Y \mid Y') &= I(K_1;E \mid Y') + I(K_1;Y \mid Y', E) \\ &\leq H(E \mid Y') + (1-p)\mathbb{E}[\log n(Y') \mid E = 0] + pH(K_1 \mid Y', E = 1). \end{aligned} \tag{16}$$

Only outputs occurring on success enter the conditional expectation, so their $n(y')$ is positive. Terms weighted by a zero-probability event are understood as zero. On success the conditional information is also bounded by $\log M$, so it is at most $a = \min\{\log N_\epsilon, \log M\}$; on failure it is at most $\log M$. Bounding $H(E \mid Y') \leq h(p)$ proves (12). The derivative of $h(p) + (1-p)a + p\log M$ is $\log((1-p)/p) + \log M - a \geq 0$ for $p \leq 1/2$. This justifies substitution of an upper bound $\eta \leq 1/2$ and proves Proposition 2.

Imperfect execution and the shadow. Suppose $\Pr[Y' \notin \mathcal{Y}_\epsilon(Y)] = p$ and τ_ϵ takes L values on S. If $L = 1$, then $I(K_1; \tau_\epsilon(Y)) = 0$. If $L \geq 2$, let $\hat{c}(Y')$ be the component of Y' when it contains an input, and a fixed value of τ_ϵ otherwise. Then $\hat{c}(Y') \neq \tau_\epsilon(Y)$ only on failures, so Fano's inequality gives $I(K_1; \tau_\epsilon(Y)) - R_{\mathrm{ret}} \leq I(K_1; \tau_\epsilon(Y) \mid Y') \leq H(\tau_\epsilon(Y) \mid \hat{c}(Y')) \leq h(p) + p \log(L-1)$ for $p \leq 1 - 1/L$, and at most $\log L$ in general.

Generality and a text-space bound. For an arbitrary joint law, the chain rule gives $I(K_1; Y) - I(K_1; Y') = I(K_1; Y \mid Y') - I(K_1; Y' \mid Y) \leq I(K_1; Y \mid Y')$. Thus the same lower bounds do not need the Markov assumption. Alternatively, bounding input rather than key uncertainty on failure gives the valid but often larger loss budget $h(p) + (1-p) \log N_\epsilon + p \log |\mathcal{Y}|$. For nonuniform keys one cannot replace $H(K_1 \mid Y', E = 1)$ by $H(K_1)$: conditioning on a rare failure may select a high-entropy part of the key distribution. The alphabet bound remains valid. □

A.4 Optimal retention under a task partition

The partition case of Theorem 2. Write $U = T(Y)$. Faithfulness means that $T(Y') = U$ almost surely, so the output determines U. Data processing therefore gives $I(K_1; Y') \geq I(K_1; U)$ for every faithful operation. This converse does not require key blindness.

To attain the bound, define the channel

$$Q_*(z \mid y) = P(Y = z \mid U = T(y))$$

on the support of Y. It samples within the input's task class, and hence is faithful. Its dependence on the input is entirely through U, giving $K_1 \perp Y' \mid U$. Since $U = T(Y')$,

$$I(K_1; Y') = I(K_1; U, Y') = I(K_1; U) + I(K_1; Y' \mid U) = I(K_1; U).$$

Furthermore,

$$P(Y' = z) = \sum_u P(U = u) P(Y = z \mid U = u) = P(Y = z).$$

Thus the same channel attains (11) under the additional constraint $P_{Y'} = P_Y$. □

The construction identifies exactly which information can be removed: its loss is $I(K_1; Y) - I(K_1; U) = I(K_1; Y \mid U)$. In particular, complete removal is possible precisely when $I(K_1; U) = 0$. This is the task-partition instance of information removal under a utility constraint (Makhdoumi et al., 2014).

Repeated resampling and intermediate retention. Every intermediate text z drawn by $Q_*(\cdot \mid y)$ has $T(z) = T(y)$. Consequently

$$(Q_*^2)(w \mid y) = \sum_z Q_*(z \mid y) P(Y = w \mid U = T(z)) = Q_*(w \mid y).$$

The channel is therefore idempotent: another application changes neither its input–output law nor its retained information. More generally, let $Q_\lambda = (1 - \lambda)\,\mathrm{Id} + \lambda Q_*$. Each Q_λ is faithful and preserves P_Y, and its retention varies continuously from $I(K_1; Y)$ to $I(K_1; U)$. These are the largest and smallest possible values among key-blind faithful channels, so every value between them is attainable. The retention is nonincreasing along this family: $Q_\lambda Q_\mu = Q_{\lambda+\mu-\lambda\mu}$, and data processing applies.

Approximate task preservation. Suppose instead that the output supplies an estimate $\widehat{U} = s(Y')$ of U, whose alphabet has size $L \geq 2$. For the actual task-error probability $p = \Pr[\widehat{U} \neq U]$, the chain rule and Fano's inequality give

$$\begin{aligned} I(K_1; U) &\leq I(K_1; Y') + I(K_1; U \mid Y') \\ &\leq I(K_1; Y') + H(U \mid \widehat{U}) \\ &\leq I(K_1; Y') + h(p) + p \log(L-1). \end{aligned}$$

Thus the task floor loses at most $h(p) + p \log(L-1)$ nats. If only $p \leq \eta \leq 1 - 1/L$ is known, replacing p by η is valid because this expression is nondecreasing on that interval. For larger upper bounds η, use the penalty $\log L$. When $L = 1$, the task statistic is constant and $I(K_1; U) = 0$. These statements require no estimate of the number of texts in a task class.

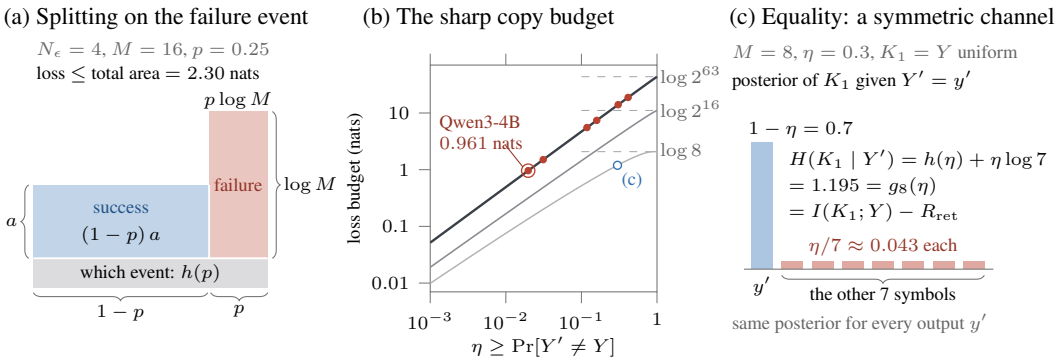

Figure 7: Failures are charged to the key alphabet, not to the text. (a) The proof of Proposition 2 splits on the failure event, as in (16). The width of a block is a probability and its height is a cost in nats, so the total area bounds the loss $I(K_1; Y) - R_{\mathrm{ret}}$. With $N_\epsilon = 4$, $M = 16$ and $p = 0.25$, a success costs at most $a = \log 4$, a failure at most $\log 16$, and learning which event occurred at most $h(0.25)$; the total is 2.30 nats. (b) The sharp copy budget $g_M(\eta)$ of (13) for $M = 2^{63}$ (dark), 2^{16} and 8; dashed lines mark the saturation value $\log M$. The dots are the six models of Table 3, placed at their η on the 2^{63} curve. Qwen3-4B copies $150/150$ documents exactly, and its budget is 0.961 nats against $\log 2^{63} = 43.67$. These are loss budgets, not absolute retention estimates. (c) Uniform $K_1 = Y$ through the 8-ary symmetric channel with $\eta = 0.3$ (open circle in (b)). Every output leaves the same posterior, so the loss is its entropy, $h(0.3) + 0.3 \log 7 = 1.195$ nats, and the bound holds with equality (Section A.5).

A.5 Sharp copying loss for a finite key

Proof of (13). Place Y and Y' in the same finite text alphabet, taking the union of their alphabets if necessary. The coupling (K_1, Y, Y') implies

$$\mathrm{TV}(P_{K_1,Y}, P_{K_1,Y'}) \leq \Pr[Y \neq Y'] \leq \eta.$$

For a key alphabet of size $M \geq 2$ and $\eta \leq 1 - 1/M$, the conditional-entropy continuity theorem of Alhejji & Smith (2020, Eq. (7)) gives

$$|H(K_1 \mid Y') - H(K_1 \mid Y)| \leq h(\eta) + \eta \log(M - 1).$$

The key marginal is identical in the two joint laws, so the left-hand side is $|I(K_1; Y) - I(K_1; Y')|$. For larger η, both conditional entropies lie between 0 and $\log M$, giving the stated continuation $g_M(\eta) = \log M$. A singleton key ($M = 1$) has zero mutual information with every observation, so its loss bound is identically zero. The continuity argument itself does not require a Markov chain; key blindness makes the loss nonnegative. □

Sharpness within the operation model. Let $K_1 = Y$ be uniform on $\{1, \ldots, M\}$, and pass it through the symmetric channel

$$P(Y' = z \mid Y = y) = \begin{cases} 1 - \eta, & z = y, \\ \eta/(M - 1), & z \neq y. \end{cases}$$

This channel is key-blind, preserves the uniform text marginal, and has copying error η. The posterior of K_1 given any output has probabilities $1 - \eta, \eta/(M - 1), \ldots, \eta/(M - 1)$, so

$$I(K_1; Y) - I(K_1; Y') = H(K_1 \mid Y') = h(\eta) + \eta \log(M - 1).$$

Hence equality holds for every $0 \leq \eta \leq 1 - 1/M$. At the upper endpoint the output is independent of the key. That same channel is admissible for any larger upper error budget, attaining the continuation $\log M$ as well. Figure 7 draws the failure split of Proposition 2, the copy budget for the models of Table 3, and this channel.

A.6 Attainable injection at a fixed channel

Fix the upstream joint law and the key-averaged channel A. Let p be the law of Y on its support $\mathcal{S}$, let $\mathcal{O}$ be the output alphabet, and write $q = pA$. We first allow any finite fresh-key alphabet and law, independent of the upstream pair. This optimization is the full-data perfect-privacy problem of Rassouli & Gündüz (2023), with the input, output, and independent auxiliary variable identified with Y, Y', K_2, respectively.

A linear program over deterministic maps. Let $\mathcal{F}$ contain all maps $f : \mathcal{S} \to \mathcal{O}$, and let $p_f(z) = \sum_{y:f(y)=z} p(y)$ be the output law under f. Define the feasible map mixtures by

$$\Lambda(A) = \left\{ \lambda_f \geq 0 : \sum_{f \in \mathcal{F}} \lambda_f \mathbf{1}\{f(y) = z\} = A(z \mid y) \text{ for all } y, z \right\}.$$

These constraints imply $\sum_f \lambda_f = 1$. The largest attainable injection is

$$J_{\text{free}}(p, A) = H(q) - \min_{\lambda \in \Lambda(A)} \sum_{f \in \mathcal{F}} \lambda_f H(p_f). \tag{17}$$

The feasible set is nonempty: independently drawing each $f(y)$ from $A(\cdot \mid y)$ gives a feasible random map. It is also compact, so the minimum exists.

For completeness, consider any stochastic keyed realization G_k. Conditional on $K_2 = k$, sample a map F by drawing $F(y)$ independently from $G_k(\cdot \mid y)$ for every y. The pair (K_2, F) is independent of the upstream pair, and setting $Y' = F(Y)$ preserves the original joint law of (Y, K_2, Y'). Conditioning on the enlarged key yields

$$I(K_2; Y') \leq I(K_2, F; Y') = H(q) - \mathbb{E}H(p_F) \leq J_{\text{free}}(p, A).$$

Conversely, draw $K_2 = f$ with an optimizing probability λ_f and output $f(Y)$. This realizes A and attains (17). The enlargement of the key is permitted because its alphabet and law are optimized; it need not preserve a prescribed key resource.

For deterministic realizations, (5) also writes the optimum as $J_{\text{free}} = F_A - \inf I(Y; K_2 \mid Y')$. The infimum is the excess functional information studied by Li & El Gamal (2018). This identifies the possible gap below the entropy cap with an established functional-representation quantity.

The achievable interval. Every realization of A induces the same joint law $P(K_1, Y') = \sum_y P(K_1, y)A(Y' \mid y)$, so retention is fixed. Injection ranges over the entire interval $[0, J_{\text{free}}(p, A)]$. The lower endpoint uses $G_k = A$ for every key. To obtain an intermediate point, use an independent Bernoulli gate with probability t to choose between this realization and an optimal map realization, and include the gate in the fresh key. Both branches have output marginal q, so the gate carries no output information. The injection is exactly $tJ_{\text{free}}(p, A)$.

The binary symmetric channel with an optimized key. Now let $Y = K_1$ be a fair bit and A the binary symmetric channel with crossover $\delta \in [0, 1/2]$. The four deterministic maps are identity, complement, constant zero, and constant one. If their probabilities are a, b, c, d, the channel constraints give

$$c = d = t, \qquad b = \delta - t, \qquad a = 1 - \delta - t, \qquad 0 \leq t \leq \delta.$$

Identity and complement have output entropy $\log 2$, while the constants have zero entropy. Consequently injection is $2t \log 2$, maximized at $t = \delta$. This proves the first equality in (7), also a special case of Rassouli & Gündüz (2023, Theorem 3). Its maximizing map probabilities are $(1-2\delta, 0, \delta, \delta)$. The channel's other coordinates are $F_A = h(\delta)$ and $R_{\text{ret}} = M_A = \log 2 - h(\delta)$.

The same channel with one prescribed fair key bit. Let K_2 be uniform on $\{0, 1\}$, and define $a_k = P(Y' = 1 \mid Y = 0, K_2 = k)$ and $b_k = P(Y' = 1 \mid Y = 1, K_2 = k)$. Averaging to A requires

$$a_0 + a_1 = 2\delta, \qquad b_0 + b_1 = 2(1 - \delta).$$

Hence $a_0 \in [0, 2\delta]$, $b_0 \in [1 - 2\delta, 1]$, and $q_0 = P(Y' = 1 \mid K_2 = 0) = (a_0 + b_0)/2$ lies in $[1/2 - \delta, 1/2 + \delta]$. The other conditional probability is $q_1 = 1 - q_0$. Thus $I(K_2; Y') = \log 2 - h(q_0)$ is maximized at either endpoint, proving the second equality in (7). The following two channels attain it; rows index Y and columns index Y':

$$G_0 = \begin{pmatrix} 1 & 0 \\ 2\delta & 1 - 2\delta \end{pmatrix}, \qquad G_1 = \begin{pmatrix} 1 - 2\delta & 2\delta \\ 0 & 1 \end{pmatrix}.$$

They average to A and have $H(Y' \mid Y, K_2) = h(2\delta)/2$. Their auxiliary randomness remains outside the prescribed key bit. The resulting injection satisfies $J_{\text{bit}}(\delta) = 2\delta^2 + O(\delta^4)$, whereas the optimized-key value is $2\delta \log 2$.

A third realization makes the distinction especially direct: set $Y' = Y \oplus K_2$ with an independent $K_2 \sim \text{Bernoulli}(\delta)$. It has the same averaged channel and retention, but zero injection because the fair input makes Y' independent of K_2. Its entire fresh-entropy budget appears in the deficit $I(Y; K_2 \mid Y') = h(\delta)$. All three realizations obey the same upper budget; their use of that budget depends on how the fresh key enters the operation.

A.7 Exact injection from a uniform finite key

Proof of Proposition 1. The preceding subsection gives the optimized-key value $J_{\text{free}}(\delta) = 2\delta \log 2$, a specialization of Rassouli & Gündüz (2023, Theorem 3). We now keep K_2 uniform on $m = 2^b$ values, with integer $b \geq 1$, while allowing stochastic keyed channels.

Feasible output biases. For each key k, write

$$a_k = P(Y' = 1 \mid Y = 0, K_2 = k), \qquad b_k = P(Y' = 1 \mid Y = 1, K_2 = k),$$
$$c_k = \frac{a_k + b_k}{2} - \frac{1}{2}.$$

The required averaged channel imposes $\sum_k a_k = m\delta$ and $\sum_k b_k = m(1-\delta)$, hence $\sum_k c_k = 0$. For a given bias $c_k \in [-1/2, 1/2]$, valid row probabilities satisfy

$$\max\{0, 2c_k\} \leq a_k \leq \min\{1, 1 + 2c_k\}, \qquad b_k = 1 + 2c_k - a_k.$$

Put $L = \sum_k |c_k|$. Since the biases sum to zero, the lower endpoints of these intervals sum to L, and their upper endpoints sum to $m - L$. Their sums fill the entire interval $[L, m - L]$. Thus a bias vector is feasible exactly when

$$\sum_k c_k = 0, \qquad |c_k| \leq \frac{1}{2}, \qquad \sum_k |c_k| \leq m\delta.$$

Here $\delta \leq 1/2$ ensures that the last inequality also gives $m\delta \leq m - L$.

The averaged output is a fair bit. Defining $\phi(c) = \log 2 - h(1/2 + c)$, we therefore have

$$I(K_2; Y') = \frac{1}{m} \sum_k \phi(c_k).$$

The function ϕ is even, convex, and nondecreasing on $[0, 1/2]$, with $\phi(0) = 0$.

Concentrating the bias gives the converse. The positive and negative biases each have total magnitude $t = L/2 \leq m\delta/2$. For nonnegative entries bounded by $1/2$ with a given sum t, convexity maximizes the sum of ϕ by filling entries to $1/2$ and leaving at most one nonzero remainder. Indeed, if $0 < u \leq v < 1/2$, transferring $s = \min\{u, 1/2 - v\}$ from u to v cannot decrease $\phi(u) + \phi(v)$. Repeating this transfer gives the upper bound

$$g(t) = \lfloor 2t \rfloor \log 2 + \phi\left(t - \frac{\lfloor 2t \rfloor}{2}\right)$$

for each sign. The function g is continuous and nondecreasing: it increases within each interval between consecutive half-integers, and its endpoint values agree. Writing $j = \lfloor m\delta \rfloor$ and $r = m\delta - j$, we obtain

$$I(K_2; Y') \leq \frac{2g(t)}{m} \leq \frac{2g(m\delta/2)}{m} = \frac{2}{m}\left[j \log 2 + \log 2 - h\left(\frac{1+r}{2}\right)\right].$$

This proves the upper bound in (8). Figure 8 shows the transfer for $m = 8$ and $\delta = 0.2$.

A matching keyed channel. Assign j keys to the constant-zero map and j keys to the constant-one map. If $r > 0$, assign two more keys to

$$G_- = \begin{pmatrix} 1 & 0 \\ r & 1 - r \end{pmatrix}, \qquad G_+ = \begin{pmatrix} 1 - r & r \\ 0 & 1 \end{pmatrix},$$

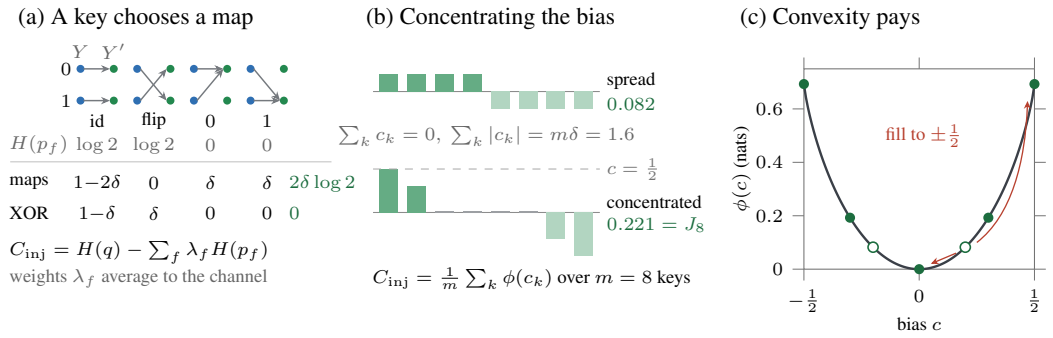

Figure 8: A uniform key must spend its bias in whole keys. (a) With a fair input, each deterministic map of a bit has output entropy $\log 2$ or 0, and injection is $\log 2$ times the weight on the two constants (Section A.6). The optimized key puts weight δ on each constant; the XOR key uses only the identity and the flip, and injects nothing. (b) For $m = 8$ uniform keys and $\delta = 0.2$, each key k contributes $\phi(c_k)$, where c_k is its output bias and $\phi(c) = \log 2 - h(\frac{1}{2} + c)$. The channel constrains only $\sum_k c_k = 0$, $|c_k| \le \frac{1}{2}$, and $\sum_k |c_k| \le m\delta$. Spreading the bias over all keys injects 0.082 nats; filling keys to $\pm\frac{1}{2}$ and leaving one remainder $\pm r/2$ on each side injects $J_8(0.2) = 0.221$ (Proposition 1). (c) Since ϕ is convex, moving bias from a smaller entry to a larger one cannot decrease the sum, so the concentrated vector is optimal (open circles: spread; dots: concentrated). Here $r = m\delta - \lfloor m\delta \rfloor = \frac{3}{5}$, where the gap to the optimized value $2\delta \log 2$ attains its sharp value $2\log(5/4)/m$.

and let every remaining key use the identity map. Rows index Y and columns index Y'. Because m is even and $r > 0$ implies $\delta < 1/2$, we have $2j + 2 \le m$, so these assignments fit the prescribed key alphabet. Their row sums satisfy

$$\frac{1}{m}\sum_k a_k = \frac{j+r}{m} = \delta, \qquad \frac{1}{m}\sum_k b_k = 1 - \delta.$$

The constant maps have zero conditional output entropy. Each of G_- and G_+ has conditional output entropy $h((1+r)/2)$, and each identity map has entropy $\log 2$. Substituting these values attains (8). When $r = 0$, omit the two fractional channels and use $m - 2j$ identity keys; this also covers $\delta = 0$ and $\delta = 1/2$. □

The exact cost of the prescribed key law. Subtracting (8) from the optimized-key value gives

$$J_{\text{free}}(\delta) - J_m(\delta) = \frac{2}{m}\left[h\left(\frac{1+r}{2}\right) - (1-r)\log 2\right] \le \frac{2\log(5/4)}{m}.$$

The bracket is nonnegative by concavity of h. Its derivative is $\frac{1}{2}\log((1-r)/(1+r)) + \log 2$, so its unique maximum occurs at $r = 3/5$ and equals $\log(5/4)$. Thus the uniform gap bound is attained, and $J_m = J_{\text{free}}$ holds exactly when $m\delta$ is an integer. For fixed m and $\delta \downarrow 0$, the entropy expansion gives

$$J_m(\delta) = m\delta^2 + O(m^3\delta^4).$$

The optimized key law has linear small-noise injection, whereas each prescribed finite uniform key has quadratic injection near zero; the gap nevertheless vanishes uniformly at rate $1/m$ as the key alphabet grows.

A.8 Proof of the substitution bound

Fix an output y'. Any y in its pre-image is determined by the set of positions at which it differs from y' — at most m of them — together with the input token at each such position, and by hypothesis at most V_ϵ input tokens admit a given output token at a given position. Hence

$$N_\epsilon \le \sum_{j \le m} \binom{n}{j} V_\epsilon{}^j \le V_\epsilon{}^m \sum_{j \le m} \binom{n}{j} \le V_\epsilon{}^m \left(\tfrac{en}{m}\right)^m = \left(\tfrac{enV_\epsilon}{m}\right)^m,$$

using $V_\epsilon \ge 1$ in the second step and the standard $\sum_{j \le m} \binom{n}{j} \le (en/m)^m$, valid for every $1 \le m \le n$, in the third. Substituting into (10) gives (18). □

A.9 Why local replay is not an exact null decomposition

The local replay statistic of Section B.2 does not induce a null decomposition over all nonmatching pairs. To see why, consider an equal-probability two-token exponential race at one context seed. Let its independent uniform coordinates be U_a, U_b, and let generation select the larger coordinate. Conditional on selecting a, the coordinate for the nonselected token b has density $2(1-u)$ on $[0,1]$, rather than the uniform density. Consequently,

$$\mathbb{E}[-\log(1-U_b) \mid U_a > U_b] = \int_0^1 2(1-u)[-\log(1-u)]\,du = \tfrac{1}{2},$$

whereas the null mean is 1. If an operation emits b at this already queried seed, its pair may be absent from the original document, but its score is not null distributed: generation used the entire vector when choosing its winner. Thus a nonmatching (seed, token) pair need not be a previously unqueried random-oracle coordinate. Replay counts remain useful measurements of local reuse, but do not become information or detector-evidence bounds through this decomposition.

A.10 Proof of the pipeline bounds

By the assumption that every hop satisfies Definition 1 with fresh total randomness, each hop satisfies (1) relative to its own input and carries no hidden inter-hop state, so $K_1 \to Y^{(0)} \to \cdots \to Y^{(k)}$ is a Markov chain.

Upper bound. Data processing along the chain gives $I(Y^{(k)}; K_1) \le I(Y^{(i)}; K_1)$ for every $i \le k$, and Theorem 1 applied at hop i gives $I(Y^{(i)}; K_1) \le I(Y^{(i-1)}; Y^{(i)}) = M_{A_i}$. Taking the minimum over i gives the first claim.

Lower bound. By the same argument as (15) applied to the composite chain, $I(Y^{(k)}; K_1) \ge I(Y^{(0)}; K_1) - H\big(Y^{(0)} \mid Y^{(k)}\big)$. Expanding the conditional entropy along the chain and using that $Y^{(i)}$ is conditionally independent of $Y^{(i+2)}, \ldots, Y^{(k)}$ given $Y^{(i+1)}$ (the reverse Markov property of the chain),

$$H\Big(Y^{(0)} \mid Y^{(k)}\Big) \;\le\; \sum_{i=0}^{k-1} H\Big(Y^{(i)} \mid Y^{(i+1)}, \ldots, Y^{(k)}\Big) \;=\; \sum_{i=0}^{k-1} H\Big(Y^{(i)} \mid Y^{(i+1)}\Big) \;\le\; \sum_{i=1}^{k} \log N_{\epsilon_i},$$

the last step by the faithfulness of hop j, for $j = 1, \ldots, k$, as in Theorem 2. □

Figure 9 draws the upper bound and the additive loss for a chain of binary symmetric channels, together with the key reuse of Example 1.

A.11 The pre-image, not the forward ball

Let $\mathcal{Y} = \{y_0, y_1, y_2, y_3\}$, $\mathcal{Y}' = \{y'_0, y'_1\}$, and let the fidelity relation be $\mathcal{Y}_\epsilon(y_i) = \{y'_0\}$ for $i \le 2$ and $\mathcal{Y}_\epsilon(y_3) = \{y'_0, y'_1\}$. Every forward ball has size at most 2, but the pre-image of y'_0 is all of $\mathcal{Y}$, so $N_\epsilon = 4$. Take K_1 uniform on $\{0,1,2,3\}$ with $Y = y_{K_1}$, so $I(Y; K_1) = \log 4$, and let the operation map every input to y'_0 — which is faithful. Then Y' is constant and $R_{\text{ret}} = 0$. The pre-image bound (10) reads $0 \ge \log 4 - \log 4 = 0$ and holds with equality; the forward-ball version would read $0 \ge \log 4 - \log 2 = \log 2 > 0$ and is false. □

B Experimental Protocol and Measurement Definitions

B.1 Fixed-key operation-study protocol

This section describes the fixed-key operation grid and copy ladder. The document-specific-key controlled studies are specified separately in Section H.

The primary grid uses Qwen3-1.7B, Qwen3-4B, and Llama-3.2-3B-Instruct, with 20 WikiText-103 topic seeds per model and documents of up to 192 generated tokens. Phi-4-mini provides a supplementary family check; the copy-fidelity ladder adds Qwen3-0.6B, Qwen3-8B, and Llama-3.2-1B. Exponential-race sampling couples each upstream document with its key.

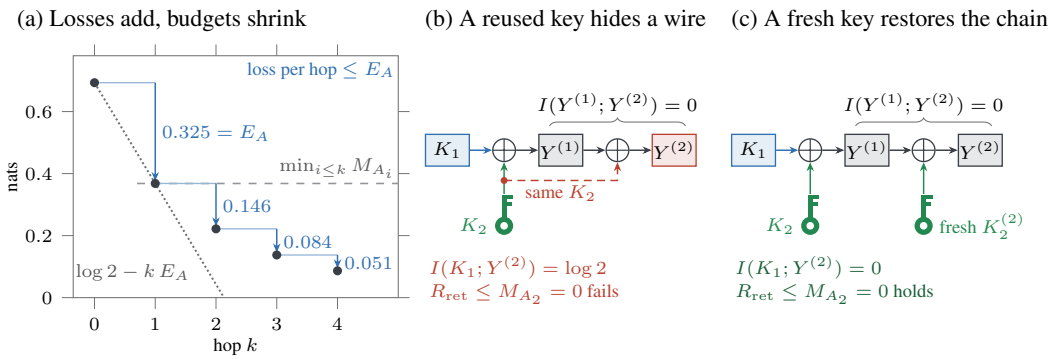

Figure 9: Along a pipeline, losses add and transmitted budgets only shrink. (a) Repeated binary symmetric channels with $\delta = 0.1$ on $Y^{(0)} = K_1$, a uniform bit, for $k = 0, \ldots, 4$ hops. Dots mark $R_{\text{ret}}^{(k)} = I(K_1; Y^{(k)})$ and blue arrows the loss of each hop. Every hop has transmitted budget $M_{A_i} = \log 2 - h(0.1) = 0.368$ and erased budget $E_{A_i} = h(0.1) = 0.325$ nats. Retention stays below $\min_{i \leq k} M_{A_i}$ (Corollary 4), which can only fall with k. It stays above $\log 2 - \sum_{i \leq k} E_{A_i}$, because the losses add hop by hop (Section A.10). The first hop attains both bounds; later hops lose less than E_A, so the additive floor is loose and is vacuous from $k = 3$. (b) Example 1 uses one key twice: $Y^{(1)} = K_1 \oplus K_2$ and $Y^{(2)} = Y^{(1)} \oplus K_2$. Then $I(Y^{(1)}; Y^{(2)}) = 0$, yet $Y^{(2)} = K_1$. The second use of K_2 is a wire around $Y^{(1)}$. The retention bound of Theorem 1 fails, because $I(K_1; K_2 \mid Y^{(1)}) = \log 2$ and the chain is not Markov. (c) A fresh key $K_2^{(2)}$ at hop 2 makes the chain Markov, and $I(K_1; Y^{(2)}) = 0 = M_{A_2}$.

The seven operation prompts request exact copying, global word replacement (local edit), constrained rewriting, English-to-French translation, paraphrase, summary, and input-independent (free) generation. The last uses a paragraph disjoint from the input document. The race and plain arms form three-hop pipelines; the green-list arm uses $\gamma = 0.25$ and a logit boost of 2. The upstream key is fixed, and generated watermarking hops use distinct fixed downstream keys. Generation stops when an output exactly matches its input token sequence or has fewer than 40 characters after whitespace trimming. Later summaries carry the last output forward. An exact reproduction under one key need not remain unchanged under a fresh key; the multi-hop curves describe this stopping policy.

B.2 Measurement scope

Detector scores and calibration. The detector records $-\ln p$ and scores each distinct (context-seed, token) pair once. For a continuous calibrated null, the mean of $-\ln p$ is 1. A period-5 control with 200 tokens and 400 wrong keys illustrates the need for de-duplication: without it, the score has standard deviation 86.6 rather than the nominal 14.1, and rejection rate 0.21 at 10^{-6}. The pooled calibration check has 6114 replicates, standardized mean 0.019, standard deviation 1.009, and rejection rate 0.056 at 0.05. This check characterizes the 0.05 threshold; resolving a false-positive probability of 10^{-6} would require a larger pool.

Binary-KL diagnostics. At the population level, applying a fixed detector to the joint law of output and key and to the product of its marginals gives $I(Y'; K_1) \geq d(\text{TPR} \| \text{FPR})$ by data processing. Our calculation combines a lower Clopper–Pearson estimate of TPR with an upper estimate of FPR, using a nominal-level floor for FPR. The largest value is 3.1 nats and the sample ceiling is 3.2, with up to 2440 wrong-key replicates. We use these values in Table 7 as binary-KL diagnostics for the evaluated sample. A finite-sample certificate would add a common target population, independent Bernoulli sampling, and simultaneous coverage; the fixed generation keys, shared wrong-key replicates, and document selection here define a different sampling design. The prefix calculation in Table 5 uses nominal $\alpha = 10^{-6}$ for FPR. For a scheme $Y \sim Q_K$ with averaged law $\bar{Q} = \mathbb{E}_K Q_K$, the standard identity $\mathbb{E}_K D(Q_K \| R) = I(Y; K) + D(\bar{Q} \| R)$ separates dependence on the secret from a public shift relative to any reference law R (see Li et al., 2026, Fact 2.11).

Conditional and marginal injection. For an ideal distortion-free deterministic keyed sampler, path surprisal and the sum of conditional next-token entropies have the same expectation. Conditional on an input Y, these quantities concern $I(Y'; K_2 \mid Y)$ and $H(Y' \mid Y)$. Marginal injection

is instead $I(Y'; K_2)$, and their difference is $I(Y; K_2 \mid Y')$ in (5). The pathwise measurements therefore diagnose how much fresh randomness the sampler uses, while (5) identifies the additional condition for attaining the marginal injection cap. For a stochastic green list, the key-averaged next-token law is $\bar{q}_t$; its entropy is estimated by jackknife-debiased Monte Carlo over keys. Using p_t would instead describe the unmodified sampler.

Positional replay. Let $S^\star$ contain output positions whose (context-seed, token) pair also appears in the upstream document. We report its fraction and an entropy-weighted replay statistic, $\widehat{M}_A$, as two measurements of local reuse. They need not equal $I(Y; Y')$: exponential-race generation compares all token coordinates at a seed, so a different nonwinning token at that seed can have a non-null detector contribution even when its pair is absent from the document. Section A.9 gives an explicit example. Conversely, an invertible transformation can preserve mutual information while changing the local matches. This is why the tables keep replay and detector columns separate.

Finite keys and ideal randomness. The mathematical identities hold for the specified joint law with any key distribution. The fixed-key operation study uses fixed master keys; its context-specific seeds do not specify a random-key law. For sensitivity analysis, we separately assume a master-key alphabet of size 2^{63}, giving $I(Y; K_1) \le H(K_1) \le 63 \log 2$, about 43.7 nats. Path-entropy budgets use the ideal fresh-randomness model of the sampler. Repeated context queries add a related distinction: determinism gives $H(Y \mid K_1, X) = 0$, whereas equality between the key-averaged sequence law and the ordinary model law is an ideal-randomness condition. We use that idealization for the empirical entropy calculations.

B.3 Finite-key loss-budget calculation

Table 3 reports exact-copy success counts after trimming leading and trailing whitespace, and a sensitivity calculation using (13). The variables in this calculation are the trimmed strings, and the information-loss claim concerns those strings. This success criterion differs from the token-sequence equality used by the multi-hop stopping rule. For f failures among N documents, the nominal one-sided 95% Clopper–Pearson upper rate is the 0.95 quantile of $\mathrm{Beta}(f+1, N-f)$ when $f < N$. Under the assumed 2^{63}-element key alphabet, we evaluate $g_{2^{63}}$ at the unrounded rate. For example, zero failures among 150 documents give $1 - 0.05^{1/150} = 0.0197734$ and a loss expression of 0.960628 nats. No path-entropy estimate or input-support proxy enters this calculation.

The rates condition on the recorded documents and fixed generation keys. Population coverage for these intervals requires independent Bernoulli trials from the target random-key law. Thus the table is a sensitivity calculation: given a valid population failure bound η and the assumed key alphabet, $R_{\mathrm{ret}} \ge I(K_1; Y) - g_{2^{63}}(\eta)$.

Global replacement is not assigned $N_\epsilon = 1$: replacing a source word by an already possible target word can merge inputs, and capitalization introduces an additional ambiguity. The operation therefore lacks the injective pre-image structure used by the exact-copy calculation.

B.4 Fresh-entropy spectrum

The per-token spectrum in Table 4 shows how strongly each operation prompt constrains local next-token distributions. Related task-dependent effects have been studied by Bahri & Wieting (2025); Le et al. (2026); He et al. (2024); Tu et al. (2023).

C Detection Length, Context Width, and Removal

Using the emitted token identifiers and initial contexts, we replay the detector at different prefix lengths, context widths, and substitution budgets. These controlled views separate the effect of the verifier from the effect of changing the text.

Table 4: Per-token fresh entropy F_A/n across the task spectrum, measured directly from logits under each operation's own prompt (base sampling, no watermark). This is the axis along which Table 7 orders its operations; it spans three orders of magnitude between free generation and exact copying.

task	qwen3-1.7b
free generation	0.6635
open-ended QA	0.4440
translate	0.3179
summarize	0.1807
paraphrase	0.1504
constrained rewrite	0.0102
local edit	0.0069
copy	0.0007

Table 5: Observed-prefix detection at nominal $\alpha = 10^{-6}$. Each cell gives the detection rate and, in parentheses, the binary-KL diagnostic formed from a lower TPR estimate and nominal α; Section B.2 specifies its calibration scope.

Prefix	Llama-3.2-3B	Qwen3-1.7B	Qwen3-4B
8	0.01 (0.0)	0.00 (0.0)	0.00 (0.0)
16	0.05 (0.3)	0.01 (0.0)	0.00 (0.0)
24	0.17 (1.3)	0.02 (0.0)	0.00 (0.0)
32	0.29 (2.6)	0.03 (0.1)	0.00 (0.0)
48	0.51 (5.3)	0.07 (0.3)	0.02 (0.0)
64	0.70 (8.1)	0.16 (1.2)	0.03 (0.1)
96	0.91 (11.6)	0.35 (3.3)	0.07 (0.3)
128	0.97 (12.7)	0.60 (6.5)	0.23 (1.9)
160	1.00 (13.4)	0.74 (8.2)	0.41 (4.0)
192	1.00 (12.9)	0.89 (8.9)	0.61 (6.2)

C.1 Detection against observed length

Table 5 traces detection as progressively longer prefixes become visible. A pair is de-duplicated only when the same pair occurred earlier in the observed prefix. Llama-3.2-3B reaches full detection by 160 tokens in these samples; Qwen3-4B reaches 61% at the full 192-token prefix. The binary-KL column combines a lower TPR estimate with nominal α and is interpreted under the calibration scope of Section B.2. The curve is useful operationally: it shows how much text this verifier needs before its decision stabilizes.

C.2 Sensitivity to context width

The context-width measurements appear in Section G.4 and Table 9. Here the output text is held fixed while the verifier width varies. The resulting power loss therefore identifies context alignment as part of the detection mechanism, rather than an effect of a stronger transformation.

C.3 Controlled random substitutions

The substitution experiment replaces selected positions with random vocabulary tokens and reports the mean number changed, detector score, rejection rate, and the edit-budget expression $m \log(enV_\epsilon/m)$ with $V_\epsilon = |\mathcal{V}|$. The edit expression is evaluated at the observed mean length and mean edit count, making it a descriptive companion to the support-wise statement in Corollary 3. Together the two columns compare how quickly this detector loses evidence with how the theorem prices the same edit scale.

On Llama-3.2-3B, changing 17.5 positions has edit price 267 nats and reduces the detector score by 51%. The rejection rate moves from 0.946 to 0.810. On Qwen3-4B, the corresponding comparison uses 4.7 edits, price 78, and rejection rates 0.507 and 0.340. These trajectories expose the model-dependent operating range of the detector under editing.

Table 6: Random-substitution response on hop-1 copy outputs. Here m is the mean number of changed positions, priced is $m \log(enV_\epsilon/m)$ with $V_\epsilon = |\mathcal{V}|$, and evidence is the mean detector score $-\ln p$.

Model	edit rate	m	priced (nats)	evidence (nats)	detect@10^{-6}
Llama-3.2-3B — $H_{\text{path}} = 352$ nats					
	0.000	0.0	0	113.3	0.946
	0.010	1.9	33	104.8	0.946
	0.025	4.5	75	94.9	0.946
	0.050	8.9	142	78.5	0.931
	0.100	17.5	267	55.7	0.810
	0.150	26.4	392	35.8	0.606
	0.200	35.0	509	22.4	0.427
	0.300	52.9	748	8.6	0.134
	0.500	87.9	1198	1.8	0.004
Qwen3-1.7B — $H_{\text{path}} = 97$ nats					
	0.000	0.0	0	18.5	0.693
	0.010	1.5	27	17.1	0.647
	0.025	3.9	64	15.4	0.569
	0.050	7.7	123	12.7	0.364
	0.100	15.5	235	8.8	0.120
	0.150	23.2	344	6.1	0.042
	0.200	30.8	448	4.3	0.007
	0.300	46.4	655	2.2	0.000
	0.500	77.2	1051	1.3	0.000
Qwen3-4B — $H_{\text{path}} = 86$ nats					
	0.000	0.0	0	14.5	0.507
	0.010	2.0	35	13.5	0.447
	0.025	4.7	78	12.1	0.340
	0.050	9.4	149	9.9	0.189
	0.100	18.4	280	7.1	0.067
	0.150	27.8	412	4.6	0.007
	0.200	36.7	534	3.6	0.002
	0.300	55.6	785	2.0	0.000
	0.500	92.4	1258	1.2	0.000

C.4 Bootstrap variation in path entropy

Resampling documents with replacement (2000 replicates) gives 95% percentile intervals for mean path entropy: $[83.2, 88.0]$ on Qwen3-4B, $[93.5, 100.6]$ on Qwen3-1.7B, and $[306.5, 397.0]$ on Llama-3.2-3B. These intervals summarize variation across the sampled topic seeds; the ideal-randomness and support conditions for the fidelity floor remain those stated in Section B.3.

D Does a Rewriter Read What It Discards?

Table 8 probes the input-dependence condition in Corollary 1. Paired upstream documents share a prefix of length λn and differ in suffixes generated under independent upstream keys. Each pair is processed with the same downstream randomness and continuation length. The prefix-only arm (TC) excludes the suffix from the prompt; the full-rewriter arm (FR) includes it. We repeat the comparison over 3 downstream keys.

Divergence measures sensitivity to the discarded suffix. The divergence rate records whether paired outputs change when only the suffix changes. It is distinct from the information deficit $I(Y; K_2 \mid Y')$: with two inputs per prefix the deficit is at most $\log 2$ nats, and it is zero whenever the output identifies the input, even if every pair diverges. A finite-alphabet construction makes the distinction concrete. As the output alphabet grows through $4, 16, 256, 10^4$, its deficits are $0.478, 0.591, 0.190, 0.000$ nats while divergence approaches one. Divergence is therefore an interpretable sensitivity test, but it does not order operations by information deficit.

Observed rates and the prefix-only control. The full rewriter diverges on 100% and 100% of its 180 and 180 paired comparisons. The prefix-only prompts are identical within each pair: the Llama control has divergence rate 0.033, and the Qwen control 0.000. These controls provide the baseline against which the full-prompt sensitivity is read; the displayed Wilson intervals summarize the corresponding pairwise rates.

Separating rewriting from copying. The upstream-key detector rejection rates at 10^{-6} are 0.283 and 0.183 on full-rewriter rows. Reporting them beside divergence distinguishes suffix-sensitive generation from simple transcription. On Qwen3-1.7B, the fresh-entropy estimate falls from 43.5 nats under prefix-only continuation to 0.3 under the full prompt: access to the reference sharply constrains the generated continuation.

E Sharp Boundary Cases and Extensions

Example 1 (Why pipelines require fresh keys)**.** Let K_1, K_2 be independent uniform bits, $Y^{(0)} = K_1$, $Y^{(1)} = K_1 \oplus K_2$ and $Y^{(2)} = Y^{(1)} \oplus K_2$. Then $I(Y^{(2)}; K_1) = 1$ bit while $I(Y^{(1)}; Y^{(2)}) = 0$: the retention bound of Theorem 1 fails outright, because $I(K_1; K_2 \mid Y^{(1)}) = 1$ bit and the chain is not Markov. Symmetrically, with $Y^{(0)}$ constant, $Y^{(1)} = K_2$ and $Y^{(2)} = Y^{(1)}$, we get $F_A = H\big(Y^{(2)} \mid Y^{(1)}\big) = 0$ while $C_{\text{inj}} = I(Y^{(2)}; K_2) = 1$ bit. This is why Corollary 4 requires a fresh key at every hop, $K_2^{(i)} \perp (K_1, K_2^{(1:i-1)}, Y^{(0:i-1)})$. The fixed-key language-model diagnostics use distinct fixed hop keys and the stopping policy described in Section G.2.

Corollary 1 (When the corner is attained)**.** *Suppose (i) the upstream scheme is deterministic given K_1; (ii) the operation is deterministic given (Y, K_2); and (iii) the regenerated material depends on the input only through the part the operation keeps. Then $C_{\text{inj}} = F_A$ and $R_{\text{ret}} = M_A$ simultaneously: both caps are met at once.*

Remark 2 (A deterministic upstream pins retention). By (6), if the upstream scheme is deterministic given its key — as the exponential race is — then $R_{\text{ret}} = M_A$ *identically*, for every operation: retention is pinned to the transmitted information and only the injection coordinate carries any slack. Slack on the retention coordinate requires an upstream scheme that samples.

Corollary 2 (Separate coordinates at a fixed channel)**.** *The two caps of Theorem 1 govern separate coordinates: F_A bounds C_{inj}, while M_A bounds R_{ret}. If the upstream scheme is deterministic given its key, then (6) sharpens the second relation to $R_{\text{ret}} = M_A$ exactly. At a fixed channel, any change in retention is then exactly a change in the operation's transmitted information.*

Corollary 3 (Removal costs edits, priced by the task)**.** *Suppose the operation is ϵ-faithful almost surely, faithfulness permits changing at most m of n token positions, and the substitution relation has* pre-image *degree at most V_ϵ: for every output token u, at most V_ϵ input tokens v admit u at that position. Then, for every $1 \le m \le n$,*

$$N_\epsilon \;\le\; \sum_{j \le m} \binom{n}{j} V_\epsilon^{\,j} \;\le\; V_\epsilon^{\,m} \sum_{j \le m} \binom{n}{j} \;\le\; \left(\tfrac{enV_\epsilon}{m}\right)^m,$$

and

$$R_{\text{ret}} \ge I(Y; K_1 \mid X) - m \log \frac{enV_\epsilon}{m}. \tag{18}$$

Hence reducing R_{ret} below $I(Y; K_1 \mid X) - r$ requires $m \log(enV_\epsilon/m) \ge r$.

Remark 3 (Why the degree is a pre-image count). The hypothesis uses a *pre-image* degree for the same reason N_ϵ is a pre-image count (Remark 4). A forward degree gives a different quantity: a redaction that replaces any token by one mask symbol has forward degree 1 at every position, yet pre-image degree $|\mathcal{V}|$, and destroys $m \log |\mathcal{V}|$ nats in m edits. Substituting a forward degree into (18) would assert a floor that such an operation violates.

Remark 4 (Injectivity and canonicalizing transforms). A "format transformation" can produce a single canonical string while remaining many-to-one. That describes the forward ball, not the pre-image, and the two differ exactly here: lower-casing, whitespace collapse or Unicode normalization are many-to-one, so the pre-image is exponentially large and the bound is correspondingly weak. In the extreme, a constant map has singleton forward balls, is faithful, and retains nothing. $N_\epsilon = 1$ holds if and only if the reference map is injective on the support of Y.

Remark 5 (When the approximate floor is informative). Proposition 2 charges failures to $\log M$, rather than $\log |\mathcal{Y}|$. With a 63-bit key, $\log M \leq 43.7$ nats; a fixed-length vocabulary bound at $n = 192$ and $|\mathcal{V}| \approx 1.5 \times 10^5$ would instead be about 2.3×10^3 nats. Zero failures in N independent Bernoulli trials gives roughly $\eta \lesssim 3/N$ at 95% confidence. For copying, the sharp bound then gives a loss budget of order $(\log N + \log M)/N$. This statement concerns loss from the true upstream information; Section B.3 separates it from the sampling assumptions in the copy ladder.

Corollary 4 (Pipelines). *For a chain $Y^{(0)} \to \cdots \to Y^{(k)}$ whose hops satisfy Definition 1 with fresh total randomness, suppose each hop is exactly ϵ_i-faithful and draws a key satisfying $K_2^{(i)} \perp (K_1, K_2^{(1:i-1)}, Y^{(0:i-1)})$,*

$$R_{\text{ret}}^{(k)} \leq \min_{i \leq k} M_{A_i}, \qquad R_{\text{ret}}^{(k)} \geq I(Y^{(0)}; K_1) - \sum_{i \leq k} \log N_{\epsilon_i}.$$

Fresh keys are essential for the upper bound, as Example 1 demonstrates; the lower bound holds for every joint law.

F Additional Results and Literature Context

F.1 Complete operation measurements

The fixed-key grid uses 20 WikiText-103 topic seeds per model (Merity et al., 2016) and fixed generation keys, as specified in Section B. It spans copying, global replacement, constrained rewriting, translation, paraphrase, summary, and input-independent generation.

Copying and translation occupy different parts of this diagnostic plane (Figure 4). On Llama-3.2-3B, copying retains score 100.54 and injects 1.61; translation gives 0.55 and 158.42. Task execution matters: Qwen3-4B largely copies under the constrained-rewrite prompt. For summary, the three primary models have $2/20$, $0/20$, and $2/20$ inherited-key detections at 1% (Llama, Qwen3-1.7B, Qwen3-4B). Their wrong-key counts are $27/2440$, $25/2440$, and $25/2440$, and all three binary-KL diagnostics are zero. A mean score above 1 alone therefore does not establish retained information.

Table 7 gives the full four-model grid. Fresh entropy measures pathwise sampler randomness; replay measures entropy-weighted local reuse; injected and retained scores are mean $-\ln p$; and the final column is the binary-KL diagnostic. Section B.2 defines the corresponding sampling scope. Reading these columns side by side reveals how the same operation can occupy different regimes of generation, reuse, and detector response.

F.2 Further literature context

The generation problem has a classical ancestry in information hiding, steganography, and covert communication (Moulin & O'Sullivan, 2003; Wang & Moulin, 2007; Bloch, 2015). Entropy-constrained embedding and finite-length detection characterizations appear in He et al. (2025; 2026). Tempered optimizers and admissible-vocabulary constraints are studied by Huang et al. (2024); information profiles organize detection, attribution, and extraction in Li et al. (2026). Related constructions and converses include Gilani et al. (2026); Tsur et al. (2025); Guo et al. (2026); Huang et al. (2026). The hypothesis-testing perspective is developed by Huang et al. (2023); Li et al. (2024); Cai et al. (2024).

State amplification (Kim et al., 2008) describes communication in the presence of state information, building on coding with state (Gel'fand & Pinsker, 1980; Costa, 1983). State masking (Merhav & Shamai, 2007) instead asks how little state information can be exposed. Our use of the document as state and the fresh key as channel input highlights an operation's simultaneous effect on fresh and inherited keys. We rely on classical information identities and distinguish this fixed-operation accounting from coding-theorem achievability.

Empirical paraphrase robustness (Kirchenbauer et al., 2023b), bounded-edit guarantees (Zhao et al., 2023; Kuditipudi et al., 2023), rewriting chains (Ameen et al., 2026), and quality-oracle attacks (Zhang et al., 2023) concern different choices of channel or adversary. Our task-based floor connects these settings whenever the fidelity relation supplies a useful pre-image bound. Non-information-theoretic evasion analyses (Hwang et al., 2025) and low-entropy cryptographic questions (Mazor

Table 7: Complete hop-1 measurements with a second race watermark. F_A estimates fresh path entropy, $\widehat{M_A}$ is entropy-weighted replay, and $|S^\star|/n$ is pair reuse. Inject and retain are mean $-\ln p$; KL is the binary-KL diagnostic defined in Section B.2.

model	operation	F_A (nats)	$\widehat{M_A}$ (nats)	$\lvert S^\star\rvert/n$	inject (nats)	retain (nats)	KL (nats)
Llama-3.2-3B	copy	1.20	273.45	0.996	1.61	100.54	3.06
	local edit	2.80	181.81	0.617	1.15	61.12	3.04
	constr. rewrite	256.45	75.20	0.357	126.67	29.34	0.56
	translate	321.42	0.39	0.005	158.42	0.55	0.00
	paraphrase	261.92	8.54	0.062	98.96	1.11	0.00
	summarize	43.62	5.84	0.101	9.16	1.90	0.00
	free generation	374.38	1.33	0.005	159.92	1.00	0.00
Phi-4-mini	copy	0.56	386.19	0.981	1.27	152.60	1.26
	local edit	148.99	100.75	0.305	46.70	28.54	0.37
	constr. rewrite	197.82	116.43	0.334	64.32	38.31	0.04
	translate	113.71	0.84	0.007	22.49	1.35	0.00
	paraphrase	207.56	12.32	0.045	57.30	2.34	0.00
	summarize	70.86	1.76	0.038	19.34	1.19	0.00
	free generation	370.72	0.00	0.000	142.11	0.93	0.00
Qwen3-1.7b	copy	0.08	89.88	0.991	0.96	17.32	2.50
	local edit	1.69	78.64	0.842	0.93	14.14	2.98
	constr. rewrite	0.76	81.66	0.926	0.97	15.58	2.73
	translate	67.67	0.49	0.020	10.96	1.36	0.00
	paraphrase	42.29	14.57	0.231	6.96	1.75	0.00
	summarize	16.88	5.31	0.196	2.63	1.26	0.00
	free generation	97.00	0.07	0.002	22.26	0.96	0.00
Qwen3-4b	copy	0.01	79.90	1.000	0.70	13.18	2.55
	local edit	1.83	49.43	0.626	0.66	7.64	1.10
	constr. rewrite	0.07	79.90	1.000	0.70	13.18	2.50
	translate	38.23	0.61	0.019	3.63	0.88	0.00
	paraphrase	35.90	13.01	0.244	4.10	1.23	0.00
	summarize	13.82	3.27	0.176	2.06	1.74	0.00
	free generation	95.22	0.01	0.002	17.85	1.29	0.00

et al., 2026) add detector- and construction-specific perspectives. Together, these strands separate three objects that are often conflated: information retained by an operation, evidence read by a verifier, and robustness engineered into a particular watermark scheme.

G Additional Operation Diagnostics

These fixed-key experiments examine how inherited detector evidence changes when we vary the downstream watermark, repeat an operation, restrict input access, or change the verifier. Detector scores and positional replay retain the measurement definitions in Section B.2.

G.1 Matched re-watermarking

The race and plain arms process the same upstream documents, allowing paired comparisons of retained scores. Figure 10 shows the resulting relation between detector evidence and positional replay for the three primary models. The statistical grid also includes Phi-4-mini, giving 28 model–operation cells with 20 paired documents. We apply exact two-sided sign tests to the untied score differences and Holm correction over all 28 cells, retaining all-tied cells with $p = 1$. No test rejects; the smallest adjusted p-value is 0.331. The tests concern sign balance, not mean differences. The cellwise bootstrap intervals remain wide: the largest magnitude of a negative endpoint is 54.63 score units. This comparison therefore does not establish equivalence between the two arms.

The sample size sets a demanding threshold for this comparison. With 20 untied pairs, an exact two-sided sign test at 5% needs at least 15 differences in the same direction. At the first Holm threshold

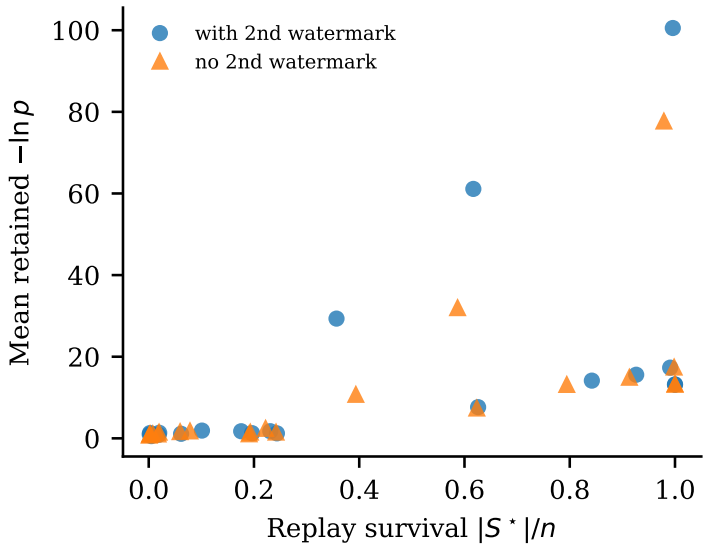

Figure 10: Mean retained detector score $-\ln p$ against positional replay $|S^\star|/n$ for each model–operation cell in the three primary models, with and without a second race watermark. The paired arms process the same upstream documents.

for 28 comparisons, at least 18 of the 20 signs must agree. A binomial calculation gives 80% power to cross this first threshold only when the probability of a positive untied difference is about 0.922; at probability 0.8, the power is about 20.6%. Ties further reduce the number of observations available to the test. This calculation describes sign-test resolution, not a minimum detectable difference in mean detector score.

G.2 Repeated transformations and the stopping policy

Repeated rewriting need not follow a constant survival factor. For Qwen3-1.7B translation, original-document replay is 0.0199, 0.0200, and 0.0200 across the three summaries. A geometric reference using only the first-hop factor gives 0.01994, 0.00040, and 0.00001. The first transformation removes most original local matches; later transformations change their remaining fraction much less. Llama's constrained-rewrite detector score continues to decrease (Figure 11). The geometric curve is a homogeneous-survival reference, not a prediction of Corollary 4.

Generation stops when an output exactly matches its input token sequence or contains fewer than 40 characters after trimming whitespace. Later summaries carry the last output forward, so each summary includes the original document set. These curves describe this stopping policy: exact reproduction under one random draw need not persist under a fresh key. They are therefore distinct from running three newly sampled operations on every document.

The Qwen translation plateau begins before any output is carried forward. All 20 documents have new outputs at each of the first two hops, with replay fractions 0.01994 and 0.02000. At hop 3, 14 documents have new outputs; their mean replay is 0.02146, compared with 0.01999 after including the six carried outputs, so carrying outputs forward slightly lowers the third-hop mean. Restricting all three hops to the same 14 documents gives replay fractions 0.0214, 0.0215, and 0.0215. Qwen3-4B shows the same pattern among its 9 complete trajectories: 0.0179, 0.0181, and 0.0180. These subsets are selected by the stopping rule; Section H.4 runs every hop on every document.

Key generations. Among generated Llama-3.2-3B translation outputs at later hops, the previous hop's key is detected on $21/40$ outputs, whereas the original key is detected on $0/40$ at the 1% level. Local editing shows the reverse pattern: the original key is detected on $33/40$ outputs and the previous key on $3/40$. These rates describe the generated outputs in the respective chains. Transla-

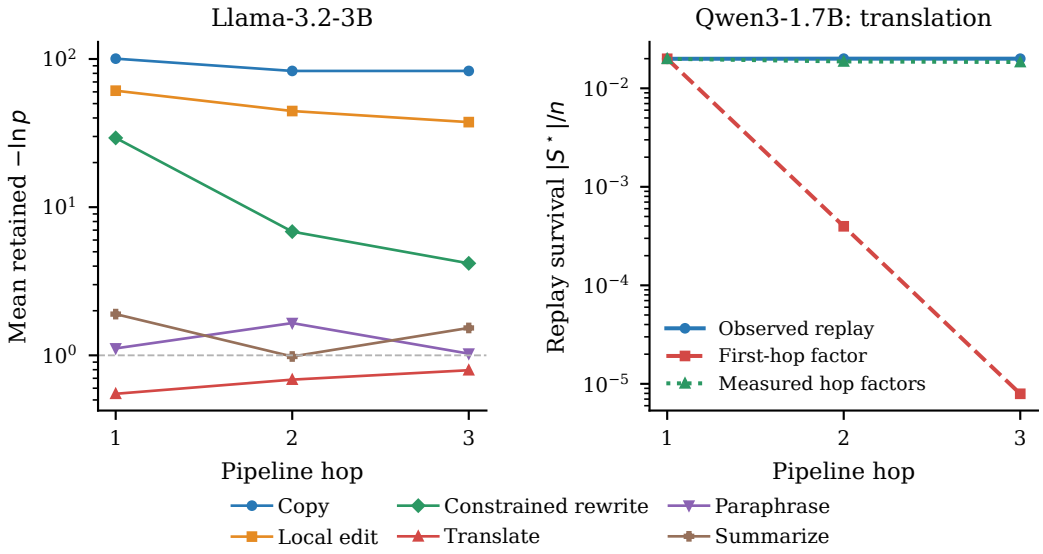

Figure 11: Repeated transformations with the stopping policy described in the text. Left: retained detector scores $-\ln p$ on Llama-3.2-3B. Right: original-document replay for Qwen3-1.7B translation, compared with a first-hop geometric reference and the product of measured mean per-hop factors. Later summaries carry stopped outputs forward.

tion favors the more recent key in this comparison, whereas local editing preserves evidence of the original one.

G.3 Discarded-suffix dependence

A rewriter can use suffix information even when that suffix does not appear in the output. We compare paired documents with identical prefixes and different suffixes, using the same downstream key and forced prefix for both members. The truncate-and-continue arm (TC) sees only the prefix; the full-rewriter arm (FR) also sees the suffix. This intervention probes the input access underlying the sufficient condition in Corollary 1.

Table 8: Paired-prefix comparison. Diverged is the fraction of unequal output pairs under the same key. The final column gives the complementary upstream-key copying diagnostic at 10^{-6}; intervals are Wilson intervals.

Model	Arm	pairs×keys	diverged	95% CI	detect@10^{-6}
Llama-3.2-3B	TC	180	0.033	[0.015, 0.071]	0.000
Llama-3.2-3B	FR	180	1.000	[0.979, 1.000]	0.283
Qwen3-1.7b	TC	180	0.000	[0.000, 0.021]	0.000
Qwen3-1.7b	FR	180	1.000	[0.979, 1.000]	0.183

The full rewriter changes every paired output. The prefix-only divergence rates are 0.033 on Llama and 0.000 on Qwen, each over 180 comparisons. The result establishes sensitivity to the discarded suffix. Divergence does not measure the conditional-information deficit $I(Y; K_2 \mid Y')$: Section D gives examples with large divergence and zero deficit, together with the complementary copying diagnostic. The experiment is therefore an input-access check rather than a numerical test of the deficit identity.

G.4 Verifier-context mismatch

Holding text fixed separates a change in the verifier from a change in the channel. We replay the detector on each copy text at widths $w \in \{1, 2, 3, 4, 6, 8\}$; generation used $w = 4$. Only the contexts used to reconstruct the pseudorandom scores change.

Table 9: Context-width sensitivity. Cells give mean detector score $-\ln p$ and rejection rate at nominal 10^{-6}. Generation uses $w = 4$ (starred). At the other tested widths, rejection rates are zero and scores are near the null reference level.

Model	detector context width w					
	1	2	3	$4^\star$	6	8
Llama-3.2-3B	1.0 (0.00)	1.0 (0.00)	1.0 (0.00)	113.3 (0.95)	1.0 (0.00)	1.0 (0.00)
Qwen3-1.7B	0.8 (0.00)	0.9 (0.00)	1.0 (0.00)	18.5 (0.69)	0.9 (0.00)	0.9 (0.00)
Qwen3-4B	0.9 (0.00)	0.8 (0.00)	1.2 (0.00)	14.5 (0.51)	1.0 (0.00)	1.0 (0.00)

Llama's matched-width mean score is 113.3, versus at most 1.0 off-width. Both Qwen models have zero observed off-width rejections at nominal 10^{-6}. This is the expected dependence of a context-seeded detector on its context convention. Its role here is a measurement control: the texts and their information about the upstream key are identical throughout this comparison. Section C gives the prefix-length and substitution studies.

G.5 A GREEN-LIST SECOND WATERMARK

The green-list arm adds a second watermark with $\gamma = 0.25$ and logit boost 2, while the upstream document still carries the exponential-race watermark. Table 10 reports all seven operations for the three primary models. Its detector scores test the upstream race key; the table therefore describes inherited evidence under a different downstream watermark family.

Across Llama-3.2-3B, Qwen3-1.7B, and Qwen3-4B, copying retains mean scores of 103.99, 17.30, and 13.18; translation gives 0.62, 1.40, and 0.88, with replay at most 0.018. Qwen3-4B still copies under the constrained-rewrite prompt (replay 1.000, score 13.18), recording prompt noncompliance rather than a successful rewriting regime.

Table 10: Inherited race-watermark evidence after a green-list second watermark, for the three primary models (20 documents per cell). Each cell reports the retained detector score $-\ln p$ (mean $\pm$ standard error), followed by positional replay. These scores test the upstream race watermark; they do not measure the injected green-list watermark.

Operation	Llama-3.2-3B	Qwen3-1.7B	Qwen3-4B
Copy	103.99 ± 34.06; 0.996	17.30 ± 1.25; 0.997	13.18 ± 1.19; 1.000
Local edit	40.31 ± 11.78; 0.577	14.27 ± 1.09; 0.847	7.80 ± 1.13; 0.619
Constrained rewrite	3.05 ± 0.75; 0.206	15.75 ± 1.38; 0.948	13.18 ± 1.19; 1.000
Translate	0.62 ± 0.12; 0.005	1.40 ± 0.30; 0.018	0.88 ± 0.21; 0.018
Paraphrase	1.12 ± 0.21; 0.044	2.08 ± 0.43; 0.216	1.42 ± 0.21; 0.239
Summarize	1.62 ± 0.19; 0.067	1.44 ± 0.32; 0.203	1.07 ± 0.24; 0.179
Free generation	1.04 ± 0.27; 0.000	0.88 ± 0.18; 0.003	1.10 ± 0.23; 0.002

H CONTROLLED LANGUAGE-MODEL STUDIES

The controlled studies extend the fixed-key operation grid in three directions: larger paired comparisons, an explicit factual-record task, and complete generation of every hop. They use Qwen3-4B and Llama-3.2-3B-Instruct, abbreviated Qwen and Llama in the tables. This section reports the full cohorts, without selecting documents on initial detection or downstream success.

H.1 DESIGN AND MEASUREMENT

For each model, the paired study uses 200 upstream documents and three operations (copy, paraphrase, summary), each with plain and race sampling. The chain study uses 64 upstream documents, two operations (paraphrase and translation), two arms, and three actual hops. The factual-record study uses 64 upstream reports and three operations with two arms. Together these comprise 656 upstream and 4,704 downstream generations. Sample sizes, prompts, scores, and tests were fixed in advance; pilot runs used separate documents.

Table 11: Upstream cohorts before any downstream operation. Key detections use 1%; null is the mean percentage of 128 wrong keys detected per document. Score intervals resample documents. Length is the mean number of emitted tokens; cap counts generations that exhaust the token budget. No document is selected on detection.

Model	Suite	Key	Mean $-\ln p$ [95% CI]	Null (%)	Length	Cap
Qwen	Paired	192/200	15.30 [14.31, 16.29]	0.95	183.6	86/200
Qwen	Three-hop	60/64	13.21 [11.85, 14.64]	0.99	187.8	40/64
Qwen	Fields	14/64	3.28 [2.64, 3.98]	0.98	111.8	0/64
Llama	Paired	200/200	146.22 [119.55, 176.00]	0.90	185.0	93/200
Llama	Three-hop	64/64	130.40 [88.73, 180.54]	0.95	185.0	29/64
Llama	Fields	61/64	27.48 [19.68, 37.86]	1.01	110.4	1/64

Inputs and generation. Unstructured documents are stories prompted by themes from WikiText-103 (Merity et al., 2016). We take the first eligible source paragraphs in corpus order, using 400–900-character excerpts and up to 160 characters of the first sentence as the theme. The paired and chain studies share this source collection but generate their own upstream texts and keys. These are fixed source collections, not independent topic samples. Each source document is shared by all its downstream comparisons. The prompts ask for exact copying, preservation of meaning in a paraphrase, concise summarization, or English-to-French translation. Translation repeats the same English-to-French instruction at later hops, including when the input is already French.

Both arms sample at temperature 1 from the full valid vocabulary, without top-k or nucleus truncation. Plain sampling uses the base model law; race sampling uses the exponential-race watermark with context width 4. Generation budgets are 192 tokens for upstream stories, 384 for their downstream outputs, and 256 for every factual report; a generation may end earlier at the end-of-sequence token. Upstream keys, wrong keys, sampling randomness, and downstream keys come from separate pseudorandom streams for each document, and the downstream race key changes with the hop. Arm order is counterbalanced across documents. Every first generation is kept, including failed, short, unchanged, and initially undetected outputs. In the chain study, each downstream input contains the actual preceding text.

Endpoints and uncertainty. As in Section B.2, the detector scores each distinct (context-seed, token) pair once; we report whole-output $-\ln p$ and rejection at $\alpha = 0.01$. The null rate uses 128 wrong keys per output, averaged within each document, so the document remains the sampling unit. All intervals below are percentile intervals from 5,000 document-bootstrap resamples. Paired contrasts resample the same document indices in both arms, and chain comparisons resample whole trajectories. These intervals describe document-level variation under the evaluated sources and generation settings, not uncertainty over a representative topic population.

Table 11 records the initial detector evidence and lengths. In particular, the Qwen factual cohort begins with only $14/64$ detections, whereas Llama begins with $61/64$; a low downstream count must be read against that baseline. A cap hit means that generation used its whole token budget without reaching the end-of-sequence token.

H.2 The paired comparison and its resolution

Table 12 adds score intervals, null rates, lengths, and cap counts to the arm-level results in Table 2. Exact two-sided sign tests discard zero differences only when counting signs. All 200 documents, including ties, remain in the mean differences and bootstrap intervals, and all six model–operation cells remain in the Holm family. Thus the mean intervals and sign tests address different summaries of the same paired observations.

With 200 untied pairs, crossing the conservative first Holm threshold $0.05/6$ requires at least 120 signs in one direction. Conditional on that untied count, 80% power requires a positive-sign probability of about 0.6262, with the symmetric condition for negative signs. Llama copying has 61 untied pairs, requiring 42 agreeing signs and probability 0.7279 for 80% power. Qwen copying has only 6: even unanimous signs give two-sided $p = 0.03125$, above $0.05/6$. These calculations describe

the resolution of the sign test; they neither bound mean-score differences nor show that the arms are equivalent. The fixed-key 28-cell analysis in Section G.1 is a separate experiment and test family.

Table 12: Arm-level paired-study measurements, all 200 documents per row. Scores have pointwise 95% document-bootstrap intervals; key and wrong-key detections use 1%. Length is the mean number of tokens; cap counts generations that exhaust the token budget. The paired comparison, rather than overlap of these marginal intervals, is in Table 2.

Model	Operation/arm	Mean $-\ln p$ [95% CI]	Key	Null (%)	Length	Cap
Qwen	Copy/plain	14.97 [14.00, 15.97]	191/200	0.93	183.2	0
Qwen	Copy/race	14.98 [14.02, 15.98]	191/200	0.93	183.2	0
Qwen	Paraphrase/plain	2.33 [2.07, 2.60]	25/200	1.00	181.6	0
Qwen	Paraphrase/race	2.28 [2.01, 2.55]	28/200	1.01	181.5	0
Qwen	Summary/plain	1.38 [1.20, 1.58]	10/200	0.91	79.2	0
Qwen	Summary/race	1.38 [1.21, 1.55]	7/200	1.03	79.2	0
Llama	Copy/plain	109.21 [90.57, 130.10]	189/200	0.95	186.1	5
Llama	Copy/race	106.52 [88.15, 126.97]	193/200	0.93	182.5	2
Llama	Paraphrase/plain	1.83 [1.46, 2.38]	9/200	0.99	185.3	5
Llama	Paraphrase/race	1.69 [1.46, 1.94]	10/200	1.09	193.2	8
Llama	Summary/plain	1.45 [1.24, 1.66]	14/200	1.04	73.2	0
Llama	Summary/race	1.56 [1.35, 1.79]	8/200	1.01	71.1	0

H.3 Factual reports: task success and input access

Each target record has four fields: organizer and city strings, and year and attendance integers. The task generator draws from four organization names, six cities, years 2001–2025, and attendance counts 31–250, separately from the watermark-key domain. The upstream prompt requests a single JSON object with exactly `facts` and `narrative`; the latter should contain 60–90 words, consistent with the facts and without invented activities, venues, or outcomes. Strict parsing rejects extra or duplicate fields, wrong types, commentary outside the JSON, and malformed objects. Four-field success additionally requires every value to equal its target. Narrative length is scored separately.

Full-context rewriting receives the complete upstream report. Task-only regeneration receives only canonically ordered parsed fields, with no source narrative or source formatting. If the source cannot be parsed, it receives the fixed instruction to return an empty object, never the hidden target record. Such an output follows the missing-input instruction but fails the primary four-field task. This design compares input-access policies as implemented, including their execution; parsing is part of the evaluated procedure, and no document is selected on it.

Full cohort and upstream-success subset. Table 1 covers all 64 reports per cell. The sensitivity columns in Table 13 keep the 38 Qwen or 49 Llama documents whose upstream reports already satisfy the field task; they select on upstream success only, never on output success. Qwen full-context rewriting succeeds on all 38 of these inputs in either arm, with zero whole-output detections; the full cohort also includes repaired malformed inputs. Copying can also change punctuation or repair JSON, so a copy instruction does not guarantee an exact copy.

The paired alternative-minus-copy comparisons in Table 14 retain all 64 documents. There are eight contrasts for each endpoint; their 95% intervals are pointwise descriptions, not simultaneous coverage or multiplicity-adjusted significance decisions. The table shows the task consequences beside the score reductions.

Narrative availability and quality. The narrative detector scores the emitted tokens of the narrative span, skipping its first four tokens so that every scored context lies inside the narrative. When parsing fails or the span cannot be located in the emitted tokens, the narrative score is missing rather than zero. The resulting detection counts in Table 13 supplement the whole-output primary endpoint. Across all upstream and downstream factual reports, narrative scores are available on 339/448 Qwen and 303/448 Llama outputs. Only 89/448 and 109/448, respectively, satisfy the requested word range under strict parsing. Field success therefore does not imply that the narrative instructions were followed.

Table 13: Factual-report diagnostics. Parse and 60–90-word counts have denominator 64; invalid JSON fails the length check. Narrative key detections use only the available narrative spans (displayed denominators equal parse counts). The last two columns restrict to documents whose upstream report passes the field test (38 Qwen, 49 Llama); no selection uses output success. All detections use 1%.

Model	Operation/arm	Parse	60–90	Narrative Key	Upstream-success subset Fields	Key
Qwen	Copy/plain	61	15	21/61	38/38	7/38
Qwen	Copy/race	61	15	21/61	38/38	7/38
Qwen	Rewrite/plain	64	10	1/64	38/38	0/38
Qwen	Rewrite/race	64	11	2/64	38/38	0/38
Qwen	Task-only/plain	27	12	0/27	27/38	0/38
Qwen	Task-only/race	24	13	0/24	24/38	0/38
Llama	Copy/plain	51	15	50/51	49/49	46/49
Llama	Copy/race	51	15	50/51	48/49	46/49
Llama	Rewrite/plain	40	17	1/40	33/49	1/49
Llama	Rewrite/race	39	21	1/39	32/49	2/49
Llama	Task-only/plain	35	12	1/35	35/49	1/49
Llama	Task-only/race	38	14	0/38	38/49	0/49

Table 14: Alternative minus copy on the same 64 factual reports. Score differences use whole-output $-\ln p$; field-success differences are percentage points. The eight contrasts per endpoint have pointwise 95% document-bootstrap intervals, not simultaneous coverage or multiplicity-adjusted decisions. All 64 reports contribute.

Model	Alternative/arm	Score difference [95% CI]	Fields (pp) [95% CI]
Qwen	Rewrite/plain	−2.11 [−2.73, −1.56]	4.7 [0.0, 10.9]
Qwen	Task-only/plain	−2.22 [−2.81, −1.68]	−53.1 [−65.6, −40.6]
Qwen	Rewrite/race	−1.99 [−2.59, −1.46]	4.7 [0.0, 10.9]
Qwen	Task-only/race	−2.13 [−2.73, −1.58]	−57.8 [−70.3, −45.3]
Llama	Rewrite/plain	−25.51 [−34.53, −18.40]	−17.2 [−31.2, −3.1]
Llama	Task-only/plain	−25.76 [−34.83, −18.64]	−25.0 [−35.9, −14.1]
Llama	Rewrite/race	−25.49 [−34.51, −18.49]	−20.3 [−32.8, −7.8]
Llama	Task-only/race	−25.69 [−34.80, −18.59]	−20.3 [−31.2, −9.4]

Reading the outputs also reveals prose errors despite exact fields. One Qwen report keeps the attendance field at 223 while describing "more than 223" attendees; other outputs introduce unsupported activities or venues. We therefore call the task metric four-field success, not semantic equivalence, and score every output as generated.

Across downstream factual cells, mean output lengths range from 69.98 to 114.00 tokens. Qwen has no cap hits among 384 downstream reports; Llama has one, as well as one upstream hit. Whole-output wrong-key detection rates across these cells range from 0.916% to 1.160%; available-narrative cell rates range from 0.684% to 1.317%. These rates, clustered within documents, check the verifier's null behavior.

H.4 Complete trajectories and the fixed-key stopping rule

Every document undergoes three actual operations. For comparison, we apply the fixed-key study's rule to the same trajectories after generation: stop at the first exact token-sequence reproduction or output shorter than 40 trimmed characters, and carry that output into later summaries. We compare this simulated policy with the actual continuation on the same documents. A stop first triggered at hop 3 has no effect within the observed horizon; the carried column in Table 15 counts stops before hop 3.

The inherited-key detection counts at hop 3 agree in every cell, even where scores change. The largest absolute mean-score change is 0.0269 on Qwen plain translation. The comparison concerns this binary endpoint on the same trajectories; it does not show that reproduced outputs stay fixed under new keys. None of the fixed-key study's outputs enter these trajectories, which also use longer generation budgets.

Table 15: Complete three-hop trajectories: 64 documents at every hop in each row. Key and cap counts are listed for hops 1/2/3. Carried counts trajectories already stopped before hop 3 under the fixed-key stopping rule; all actual outputs were newly generated. The final column gives the actual minus the simulated carry-forward hop-3 score, with a pointwise 95% paired interval. Every hop-3 detection-count difference at 1% is zero.

Model	Operation/arm	Key 1/2/3	Cap 1/2/3	Carried	Score difference [95% CI]
Qwen	Paraphrase/plain	3/13/4	0/0/0	0	0.000 [0.000, 0.000]
Qwen	Paraphrase/race	8/12/7	0/0/0	0	0.000 [0.000, 0.000]
Qwen	Translate/plain	1/1/1	0/0/0	22	−0.027 [−0.065, 0.002]
Qwen	Translate/race	1/1/1	0/0/0	14	−0.003 [−0.009, 0.000]
Llama	Paraphrase/plain	6/2/1	2/0/1	1	−0.021 [−0.063, 0.000]
Llama	Paraphrase/race	3/2/2	3/2/0	0	0.000 [0.000, 0.000]
Llama	Translate/plain	0/1/0	5/3/1	1	0.001 [0.000, 0.004]
Llama	Translate/race	0/0/0	12/0/0	0	0.000 [0.000, 0.000]

Cellwise mean output lengths range from 181.06 to 284.48 tokens. There are no downstream Qwen cap hits; Llama has $29/768$, including 21 translations. All are included in the trajectory comparison. Whole-output wrong-key rates across cells range from 0.842% to 1.221%. The repeated translation prompt can yield unchanged French text, malformed expressions, or meaning changes, so completing all hops does not indicate translation quality.

I THE EPIGRAPH

The epigraph is Heraclitus's river fragment in Burnet's translation (Burnet, 1920). It asks of a river the question this paper asks of a document: what is still the same after the water has been replaced? A rewrite replaces the water of a text. We read the fragment through the coupling (Y, Y') of Definition 2; it serves as an image for the formal statements, not as an interpretation of Heraclitus.

Old, kept and fresh water. The water that has flowed away is the erased budget $E_A = H(Y \mid Y')$, the water that is still there is the transmitted budget $M_A = I(Y; Y')$, and the fresh water is $F_A = H(Y' \mid Y)$. Together they account for both texts: $H(Y) = E_A + M_A$ and $H(Y') = F_A + M_A$.

A key moves with the water. By Theorem 1, information about any key K enters Y' only with the fresh water and leaves only with the erased water: $I(K; Y') - I(K; Y)$ is an inflow at most F_A minus an outflow at most E_A. The rewriter's own key is independent of the old text, so it arrives only with the fresh water and $C_{\text{inj}} \leq F_A$. The inherited key is carried by the old water. The rewriter never sees it, so it has no inflow: what is not carried away with the erased water can remain only in the water that stays, and $R_{\text{ret}} \leq M_A$.

What makes it the same river. The fragment still speaks of the same rivers. Which rewrites count as the same document is decided by the task, not by the water. By Theorem 2, every admissible output reveals the component of its input, the shadow $\tau_\epsilon(Y)$, so the key information in the shadow survives every faithful rewrite: $R_{\text{ret}} \geq I(K_1; \tau_\epsilon(Y))$. When the task preserves a statistic T, this floor is exactly $I(K_1; T(Y))$. Conditional resampling attains it: it replaces all the water the task lets go and keeps the distribution of the text. A key independent of T can therefore be erased completely.

Two questions kept apart. A rewrite can thus be the same document in the task's sense and still erase the old key. Section 6 shows this separation in a language model: Qwen3-4B rewrites all four target facts correctly in $64/64$ reports, while inherited-key detections fall from $14/64$ upstream to $1/64$. Whether a rewrite is faithful is a question about the task; whether the old key survives it is a question about the three budgets.