# The Price of a Summary:

Information-Theoretic Limits of Agent Oversight

Erich Owens
The Harbor Research Program erich.owens@gmail.com

August 2026

**Abstract**

A human overseeing a fleet of software agents reads digests, not transcripts. We show that this practice obeys exact information-theoretic limits, and we price them. Four results: (i) any digest that *guarantees* the overseer catches all $k$ critical artifacts among $N$ while opening only $m$ must carry at least $\log_2 \binom{N}{k} - \log_2 \binom{m}{k}$ bits — a floor that survived a pre-registered falsification experiment (0/16 violations, including an oracle encoder); (ii) one scalar summary head serves two readers if and only if their preference orders are comonotone — and the overseer and the successor agent provably are not, with the joint floor *super*-additive ($\approx 2.13\times$, not $2\times$); (iii) the folklore ranking key stakes $\times$ irreversibility $\times$ anomaly is the Bayes-optimal surfacing rule iff anomaly is a calibrated posterior; (iv) a two-constraint rate–distortion program $R(\delta, f)$ — false-negative rate priced separately from flag rate — has a closed form for Bernoulli sources, with the covering floor as its zero-miss corner, and a new zoom-advantage theorem: adaptive halving identifies the $k$ criticals inside $F$ flagged items in at most $2k\lceil \log_2(F/k) \rceil + 4k$ group opens, a bound whose leading constant is achieved. Every number is tagged as externally recomputable or as regenerating from a named script at a fixed seed.

## 1 Introduction

**The scene.** An operator supervises forty coding agents overnight. By morning the fleet has produced hundreds of artifacts — branches, migration scripts, config edits, deleted test suites. Two of them are critical: if the operator does not open exactly those two before the morning merge, an irreversible mistake ships. The operator will read one screen of digest and open perhaps eight items. The question this paper answers with theorems rather than taste: *what must that screen carry, what can it never do, and what does each relaxation buy?*

**Express lane.** One sentence per result, formal statements in the boxes. A zero-miss digest is a message that must single out a covering $m$-subset, so it costs $\log_2 \binom{N}{k} - \log_2 \binom{m}{k}$ bits (box in §2); one scalar head serves two readers iff their orders are comonotone, and oversight's two readers are not (box in §3); the inspect-or-pass rule is a likelihood-ratio test, making "anomaly" a calibration obligation (box in §4); and pricing misses and flags separately yields a closed-form frontier $R(\delta, f)$ whose flagged set is then searched in $2k\lceil \log_2(F/k) \rceil + 4k$ adaptive opens — the one theorem newly proved in this paper (boxes in §5).

*The idea in one breath: a summary is a channel, an overseer is a decoder with an open budget, and every oversight guarantee therefore has an exact price in bits, opens, and misses — prices we compute, and one of which we bound with a new, constant-explicit group-testing theorem.*

**Contributions.**

1. **The information floor, falsified-against** (§2): the covering bound $B^\star = \log_2 \binom{N}{k} - \log_2 \binom{m}{k}$, plus a pre-registered experiment in which no encoder — oracle, noisy, or random

— beat it (0/16 violations at $N$=60, $k$=2, $m$=8 [internal, `a7_experiment.py`]). We also report the instructive failure: a first meter charged bits only for tie-breaking and produced 8/14 spurious "violations," a worked example of the unpriced-side-channel trap.

2. **The split-digest characterization** (§3): a scalar head serves two readers iff their induced orders are comonotone; the operator (regret-shaped) and successor (fit-shaped) readers order constructible pairs oppositely, so no shared head exists — and the joint zero-miss floor is super-additive: $\mathrm{Floor}(N, 2k, m) \approx 2.13 \times \mathrm{Floor}(N, k, m)$, a ratio-over-double of 1.067 at $N$=60, $k$=2, $m$=8 that exceeds 1 at every regime and grows without bound in $k$ (1.224 at $k$=4, $m$=8; 1.392 at $k$=10, $m$=20) [verified].
3. **The derived regret head** (§4): stakes × irreversibility × anomaly is exactly expected unrecoverable loss iff anomaly is the calibrated posterior $\Pr(\texttt{bad} \mid \text{evidence})$; the optimal surfacing rule is a likelihood-ratio threshold, and raising the miss-to-false-alarm cost ratio *lowers* the bar to inspect.
4. **The two-constraint frontier and the zoom theorem** (§5): a custom rate–distortion program $R(\delta, f)$ with false negatives priced separately from flag rate, solved in closed form for Bernoulli sources (the optimizing joint is fully pinned); and a new worst-case bound for the second stage — adaptive halving finds all $k$ positives among $F$ flagged items in at most $2k\lceil \log_2(F/k) \rceil + 4k$ group opens versus $F$ flat, proved by a charging argument, shown tight in its leading constant, and consistent with the measured 15.3× advantage at $(F, k) = (2500, 10)$ [internal, `b1_frontier.py`].

Results 1–3 were derived and verified in the program's execution reports; result 4's frontier was computed there but its zoom bound existed only as a simulation regularity — the theorem and proof in §5.2 are new to this paper. Throughout, numbers are tagged **[verified]** (externally recomputable from the closed form) or **[internal]** (regenerates from the named script at seed 20260816).

## 2 The information floor, and its falsification test

**The analogy, mapped by relations.** A digest is not a description of the night's work; it is a *message* that selects, from a codebook shared with the overseer, which $m$ items to open (Figure 1). The relation that transfers: a $B$-bit message distinguishes at most $2^B$ codewords, and each codeword — an $m$-subset — covers only $\binom{m}{k}$ of the $\binom{N}{k}$ ways the $k$ critical items could be placed. The predictable misread, preempted: the floor does *not* say a digest needs $B^\star$ bits to be useful — it prices the *zero-miss guarantee* only; cheaper digests are fine the moment misses are tolerated (§5 prices exactly that).

**Theorem 1 (information floor).** Fix $N$ items containing an arbitrary unknown $k$-subset $T$ of critical items, and an open budget $m$. Any digest scheme — an encoder $e$ mapping the corpus to $B$ bits and a data-independent decoder $d$ mapping $B$ bits to an $m$-subset to open — that guarantees $T \subseteq d(e(\cdot))$ for *every* placement of $T$ must have

$$B \;\geq\; B^\star(N, k, m) \;=\; \log_2 \binom{N}{k} - \log_2 \binom{m}{k}.$$

*Proof.* The decoder realizes at most $2^B$ distinct $m$-subsets; an $m$-subset contains $\binom{m}{k}$ of the $\binom{N}{k}$ possible $k$-subsets; zero miss requires the realized $m$-subsets to jointly cover all placements, so $2^B \binom{m}{k} \geq \binom{N}{k}$. □

**Numbers by hand.** At $N$=60, $k$=2, $m$=8: $\log_2 \binom{60}{2} = \log_2 1770 = 10.79$, $\log_2 \binom{8}{2} = \log_2 28 = 4.81$, so $B^\star = 5.98$ bits [verified] (Table 1 collects this and every other headline

| Quantity ($N$=60, $k$=2, $m$=8) | Value | Provenance |
|---|---|---|
| Floor $B^\star$ | 5.98 bits | [verified] |
| Floor, disjoint readers ($2k$=4; §3) | 12.77 bits | [verified] |
| Ratio $\text{Floor}_{2k}/\text{Floor}_k$ | 2.13× | [verified] |
| Super-additivity $\text{Floor}_{2k}/(2\,\text{Floor}_k)$ | 1.067 | [verified] |
| Oracle violations of the floor | 0/16 | [internal, `a7_experiment.py`] |
| Wrong-turn v1 spurious violations | 8/14 | [internal, `wrong-turns/`] |

Table 1: Every headline number in §2–§3, with its evidence tier. The four rows above the inner rule are closed-form quantities recomputable by hand from Theorem 1; the two below it are simulation counts, including the wrong turn's 8/14 spurious violations, retained rather than deleted. The pairing that matters: the joint floor is 12.77 bits where naive doubling budgets 11.96.

number with its evidence tier). Now you try: $N$=100, $k$=2, $m$=10 — compute $\log_2 4950$ and $\log_2 45$ and check you get a floor just under 6.8 bits.

**The falsification experiment.** Theorem 1 makes a sharp prediction: a $B$-bit digest's best zero-miss probability is $\min(1, 2^B\binom{m}{k}/\binom{N}{k})$, so *any* scheme dipping below the implied miss curve refutes the formalization. We implemented three encoders — an *oracle* that sees the true set and addresses the best codeword its $B$ bits can reach, a *noisy* encoder (feature SNR $\approx 1$), and a *random* one — against a shared randomized-cover codebook, 4000 placements per budget point, sweeping $B$ through the floor at $N$=60, $k$=2, $m$=8. Result: **0/16 violations**; the oracle hugs the floor from above, the others pay a premium [internal, `a7_experiment.py`]. The floor's regime behavior is Figure 2: it falls as the open budget grows and reaches zero only at $m = N$.

**A wrong turn, reported.** A first implementation handed the oracle a perfect per-item score vector and charged $B$ bits only for resolving rank ties: 8/14 apparent violations. That was not a refutation but an unpriced side channel — identity of the critical set leaked outside the metered message. Rebuilding the model so $B$ is a literal message length gating an encoder/decoder pair restored agreement. The transferable rule: a simulation that appears to beat an information bound is leaking through an unpriced channel; audit the meter before doubting the theorem — and before trusting the "refutation."

## 3 One summary cannot serve two readers

**The analogy, mapped by relations.** A scalar summary head is a thermometer: one number, one order, every reader who consults it ranked the same way (Figure 3). That is fine when the readers want the same thing — a nurse and a pharmacist both want the hottest patient first. It fails the moment two readers want quantities that do not move together. Every compaction here is read twice: by the *successor* — the next agent instance, which wants what is worth continuing (fit-shaped, forgone continuation value) — and by the *operator*, who wants what is dangerous to ignore (risk-shaped, regret-if-ignored). The relation that transfers is not "a summary resembles a temperature" but the harder fact underneath it: a single scalar induces exactly one total order, so it can serve two readers only if the two readers' own orders never cross. Standard practice stores one summary for both, and therefore silently asserts that they never do. The misread to preempt: this is not a claim that scalars are too coarse — a scalar with *infinite* precision fails identically, because the obstruction is ordinal, not numeric.

**Theorem 2 (comonotone characterization).** A scalar head $h : X \to \mathbb{R}$ *serves* a reader with loss $L_r$ if ranking by $h$ realizes $r$'s optimal selection at every budget. A single head

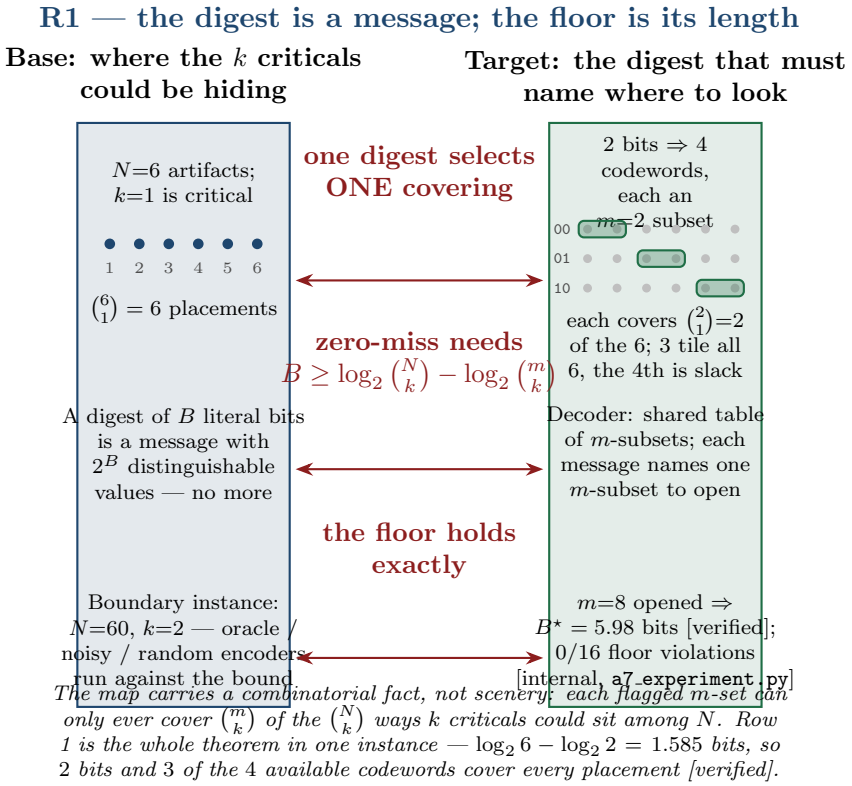



Figure 1: The relation-map for Theorem 1. A digest is a $B$-bit message selecting one of the decoder's covering $m$-subsets (base domain: a shared codebook); each covering accounts for $\binom{m}{k}$ of $\binom{N}{k}$ placements, which is the whole proof. The map is of *relations*, not surface features: from it the reader can re-derive the floor without the text.

serves two readers iff their induced preference orders are comonotone (agree on every pair). The successor and operator orders are not comonotone: a routine-but-essential working state (high continuation value, negligible regret) and an anomalously abandoned dead-end with irreversible side effects (zero continuation value, maximal regret) order oppositely. Hence no shared head serves both.

*Proof.* ($\Leftarrow$) Comonotone total preorders admit a common refinement; any utility representing the refinement serves both. ($\Rightarrow$) A shared head induces one order; each reader's optimum at every budget is a top-$m$ set of that order, so both readers' orders equal the $h$-order and hence agree pairwise. The displayed pair — constructible in any real fleet — disagrees, so no shared head exists. □

**Claim 1** (Super-additive split penalty)**.** *Let* $\mathrm{Floor}(N,k,m) = B^\star(N,k,m)$ *from Theorem 1. A joint zero-miss digest serving two readers with* disjoint *critical $k$-sets must catch their union and so needs* $\mathrm{Floor}(N,2k,m)$ *bits, and*

$$\frac{\mathrm{Floor}(N,2k,m)}{2\,\mathrm{Floor}(N,k,m)} > 1 \quad \textit{for every } N,k,m \textit{ with } 2k \le m < N,$$

*because* $\log_2\binom{N}{2k}$ *grows faster than* $2\log_2\binom{N}{k}$*. The penalty is not a fixed band: it is smallest where the critical sets are smallest and grows without bound in $k$. At $N$=60, $m$=8 it runs* 1.029 *($k$=1),* 1.067 *($k$=2),* 1.122 *($k$=3),* 1.224 *($k$=4); at $N$=60, $m$=20 it reaches* 1.392 *by $k$=10 [verified]. At the reference regime $N$=60, $k$=2, $m$=8 (Table 1): joint floor* 12.77 *bits versus* $2 \times 5.98 = 11.96$*; ratio* 2.13×*, super-additivity* 1.067 *[verified]. One stored compaction underprovisions divergent readers by* more *than a factor of two, and by increasingly more as each reader's critical set grows (Figure 4).*

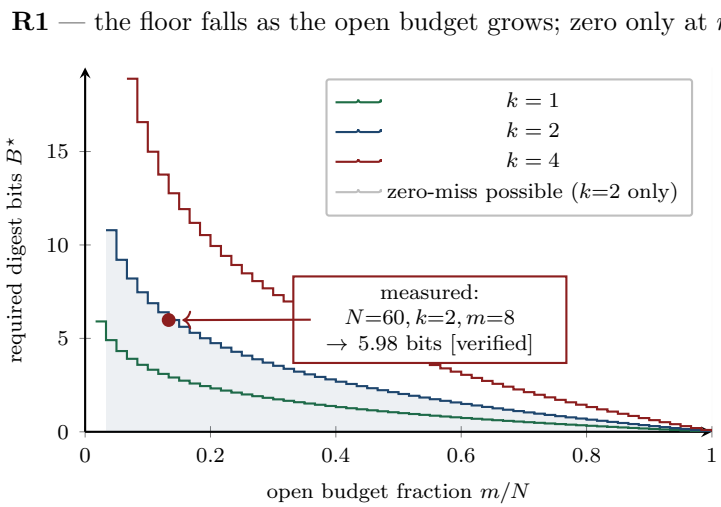



Figure 2: Regime diagram for the floor: required digest bits $B^\star$ against the open-budget fraction $m/N$ (here $N$=60, $k$=2; the marked point is the experiment's $m$=8, $B^\star$=5.98 bits [verified]). The floor is steep where budgets are tight and vanishes only when the overseer opens everything — the price of the guarantee, not of usefulness.

**What it buys.** Two compaction heads — one per reader class, compacted from durable artifacts — are a necessity of the loss structure, not an architectural taste. The misread to preempt: comonotone readers *can* share a head; the theorem forbids sharing only when the crossing pair actually occurs in the domain, and in agent oversight it does.

## 4 The regret head is derived, not designed

Which flagged items should actually interrupt a human? Fleet dashboards rank by stakes × irreversibility × anomaly — three tuned knobs. We show the product form is a theorem under one commitment, which converts "anomaly" from a vibe into a measured obligation.

**The analogy, mapped by relations.** A surfacing rule is a smoke detector (Figure 5). It has one dial, and the right setting of that dial is fixed not by how much smoke there is but by the *ratio* of two costs: the house burning down against burnt toast waking you at 3am. The relations that transfer are three. Expected loss is magnitude times probability, so the alarm must weigh what it would cost to be wrong, not merely how unusual the reading is. Raising the cost of the fire *lowers* the threshold at which the alarm should sound — the dial moves toward hair-trigger, not away from it. And a detector that cries wolf is not miscalibrated in *sensitivity*; it is miscalibrated in *probability*, because the number it reports is not the chance of a fire. That third relation is this section's thesis: once "anomaly" is required to be $\Pr(\mathsf{bad} \mid \text{evidence})$, tuning stops being the job and calibration starts. The misread to preempt: the detector analogy invites the thought that a better sensor removes the trade-off. It does not — a perfect sensor still needs the cost ratio to know where to trip, and the threshold below is a statement about costs, not about evidence quality.

**Theorem 3 (regret head).** For item $x$ let $C_{\text{miss}}(x) = \text{stakes}(x) \cdot \text{irrev}(x)$ be the unrecoverable loss of passing a bad $x$, let $a(x) = \Pr(\mathsf{bad} \mid \text{evidence}(x))$ be a *calibrated* posterior, $c_{\text{att}}$ the marginal attention cost, and $C_{\text{fa}}$ the non-attention cost of inspecting a good item. The expected unrecoverable loss of passing is exactly $C_{\text{miss}}(x)\, a(x)$, and the Bayes-optimal

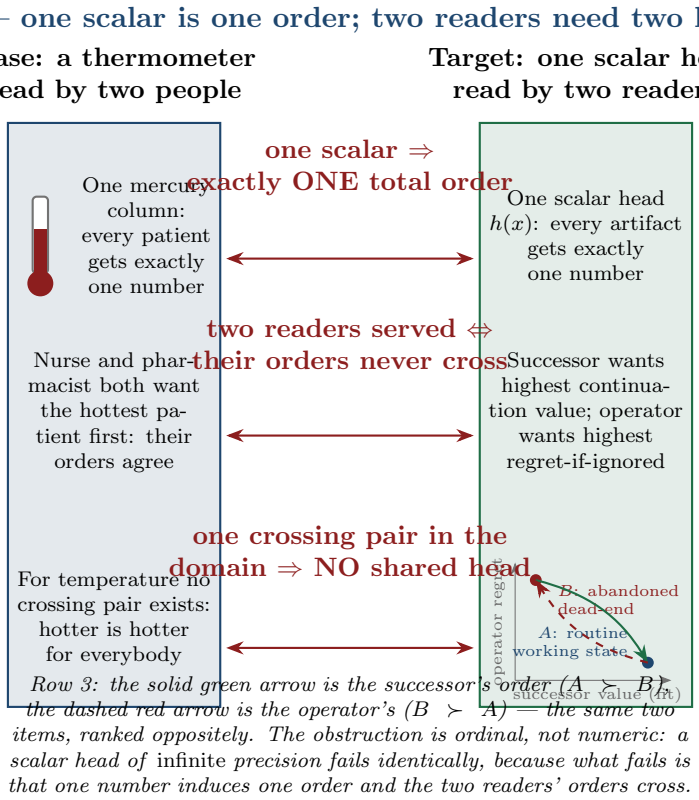



Figure 3: The relation-map for Theorem 2. The transferred relation is not that a summary resembles a temperature; it is that a single scalar induces exactly one total order, so a shared head exists iff the two readers' induced orders are comonotone. Row 3 draws the crossing pair the proof constructs — a routine working state (high continuation value, negligible regret) against an abandoned dead-end with irreversible side effects (zero continuation value, maximal regret) — which is what makes the impossibility bite in real fleets rather than merely in principle.

policy inspects iff

$$C_{\text{miss}}(x)\,a(x) \;\geq\; c_{\text{att}} + C_{\text{fa}}\big(1 - a(x)\big), \qquad \text{equivalently} \qquad \frac{a(x)}{1-a(x)} \;\geq\; \frac{c_{\text{att}} + C_{\text{fa}}}{C_{\text{miss}}(x)},$$

the classical likelihood-ratio criterion. The product form stakes × irrev × anomaly is the exact ranking key iff anomaly $= a$; any other anomaly scalar $s$ requires the correction $C_{\text{miss}} \cdot g(s)$ with $g$ the empirical calibration map fit from the audit log.

*Proof.* Expected loss of passing: $C_{\text{miss}}(x)a(x)$. Expected cost of inspecting: $c_{\text{att}} + C_{\text{fa}}(1 - a(x))$. Compare; rearrange into the odds form. □

**Numbers by hand.** $C_{\text{miss}}$=100, $c_{\text{att}}$=1, $C_{\text{fa}}$=5: inspect when $a \geq 6/105 \approx 0.057$ [verified]. Raising $C_{\text{miss}}$ to 1000 drops the bar to $6/1005 \approx 0.006$ — the direction that matters: as the miss-to-false-alarm ratio grows the threshold *falls*, driving the operator toward inspecting everything ambiguous (Figure 6); an earlier draft of this program's signal-detection figure had that direction reversed, and the likelihood-ratio statement leaves no room for the error. *Now you try the other knob:* hold $C_{\text{miss}}$=100 and raise the false-alarm cost to $C_{\text{fa}}$=20 — does the bar rise or fall, and by how much? (It rises, to $21/120 = 0.175$: attention that is expensive to waste buys back the operator's right to pass.)

**What this does not say.** The theorem does not say the operator should inspect more. It says the *threshold* moves; whether a moving threshold means more opens depends entirely on where the posterior mass sits, and an uncalibrated $a$ moves the threshold without moving the

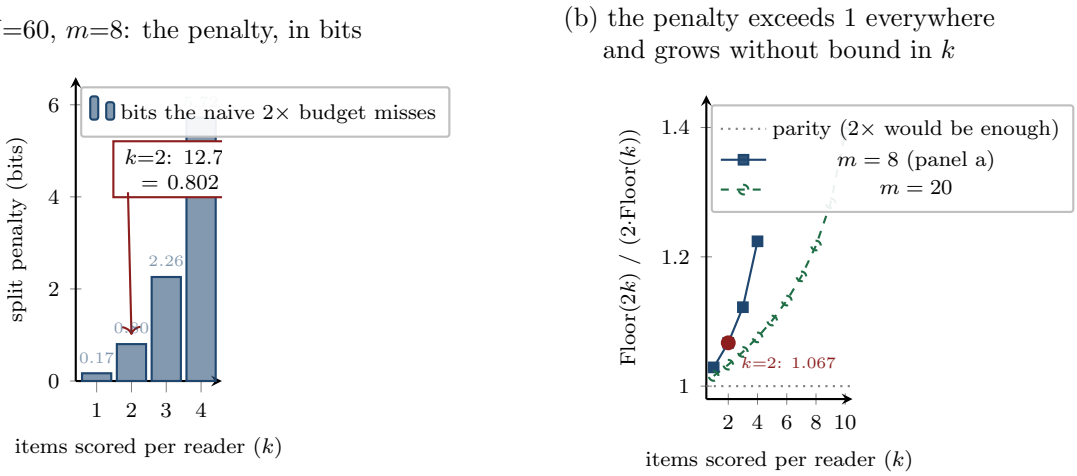



Figure 4: Regime diagram for the split penalty. Panel (a): the bits a "two summaries = two × one summary" budget is short, at $N$=60, $m$=8 — 0.80 bits at the reference $k$=2, 5.72 bits by $k$=4. Panel (b): the same penalty as a ratio over double. The finding is not a band: $\mathrm{Floor}(2k)/(2\,\mathrm{Floor}(k))$ exceeds 1 at every regime and grows without bound in $k$ — 1.067 at the reference regime, 1.224 at $k$=4 ($m$=8), 1.392 by $k$=10 ($m$=20) [verified]. Budgeting "two summaries" by doubling one summary's budget under-provisions, and under-provisions worse the more each reader must catch.

decision it was supposed to make. Nor does it license tuning: with a miscalibrated anomaly score the product form is wrong by precisely the miscalibration, so fitting $g$ from the audit log is a measured obligation and not a footnote (§7).

**Claim 2** (Reputation enters only through the posterior)**.** *If agent reputation $r$ is admitted, it enters as $a(x) = \Pr(\mathsf{bad} \mid \textit{evidence}, r)$ — never as a free multiplier on the score. High reputation lowers surfacing frequency only where the calibrated posterior actually falls, and cannot zero a term: with* $\mathrm{irrev}(x) = 1$ *and catastrophic stakes, $C_{\mathrm{miss}}(x)a(x)$ stays above threshold for any $a(x)$ bounded away from zero. This is the formal content of "reputation must not defeat hard stakes gates."*

## 5 The digest–zoom frontier, and the zoom theorem

Theorem 1 prices one corner: zero misses at a fixed open budget. The operator's real dial has three settings — digest bits, expected opens, tolerated misses — and the exchange rates among them are this section's subject. Its second half proves the theorem that was, until now, the program's one missing piece.

### 5.1 The two-constraint program and its pinned closed form

Model each artifact's critical status as i.i.d. $X \sim \mathrm{Bern}(p)$; a digest emits a per-item flag $\hat{X} \in \{0, 1\}$. Price the two failure modes separately:

$$R(\delta, f) \;=\; \min_{P_{\hat{X}|X}} I(X;\hat{X}) \quad \text{s.t.} \quad \Pr(X{=}1, \hat{X}{=}0) \le \delta, \qquad \Pr(\hat{X}{=}1) \le f.$$

This is a *custom formulation*, deliberately: classical rate–distortion charges one scalar distortion, which cannot express that oversight prices a missed critical differently from a wasted open. An August-2026 literature sweep found no prior false-negative-constrained lossy compression for

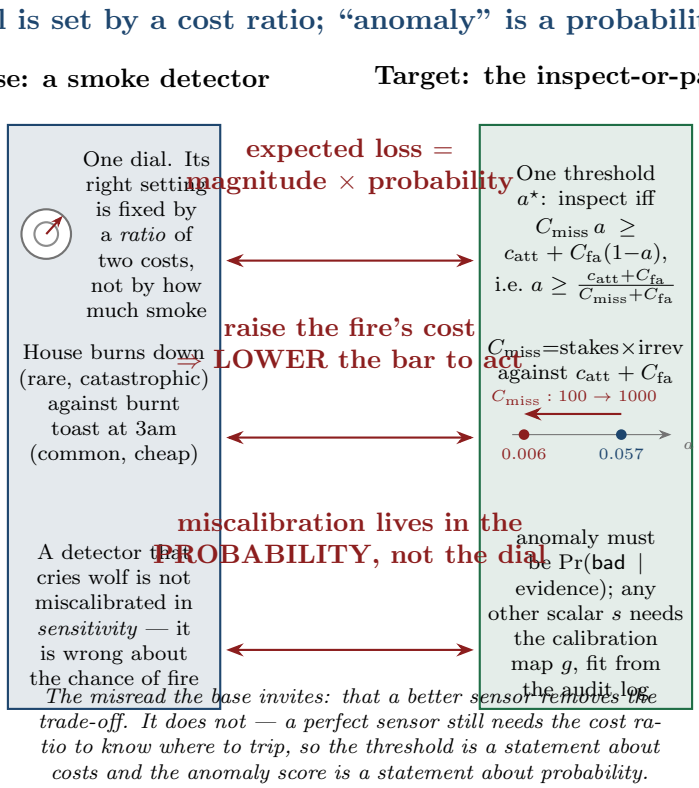



Figure 5: The relation-map for Theorem 3. Three relations transfer, and they are the section in order: expected loss is magnitude times probability, so the rule must weigh what being wrong costs and not merely how unusual the reading is; raising the miss cost *lowers* the threshold at which to act (drawn in row 2, $a^\star$ moving from 0.057 to 0.006 as $C_{\text{miss}}$ goes $100 \to 1000$ [verified]); and a rule that surfaces too much is miscalibrated in probability rather than over-sensitive, which is why anomaly = Pr(bad | evidence) is an obligation and not a naming convention.

binary sources under this framing; the nearest neighbors are expansion coding with one-sided error for continuous sources [6] and the rate–distortion–classification line for Bernoulli sources [7]. We position $R(\delta, f)$ against that literature as a formulation, not as a textbook import — and "not found" is not "proven nonexistent." What the second constraint changes, and where it collapses back to Shannon, is Figure 7.

**Proposition (pinned joint).** For $X \sim \text{Bern}(p)$ with $0 \le \delta < p$ and $p - \delta \le f < 1 - \delta$, if $f$ is below the flag rate at which an $X$-independent flagger meets the miss constraint (i.e. $f < 1 - \delta/p$), both constraints bind at the optimum and the minimizing joint is fully pinned:

$$q_{11} = p - \delta, \qquad q_{10} = \delta, \qquad q_{01} = f - (p - \delta), \qquad q_{00} = 1 - f - \delta,$$

feasible iff all four are nonnegative, and $R(\delta, f) = I(X; \hat{X})$ evaluated at this joint. At the zero-miss tightest-flag corner, $R(0, f{\to}p) = H(p)$: the digest must spend the source's full entropy.

*Proof.* $I(X; \hat{X})$ is convex in $P_{\hat{X}|X}$ and decreases along the feasible ray toward the independent coupling ($I = 0$ iff independent). An independent flagger $\hat{X} \sim \text{Bern}(\phi)$ has miss mass $p(1 - \phi)$, so it meets the constraint only when $\phi \ge 1 - \delta/p$; when $f < 1 - \delta/p$ that coupling is outside the flag budget, so the minimum sits on the boundary where any

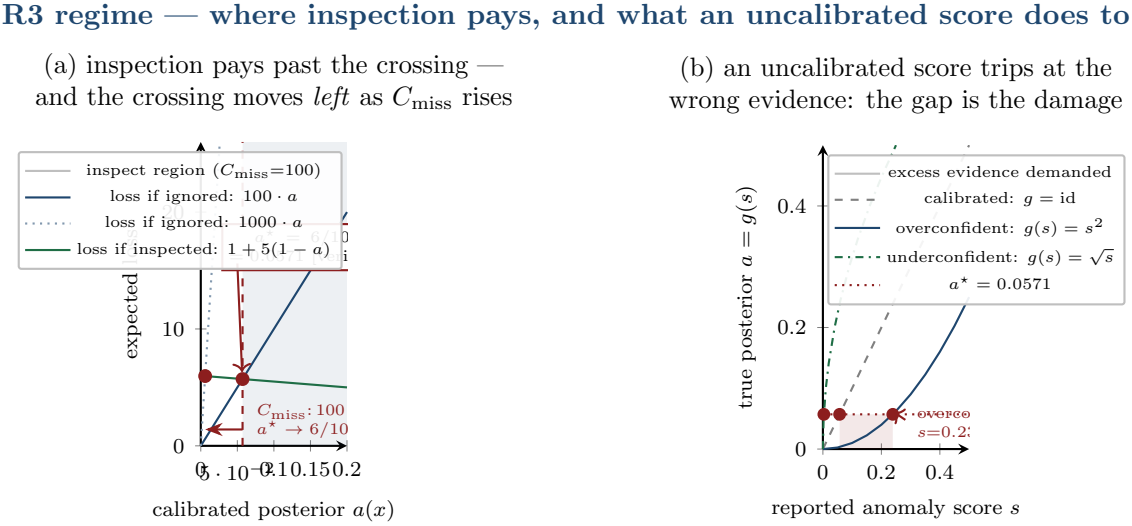



Figure 6: Regime diagram for the surfacing rule. Panel (a): expected loss of pass against expected loss of inspect, as a function of the calibrated posterior $a(x)$. Inspection pays exactly where the two loss lines cross, at $a^\star = (c_{\text{att}} + C_{\text{fa}})/(C_{\text{miss}} + C_{\text{fa}})$ — the closed form obtained by solving Theorem 3's inequality directly. At the shown costs ($C_{\text{miss}}$=100, $c_{\text{att}}$=1, $C_{\text{fa}}$=5) that is $6/105 \approx 0.057$; raising $C_{\text{miss}}$ to 1000 moves the crossing *left*, to $6/1005 \approx 0.006$ [verified] — the counterintuitive direction, drawn rather than asserted. Panel (b): what the theorem's calibration hypothesis is worth. If the reported score $s$ is thresholded at $a^\star$ but the true posterior is $g(s) \neq s$, the rule trips at the wrong evidence: an overconfident scorer ($g(s) = s^2$) demands $s$=0.239 before acting where 0.057 was warranted, and the shaded horizontal gap is the excess evidence the operator is silently made to wait for.

remaining slack in either constraint could otherwise be spent moving further along the ray — both constraints bind. Binding constraints fix $q_{10} = \delta$ and $q_{11} + q_{01} = f$; with the source margin $q_{10} + q_{11} = p$ this pins all four cells. At $(\delta, f) = (0, p)$ the joint is diagonal, $\hat{X} = X$, and $I = H(p)$. □

**Numbers** ($p = 0.05$): $R(0, 0.05) = H(0.05) = 0.286$ bits/symbol [verified, Cover–Thomas [5]]; $R(0, 0.10) = 0.186$; $R(0.01, 0.10) = 0.110$; $R(0.04, 0.06) = 0.009$ [internal, `b1_frontier.py`]. Reading: loosening the open budget cheapens the digest; tolerating misses cheapens it dramatically — two orders of magnitude between the guarantee corner and $(\delta, f) = (0.04, 0.06)$ (Figure 8 shows the whole surface these four numbers sit on). Consistency with Theorem 1: at $N$=1000, $R \cdot N \approx 186$ bits at $(0, 2p)$, below the worst-case covering floor $\log_2 \binom{1000}{50} \approx 282$ bits — average-case rate under the worst-case floor, as it must be. One geometric honesty note, corrected here: the pinned closed form holds only while $f < 1 - \delta/p$, the Proposition's stated regime where both constraints bind; past that line an $X$-independent flagger already meets the miss budget on its own and the true rate is exactly 0, not rising. An earlier plot that evaluated the pinned formula past this line showed a spurious uptick near $\delta/p \to 1$ — a domain-of-validity error, not a feature of the true frontier, and we flag it with the same audit-the-meter discipline as §2's wrong turn.

### 5.2 The zoom-advantage theorem

A conservative digest over-flags to be safe: it emits a flagged set of size $F$ containing the $k$ true criticals with $k \ll F$. The naive second stage opens all $F$. The adaptive second stage plays twenty questions with group queries (Figure 9): open a *group* of flagged items at once — an aggregate check cheap at group level, such as a combined diff scan or batched test run —

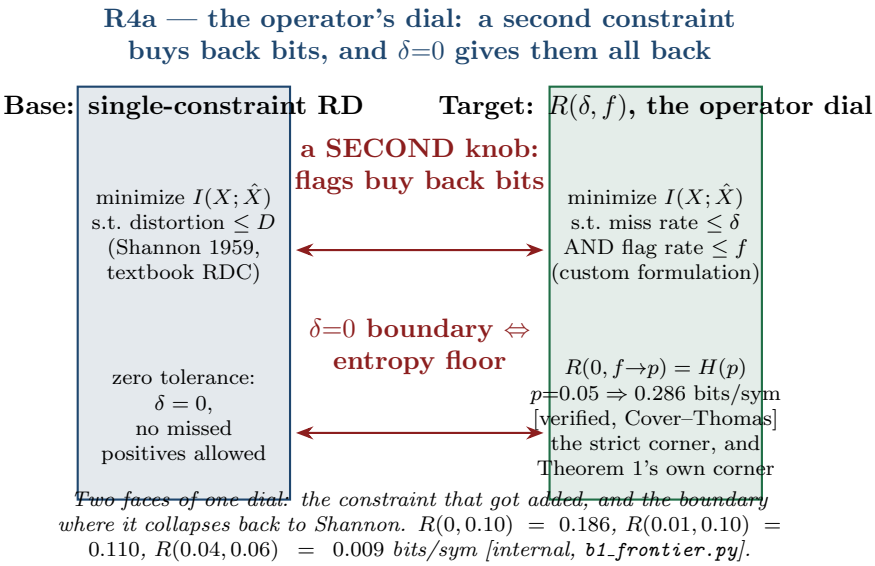



Figure 7: The relation-map for the two-constraint program. The relation that transfers from classical rate–distortion is not the objective — that is unchanged, $I(X;\hat{X})$ — but the shape of the feasible set: pricing false negatives separately from flag rate adds a second binding constraint, and every bit the digest saves is bought by loosening one of the two. Row 2 is the seam with Theorem 1: at $\delta{=}0$ and $f \to p$ the program returns the source's entire entropy, which is the frontier's own version of the covering floor.

learn whether the group contains any critical, and split only the groups that do. The analogy transfers because its critical relation transfers: halving a set costs one question and eliminates half the candidates *only when positives are rare in the set.* The misread to preempt: twenty questions finds *one* answer in $\log_2 F$ queries; with $k$ answers the tree forks, and the accounting below is precisely the price of the forking.

**Theorem 4 (zoom advantage; new).** Let a flagged set of $F$ items contain an unknown set $P$ of $k \geq 1$ positives, and let a group query on $S$ report whether $S \cap P = \emptyset$. Adaptive halving — query the flagged set; discard any group that tests negative; report any positive singleton; otherwise split the group into halves of sizes $\lfloor |S|/2 \rfloor$, $\lceil |S|/2 \rceil$ and recurse on both — identifies $P$ exactly, using a total number of group queries

$$Q \;\leq\; 2k\,\lceil \log_2(F/k) \rceil + 4k \qquad (\text{and always } Q \leq 2F-1),$$

for every placement of the $k$ positives, versus $F$ opens for flat inspection. The leading constant is tight: for $F$ and $k$ powers of two, evenly spread positives force $Q = 2k\log_2(F/k) + 2k - 1$.

*Proof.* Every group the algorithm ever queries is a node of the *halving tree* $T$ of $[F]$: the root is the full flagged set, each internal node's children are its two halves, leaves are singletons. Since a node of size $s$ has children of size at most $\lceil s/2 \rceil$, every node at depth $d$ has size at most $\lceil F/2^d \rceil$, so $T$ has depth $D \leq \lceil \log_2 F \rceil$, and $T$ has exactly $2F-1$ nodes (a binary tree with $F$ leaves), giving the unconditional bound.

Call a node *positive* if its group meets $P$. The algorithm splits exactly the positive non-singleton nodes it visits, and every queried node is either the root or a child of a split node. Hence, writing $S^+$ for the number of positive non-singleton nodes of $T$,

$$Q \;\leq\; 1 + 2S^+.$$

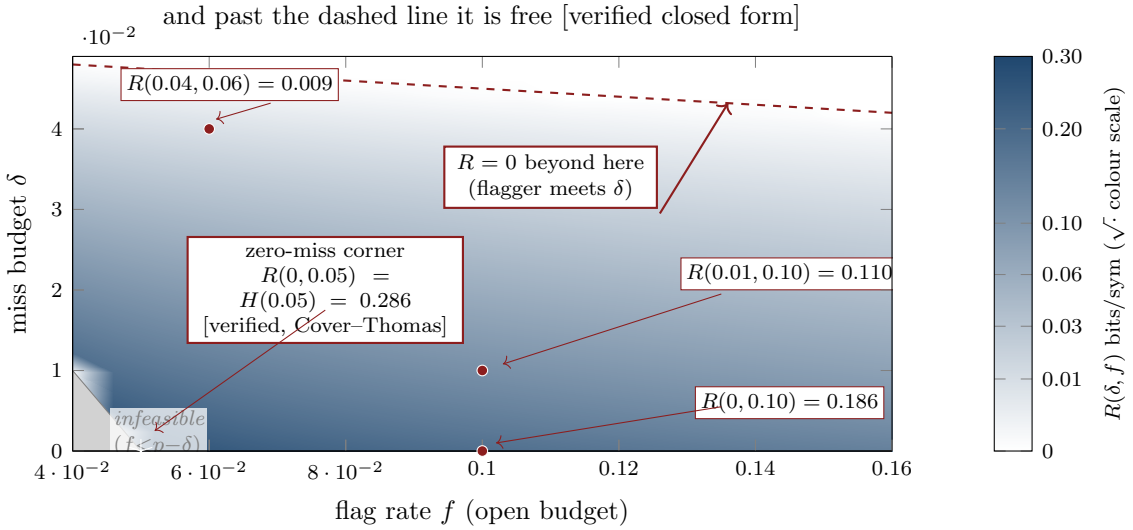



Figure 8: The two-constraint frontier $R(\delta, f)$ over the whole $(f, \delta)$ plane near the guarantee corner, from the Proposition's closed form (same formula as `b1_frontier.py`). The four numbers of the "Numbers" paragraph are marked directly on the surface, including the zero-miss corner that ties this section back to Theorem 1. The dashed line is the Proposition's own boundary $f = 1 - \delta/p$: below it both constraints bind and the pinned joint applies; above it an $X$-independent flagger already meets the miss budget and $R = 0$ exactly, not rising — the region the corrected geometric-honesty note above refers to. The small grey wedge is simply infeasible ($f < p - \delta$: not enough flag budget to cover the source rate at all).

It remains to bound $S^+$ by a charging argument. The nodes at any fixed depth $d$ of $T$ are pairwise disjoint subsets of $[F]$, so each positive item lies in at most one of them; therefore at most $\min(k, 2^d)$ nodes at depth $d$ are positive. Charge the positive nodes in two regimes:

- *Top of the tree* ($d \le \lfloor \log_2 k \rfloor$): charge each positive node to its level. The total over these levels is at most $\sum_{d=0}^{\lfloor \log_2 k \rfloor} 2^d \le 2k - 1$.
- *Below the fan-out* ($d > \lfloor \log_2 k \rfloor$): charge each positive node to one positive item it contains. By disjointness within a level, each of the $k$ positives absorbs at most one charge per level. Non-singleton nodes exist only at depths $d \le D - 1$, so the number of charging levels is

$$D - 1 - \lfloor \log_2 k \rfloor \;<\; (\log_2 F + 1) - 1 - (\log_2 k - 1) \;=\; \log_2(F/k) + 1$$

(using $D \le \lceil \log_2 F \rceil \le \log_2 F + 1$ and the strict $\lfloor x \rfloor > x - 1$); an integer strictly below $\log_2(F/k)+1$ is at most $\lceil \log_2(F/k) \rceil$, so these levels contribute at most $k\lceil \log_2(F/k) \rceil$ charges.

Summing, $S^+ \le k\lceil \log_2(F/k) \rceil + 2k - 1$, so $Q \le 1 + 2S^+ \le 2k\lceil \log_2(F/k) \rceil + 4k - 1$, which gives the bound. Correctness is immediate: a discarded group is certified free of positives, and every positive survives discarding at every level, so each is eventually isolated and reported as a positive singleton.

For tightness, take $F = 2^a$, $k = 2^b$, $b \le a$, positives at positions $\{iF/k : 0 \le i < k\}$. Every node at depth $d \le b$ contains a positive ($2^{b-d}$ of them), so all $2^{b+1} - 1 = 2k - 1$ nodes down to depth $b$ are queried and split (or, at depth $b$, split further); below depth $b$, exactly $k$ nodes per level are positive, each is split at levels $b, \ldots, a-1$, and each split queries two children. Counting queried nodes: all $2k - 1$ nodes at depths $0..b$, plus $2k$ children per level for levels $b+1, \ldots, a$, of which per level $k$ are positive; total queried $= (2k - 1) + 2k(a - b) = 2k \log_2(F/k) + 2k - 1$. This configuration is realized numerically below. □

*Corollary* 1 (The advantage, priced). Flat inspection opens $F$; adaptive halving opens at most $2k\lceil\log_2(F/k)\rceil + 4k$, so the guaranteed advantage is at least $F/(2k\lceil\log_2(F/k)\rceil + 4k)$ — at $(F, k) = (2500, 10)$, at least $2500/200 = 12.5\times$, against a measured mean of $15.3\times$ [internal, `b1_frontier.py`]. Written in the flagged-set density $d = k/F$ the advantage depends on nothing else:

$$\mathrm{adv}(d) \;=\; \frac{1}{d\big(2\lceil\log_2(1/d)\rceil + 4\big)},$$

which exceeds 1 exactly on $d \le 1/12$ (the step function $d(2\lceil\log_2(1/d)\rceil + 4)$ first touches 1.0000 at $d = 1/12$) [verified]. The worked point above is $d$=0.004 — a far sparser flagged set than §5.1's $p$=0.05 frontier can deliver, which is the composition window priced below.

**Positioning against group testing.** None of the group-testing machinery here is new mathematics: Dorfman introduced pooled testing in 1943 [1], and Hwang's generalized binary splitting [2] finds $k$ defectives in $\log_2\binom{F}{k} + O(k)$ tests — essentially matching the counting lower bound $\lceil\log_2\binom{F}{k}\rceil \ge k\log_2(F/k)$ that any adaptive strategy with binary outcomes must pay, with the full theory in Du and Hwang [3]. Plain halving pays a factor $2 + o(1)$ over Hwang because it re-queries both children of every split rather than inferring one subgroup and binary-searching within the other. Theorem 4's contribution is therefore *not* a better group-testing algorithm; it is the explicit worst-case constant for the exact procedure the digest architecture runs, proved tight — which converts the simulation's observed $\approx 2\times$ gap over the ideal $k\log_2(F/k)$ from an empirical curiosity into the algorithm's provable constant, and makes the frontier's second-stage cost a theorem rather than a fit.

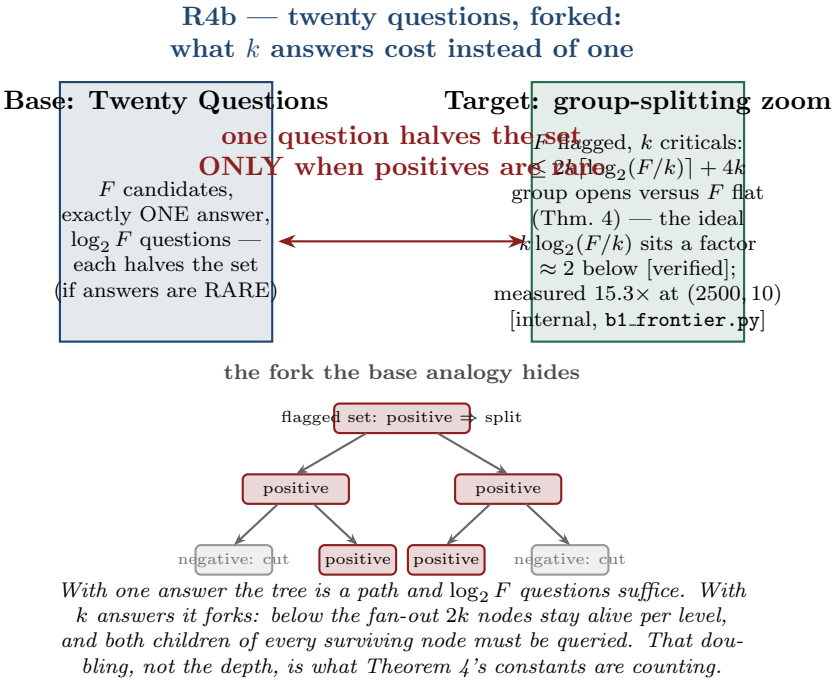
**Numbers, verified.** A dedicated check (seed 20260816) re-runs the exact halving procedure of `b1_frontier.py`: at $F$=2500, $k$=10, mean 162.6 opens over 150 random placements ($15.4\times$ advantage, reproducing the compendium's $15.3\times$; ideal $k\log_2(F/k) = 79.7$; empirical overhead $2.04\times$), maximum observed 183, all under the bound $2 \cdot 10 \cdot 8 + 40 = 200$; random sweeps over $F \in [16, 5000]$, $k \in [1, 50]$ and adversarial evenly-spread and clustered placements never exceed the bound; and the tightness configuration $F$=4096, $k$=32 lands exactly on $2k\log_2(F/k) + 2k - 1 = 511$ [internal, `zoom_bound_check.py`]. Now you try the $k = 1$ sanity case by hand: $F = 1024$, one positive — halving queries two children per level for ten levels plus the root, 21 opens, under the bound $2 \cdot 10 + 4 = 24$, versus 1024 flat.

**What each stage buys, priced separately.** Together, §5.1 and Theorem 4 give the operator a three-way dial. Stage one, the digest, spends $R(\delta, f) \cdot N$ bits to cut $N$ artifacts to $F \approx fN$ flags at miss budget $\delta$; stage two, zoom, spends at most $2k\lceil\log_2(F/k)\rceil + 4k$ group opens to isolate the $k$ criticals among the flags. Each stage's own price is exact: loosening flags from $f$=$p$ to $f$=$2p$ cuts the digest from 0.286 bits/symbol [verified] to 0.186 [internal] at zero miss, and tolerating $\delta$=0.01 misses cuts the rate almost in half again. The stages also close the loop with Theorem 1: the worst-case covering floor is the frontier's ($\delta$=0, $f \to p$) corner.

**The composition window, stated.** The two prices are drawn from one budget, and they compose only inside a window this paper can state exactly. Stage one sets stage two's density: by the Proposition both constraints bind, so the flagged set holds $(p - \delta)N$ real criticals among $fN$ flags, and

$$d \;=\; \frac{k}{F} \;=\; \frac{p - \delta}{f}, \qquad \text{guaranteed advantage} \;=\; \mathrm{adv}(d) \;=\; \frac{1}{d\big(2\lceil\log_2(1/d)\rceil + 4\big)}.$$

Since $\mathrm{adv}(d) > 1$ exactly on $d \le 1/12$, *the two stages compose profitably if and only if the flag budget is at least twelve times the residual positive rate*, $f \ge 12(p - \delta)$ [verified from the two closed forms]; the measured mean crossover is about twice as forgiving, $d \approx 0.165$, i.e. $f \gtrsim 6(p - \delta)$ [internal, `zoom_bound_check.py`]. At $p$=0.05, $\delta$=0 the window opens only at $f \ge 0.60$ — the digest must flag 60% of the night's artifacts before the second stage is guaranteed to beat simply reading the flags. §5.1's recommended loosening does not reach it: at $f$=$2p$ the density is 0.500 and the guarantee is $0.33\times$ (a $3\times$ loss, not a saving); at $(\delta, f) = (0.01, 0.10)$ the density is 0.400




*With one answer the tree is a path and $\log_2 F$ questions suffice. With $k$ answers it forks: below the fan-out $2k$ nodes stay alive per level, and both children of every surviving node must be queried. That doubling, not the depth, is what Theorem 4's constants are counting.*

Figure 9: The relation-map for zoom: twenty questions (base) against group-split inspection of the flagged set (target). The correspondence that does the work is "one question halves the candidate set *when answers are rare*"; the misread it invites — that $\log_2 F$ questions suffice — is exactly what Theorem 4's $2k\lceil\log_2(F/k)\rceil + 4k$ accounting corrects for $k > 1$, and the drawn fork is where the extra factor comes from. Note the target cell states the *deployed* bound: the ideal $k\log_2(F/k)$ is a factor $\approx 2$ cheaper and is not what the digest architecture runs.

and the guarantee is 0.31×. *Inside* the window the flags stage one adds are cheap in the second stage — $k$ is fixed by the source, so doubling $F$ costs only the logarithmic term, $2k$ further opens. *Outside* it, the honest second-stage cost is linear in $F$, because reading all $F$ flags is then the better procedure. The Corollary's 12.5× belongs to $d$=0.004, two-and-a-half orders of magnitude sparser than anything the $p$=0.05 frontier produces: a source with $p \approx 0.004$ and a generous flag budget. Stated plainly: cheap digests and cheap zooms are bought from the same budget, and the operating points where both are cheap are exactly the sparse-source ones.

**The dense-flag boundary, stated.** At the other end the advantage inverts outright. When the flagged set is already dense in positives ($k/F$ large — e.g. a flag budget barely above the residual positive rate, $f \approx 1.11(p-\delta)$, so nine flags in ten are real), group overhead dominates and adaptive halving opens *more* than $F$: at $F$=100, $k$=90, density $d$=0.9, the run costs 199 opens against 100 flat [internal] — exactly Theorem 4's unconditional $2F-1$, reached because at that density essentially every node of the halving tree is positive. The regime where zoom pays is exactly the sparse-flagged corner $d \leq 1/12$ (Figure 10). A miss-averse digest moves toward that corner by over-flagging, but only arrives if it over-flags by at least a factor of twelve; the modest loosening §5.1 prices does not get there. The theorem supports the aggressively over-flagging operating point, and honestly does not support the others.

## 6 Related work

**Group testing.** Dorfman [1], Hwang's generalized binary splitting [2], and the Du–Hwang monograph [3] supply the second stage's mathematical substrate; §5.2 states precisely what is imported (the algorithmic idea and the counting lower bound) and what is contributed (the constant-explicit worst-case bound for plain halving, its tightness, and its role as the frontier's

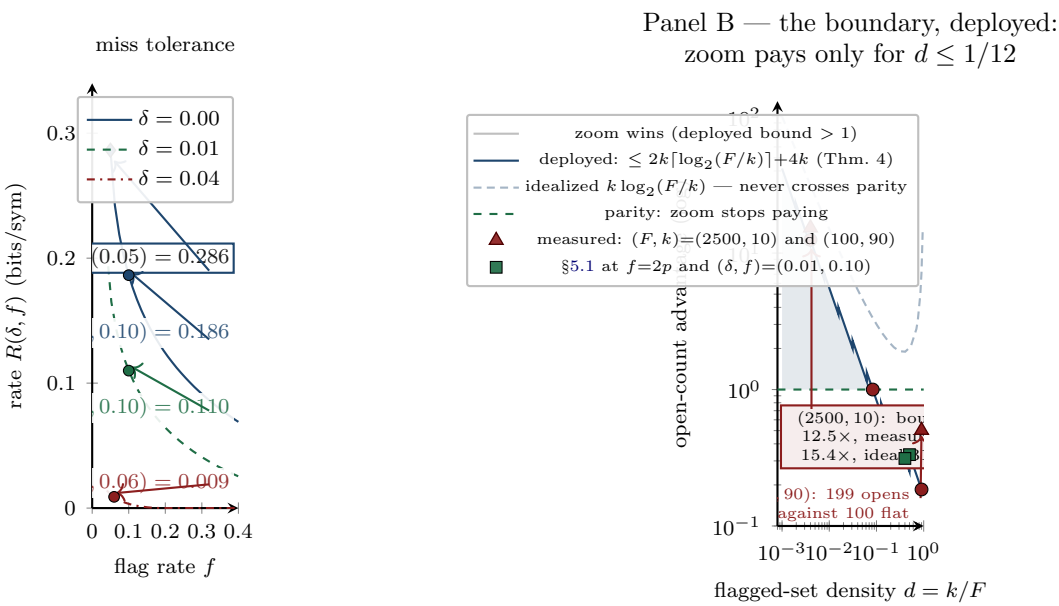



Figure 10: Regime diagram for the zoom advantage: the idealized open-count advantage $F/(k\log_2(F/k))$ against flagged-set density $k/F$, over the sparse-to-moderate range this digest architecture targets. The advantage is steep where flags are sparse ($k \ll F$) and falls toward parity as density grows; the plotted idealization never crosses parity, but the deployed algorithm's constant overhead does invert the ordering at higher density still, exactly as measured at $(F, k) = (100, 90)$ in §5.2 (199 adaptive opens against 100 flat) — both halves of the claim are critical.

second-stage cost). **Rate–distortion.** The single-constraint theory is classical [4, 5]. The two-constraint $R(\delta, f)$ is a custom formulation for oversight; the nearest published lines are expansion coding with one-sided error for continuous sources [6] and rate–distortion–classification theory for Bernoulli sources [7], neither of which prices false negatives as a separate constraint for binary sources. We claim the formulation and its pinned closed form as this paper's framing contribution, subject to the standard caveat that a survey's "not found" is defeasible. **Attention and oversight.** That attention is the scarce resource an information-rich environment consumes is Simon's observation [8]; competitive analysis of paging [9] inspires the budgeted-opens framing. On the AI-oversight side, debate and related protocols [10] and delegated-computation results [11] ask how a weak verifier can check strong provers; this paper is complementary — it prices the *interface* (the digest) rather than the protocol, and its floors apply to any oversight scheme whose human reads a bounded summary. **Signal detection.** Theorem 3 is a Bayes-risk likelihood-ratio test, textbook decision theory; the contribution is the exact identification of the folklore product form with it and the resulting calibration obligation, which also caught and corrected a reversed-direction figure in an earlier draft.

## 7 Limitations and honest boundaries

**Honest boundary** — stated at the same prominence as the claims.

- *Theorem 1* lower-bounds the *zero-miss guarantee* in a worst-case combinatorial model with a data-independent decoder; it says nothing against cheaper digests at nonzero miss (that is §5.1's job) and nothing average-case.

- *Theorem 2* forbids a shared head only for non-comonotone readers; comonotone readers can share. The impossibility bites because the crossing pair (routine-essential state vs. abandoned irreversible experiment) actually occurs in fleets.
- *Theorem 3* is exactly as good as its calibration: with an uncalibrated anomaly score the product form is wrong by precisely the miscalibration, and fitting $g$ from the audit log is a measured obligation, not a footnote.
- *The frontier* $R(\delta, f)$ is an i.i.d. Bernoulli model of critical status — a modeling commitment, not a law; and the formulation is custom, positioned against [6, 7], not a textbook theorem.
- *Theorem 4* assumes a reliable group query at unit cost; correlated or noisy group checks, and group costs growing with $|S|$, change the accounting. The advantage inverts in the dense-flag regime, as stated and measured, and it composes with §5.1's frontier only for $f \geq 12(p-\delta)$ — at $p$=0.05 that is $f \geq 0.60$, well outside the flag budgets §5.1 tabulates, so the two halves of §5 are not jointly instantiated at the same source rate anywhere in this paper. Plain halving is a factor $\approx 2$ from Hwang's optimum by design — we bound the deployed algorithm, not the best one.
- *Falsification scope:* the 0/16 result tests the floor's encoder/decoder formalization at one swept regime ($N$=60, $k$=2, $m$=8); the noisy-encoder curve in that experiment barely beats random at SNR $\approx 1$ and must not be read as evidence noisy digests are near-optimal. The feature-corruption panel of the same experiment is visually compressed at $m$=8 (the budget is generous enough that corruption barely moves in-range miss) and would need a tighter-$m$ rerun for a publication-grade figure.

## 8 Reproducibility

Every [internal] number regenerates deterministically at seed `20260816` from self-contained scripts: `a7_experiment.py` (the falsification experiment; 0/16, with the wrong-turn v1 and its 8/14 spurious violations retained under `wrong-turns/`) and `b1_frontier.py` (the $R(\delta, f)$ table and the zoom simulation), both bundled with the research program's results skill, plus this paper's companion check `zoom_bound_check.py` (Theorem 4's bound against random and adversarial placements, the exact tightness configuration, and the dense-flag boundary). Every [verified] number is recomputable from the stated closed forms with a hand calculator. Figures regenerate from `figures/src/` at the same seed.

## References


[1] R. Dorfman. The detection of defective members of large populations. *Annals of Mathematical Statistics*, 14(4):436–440, 1943.

[2] F. K. Hwang. A method for detecting all defective members in a population by group testing. *Journal of the American Statistical Association*, 67(339):605–608, 1972.

[3] D.-Z. Du and F. K. Hwang. *Combinatorial Group Testing and Its Applications*, 2nd ed. World Scientific, 2000.

[4] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record*, part 4, 142–163, 1959.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*, 2nd ed. Wiley, 2006.

[6] H. Si, O. O. Koyluoglu, and S. Vishwanath. Lossy compression of exponential and Laplacian sources using expansion coding. arXiv:1308.2338, 2013.

[7] Rate-distortion-classification representation theory for Bernoulli sources. arXiv:2601.11919, 2026; see also the universal rate-distortion-classification line, IEEE ITW 2025.

[8] H. A. Simon. Designing organizations for an information-rich world. In M. Greenberger, ed., *Computers, Communications, and the Public Interest*. Johns Hopkins Press, 1971.

[9] D. D. Sleator and R. E. Tarjan. Amortized efficiency of list update and paging rules. *Communications of the ACM*, 28(2):202–208, 1985.

[10] G. Irving, P. Christiano, and D. Amodei. AI safety via debate. arXiv:1805.00899, 2018.

[11] S. Goldwasser, Y. T. Kalai, and G. N. Rothblum. Delegating computation: interactive proofs for muggles. In *STOC 2008*, 113–122.