BYZANTINE-ROBUST FEDERATED RAG VIA ALIGNED CALIBRATION AND FIXED-MEMBERSHIP CONFORMAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models answer questions more accurately when they can consult relevant documents, an approach called retrieval-augmented generation (RAG). Many valuable collections, such as medical records or company files, cannot be pooled in one place because of privacy rules or ownership. Federated RAG leaves each collection with its owner: each owner, or *node*, searches its own documents and scores candidate answers, and a central hub combines the scores. Because the hub cannot inspect the nodes, some of them, called Byzantine, may be compromised, faulty, or misled by malicious instructions hidden in documents, and report arbitrary scores. We want the hub to return a set of candidate answers that contains the correct one with a chosen probability, the guarantee offered by conformal prediction. Conformal prediction keeps every answer whose score passes a cutoff, and the hub sets that cutoff in a calibration step, using questions with known answers. An unknown group of nodes, no larger than a declared bound, may misreport both in this step and when new questions are answered. Existing methods assume every node is honest or protect only the calibration step. Our method rests on a simple observation: the honest nodes are the same in both steps. The hub therefore has all nodes score the same calibration questions. It keeps a candidate answer only if some plausible group of honest nodes, using its own scores in both steps, would keep it. We prove three guarantees. The resulting sets contain the correct answer with the chosen probability in finite samples, whatever the Byzantine nodes report. No method using the same information can return smaller sets without risking the loss of an answer the honest nodes support. If nodes fail at random, the guarantee weakens only by the probability that more nodes fail than declared. We tested the method in simulations and on real question-answering tasks, including medical exams. We also used a panel of language models as nodes, some of them hijacked by injected instructions. Our answer sets reached the target probability whenever no more nodes misbehaved than declared. They were also clearly smaller than those of simpler methods offering the same protection, most of all when the declared bound was generous. Simply averaging the nodes' scores could miss the target. In practice, an operator can therefore declare a cautious bound on the number of bad nodes without paying much for it in the size of the answer sets.

1 INTRODUCTION

Many of the document collections that could make language models more reliable cannot be pooled in one place. Retrieval-augmented generation (RAG) lets a language model answer a question using documents retrieved for it (Lewis et al., 2020). Yet valuable collections, such as the records of different hospitals or the files of different firms, are often kept apart by privacy rules or ownership. Federated RAG leaves each collection with its owner (Chakraborty et al., 2025; He et al., 2025; Dhasade et al., 2026): each owner, or *node*, retrieves from its own collection and scores candidate answers, and a central hub combines the scores. For its output to be trusted, the hub should return a set of answers that contains the correct one with a chosen probability. Conformal prediction provides such sets (Vovk et al., 2005; Angelopoulos & Bates, 2023). It sets a cutoff from the scores of the

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.

correct answers to *calibration questions*, whose answers are known, and keeps every candidate scoring at or below it (a lower score means a more plausible answer).

The challenge is that the hub cannot inspect the nodes, so some may misreport, whether compromised, faulty, or misled by instructions planted in retrieved text (*prompt injection*). Consider "Which dry material conducts electricity: copper, wood, glass, or rubber?" sent to 16 nodes with separate corpus shards. At most two nodes may report arbitrary scores, both when the cutoff is calibrated (the *calibration phase*) and when the question is answered (the *query phase*). Such nodes are *Byzantine*, the rest are honest, and the hub does not know which are which. The natural baseline, a plain average of all nodes' reported scores (their *reports*), is vulnerable. Small calibration reports lower the cutoff and large query reports push the correct answer out; in our experiments this makes the set miss the correct answer far more often than the target allows (Figure 3a in Section 5). Sending scores with few bits to save communication also perturbs even honest scores by rounding.

We cast the problem as split conformal prediction with contaminated inputs (Section 2). If the hub knew which nodes were honest and saw their exact scores, it could average those scores and apply standard split conformal prediction; we call the resulting set the *honest-mean oracle*. The hub cannot compute it, but it knows a declared *budget* A: an upper bound, fixed in advance, on the number of Byzantine nodes. We seek a prediction set that meets three requirements. (i) It contains the correct answer with probability at least $1 - \alpha$, whatever the Byzantine nodes report. (ii) It contains every answer the oracle might contain, given what the reports reveal. (iii) It is otherwise as small as possible. Requirement (ii) is stronger than (i). Coverage holds only on average over questions, so a set can meet it while dropping answers the honest nodes support, as does discarding nodes that look suspicious (Section 5).

Existing methods guarantee parts of this problem but not the whole. TRAQ and C-RAG certify answer sets or risk for RAG (Li et al., 2024; Kang et al., 2024a), and federated conformal methods assume honest reporting (Lu et al., 2023; Humbert et al., 2023; Zhu et al., 2024). Rob-FCP and PRISM-FCP defend calibration but not query scores (Kang et al., 2024b; Lari et al., 2026b). FC-RAG calibrates each node on its own questions and assumes the resulting summaries approximate the all-node mean's cutoff (Dubey & Huo, 2026b). Robust conformal methods certify sets against bounded perturbations or data poisoning (Gendler et al., 2022; Zargarbashi et al., 2024; Scholten & Günnemann, 2025). None says which answers to keep when one unknown honest set is shared by every calibration question and the query (Appendix A). Together, these methods leave three gaps. First, most assume honest reporting. Second, current defenses protect only calibration or bound each score separately, though nodes can misreport in both phases. Third, calibrating each node separately, as FC-RAG does, cannot determine the honest-mean oracle's cutoff (Section 3.1).

We propose *fixed-set* inference, which exploits the fact that the same nodes are honest in the calibration phase and in the query phase. All nodes score the same calibration questions. With K nodes in total, any set of at least $K - A$ of them could be the honest group; call it a *possible honest group*. Such a group *supports* a candidate answer if the group's average query score for it falls at or below the cutoff computed from that same group's calibration scores, allowing for known rounding error. Fixed-set keeps exactly the candidates that some possible honest group supports. It closes each gap in turn: its coverage holds whatever up to A nodes report, one possible honest group must support a candidate in both phases, and shared calibration questions retain how the nodes' scores vary together. Rules that let the group change between the two phases can keep answers that no single honest group supports, such as wood in Figure 1. Unlike robust conformal methods that bound each score separately, fixed-set holds one honest group fixed throughout. Section 3 describes the method and two simpler relaxations of it.

The method brings three practical benefits, which are our contributions.

1. *Reliable answers without trusting any single node.* The answer set contains the correct answer with the chosen probability whatever the misbehaving nodes report, provided no more of them misbehave than declared. If nodes instead fail at random, the guarantee drops only by the chance that too many fail, which an operator can compute and advertise in advance (Section 4). Plain averaging offers no such protection: in our experiments it misses the target both under constructed attacks and when language models are hijacked by injected instructions (Section 5).

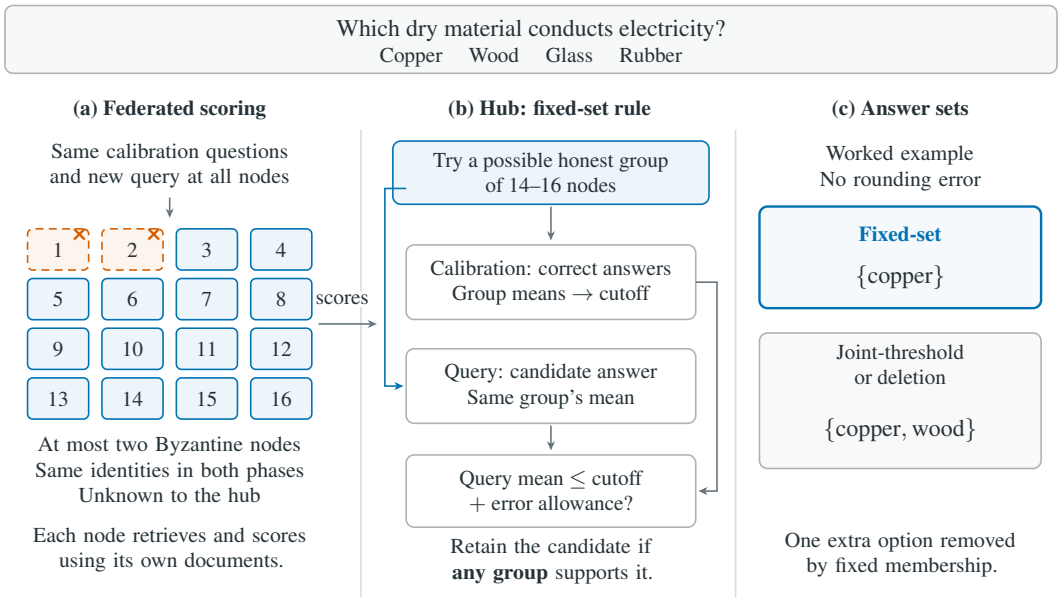


Figure 1: Requiring one possible honest group to support a candidate in both phases (fixed membership, b) removes an answer that two relaxations keep because they may use different groups for the cutoff and the query (c; Section 3.3). The hub does not know which two nodes (marked ×) are Byzantine. (a) All 16 nodes score the same calibration questions and the query. (b) A group supports a candidate if its mean query score is at most the cutoff set by its own calibration scores, plus a rounding allowance; the proposed fixed-set rule keeps exactly the candidates that some group supports. (c) No group supports wood, so fixed-set returns {copper}, while joint-threshold and deletion return {copper, wood}.

2. *The smallest answer sets that are safe.* Among methods that see only the nodes' reports, none can return fewer answers without risking the loss of an answer the honest nodes support (Section 4). Every answer removed is one fewer candidate for a person or a downstream system to check.
3. *Caution at little cost.* Because the method tracks which groups of nodes could be honest, declaring a generous budget for the number of bad nodes adds few answers. In simulations where Byzantine nodes send the largest score in both phases, raising the budget from two to six adds about one fifth of an answer per question, against about one full answer for a simpler robust method. Operators can therefore err on the side of safety. On real question-answering tasks, including medical exams, the method also returns smaller sets than the simpler one; with language-model judges as nodes, its advantage again grows with the budget (Section 5).

2 Problem Formulation

Frozen federation. Each of K nodes scores a finite list of candidate answers with a frozen pipeline, one whose retrieval and scoring settings are fixed before calibration, and an unknown subset of the nodes is Byzantine. Freezing makes a node's *clean* score, the one it computes before any rounding or tampering, mean the same thing during calibration and prediction. Let $\mathcal{X}$ be the set of questions and $\mathcal{Y}$ a finite set of M candidate answers. In a multiple-choice question $x \in \mathcal{X}$ with M listed options, $\mathcal{Y} = \{1, \dots, M\}$ and answer y is the yth listed option. Write $[q] = \{1, \dots, q\}$. Among K authenticated nodes, a fixed unknown set $\mathcal{A} \subseteq [K]$ is Byzantine and $\mathcal{H} = [K] \setminus \mathcal{A}$ honest. The actual number of Byzantine nodes, $a = |\mathcal{A}|$, obeys the declared integer budget A ($0 \le a \le A < K$), and $h = |\mathcal{H}| = K - a$. Each honest node freezes its corpus, retriever, model, prompt, and normalization before calibration. Their composition is a score $s_i : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, smaller favoring inclusion; in our testbeds it is one minus the model's probability for the answer given the node's passages. The components' training data are independent of the calibration split. The honest-mean score is

$$s_{\mathcal{H}}(x, y) = \frac{1}{h} \sum_{i \in \mathcal{H}} s_i(x, y). \tag{1}$$

Directional reconstruction errors. Known bounds on transmission error let the hub protect coverage even when honest scores are rounded to a few bits before they are sent. Each phase has its

own transmission *channel* $r \in \{\mathrm{cal}, \mathrm{qry}\}$ (calibration or query). The hub decodes the bits it receives into a number, the *decoded report*; for honest node i's clean score u it is $Q_{i,r}(u)$. Here $Q_{i,r}$ is a deterministic quantizer *registered* (fixed and recorded) before calibration. Nonnegative ρ_r^- and ρ_r^+ bound how far the report falls below or exceeds u (the *reconstruction error*):

$$-\rho_r^- \le Q_{i,r}(u) - u \le \rho_r^+. \tag{2}$$

Quantizers may differ by node and phase; scores need not be bounded (Appendices C and H).

Arbitrary messages, fixed identities. Byzantine reports may depend on anything, labels and honest reports included, and may switch strategy between phases. We assume honest nodes respond and that the hub, corpora, and honest score functions are trusted and frozen. A node misled by an injected prompt counts as Byzantine. Out of scope are poisoning of calibration data or of honest nodes' corpora or models, changing membership, Sybil attacks (forged identities), and unmodeled distribution shift (Appendix C). Corollary 4.2 relaxes the cap $a \le A$.

The honest-mean oracle and the clean rank event. The *honest-mean oracle* is the set the hub would obtain if it knew which nodes were honest and could read their exact scores. Fix $n \ge 1$ and let $Z_j = (X_j, Y_j)$, $j \in [n+1]$, be clean examples exchangeable conditional on the frozen system and honest set: their joint law is unchanged by reordering, and i.i.d. examples are a special case. The first n calibrate the rule and the last is the future question. Write $R_j = s_{\mathcal{H}}(X_j, Y_j)$, fix $\alpha \in (0,1)$, and set $k = \lceil (n+1)(1-\alpha) \rceil$. For $k \le n$, let $R_{(k)}$ be the kth smallest calibration score, an *order statistic*. The split-conformal honest-mean oracle (Lei et al., 2018) is $\mathcal{C}^\circ_\alpha(x) = \{y : s_{\mathcal{H}}(x,y) \le R_{(k)}\}$; for $k = n+1$, it and all three rules proposed in Section 3 return $\mathcal{Y}$.

One event drives every guarantee below. Call

$$\{R_{n+1} \le R_{(k)}\} \quad \text{equivalently} \quad \{Y_{n+1} \in \mathcal{C}^\circ_\alpha(X_{n+1})\} \tag{3}$$

the *clean rank event*: the future question's clean honest-mean score falls at or below the clean calibration cutoff. Exchangeability makes R_{n+1} equally likely to take each of the $n+1$ ranks (ties only help), so the event has probability at least $k/(n+1) \ge 1-\alpha$ (Lemma L.1).

Problem statement. Using only the decoded reports, the budget A, and the error bounds, the hub must return a set $\mathcal{C}_\alpha(x)$ that meets three requirements. (i) *Coverage*: $\mathbb{P}\{Y_{n+1} \in \mathcal{C}_\alpha(X_{n+1})\} \ge 1-\alpha$, whatever the Byzantine nodes send. (ii) *Clean-score integrity*: the set keeps every y for which the clean rank event could hold under some possible honest group (any set of at least $K - A$ nodes) and clean scores consistent with the reports (made precise in Theorem 4.3). (iii) *Minimal size*: the set is as small as possible given (i) and (ii). Integrity implies coverage, because the oracle's answers are always kept. Appendix Q lists the assumptions of Theorems 4.1 and 4.3 and what they do *not* require. Appendix I tabulates the notation.

3 Proposed method: aligned fixed-membership inference

The method exploits *fixed membership*: the honest set $\mathcal{H}$ is the same in the calibration and query phases. Section 3.1 specifies the data the hub collects, Section 3.2 the fixed-set rule, and Section 3.3 two relaxations that let the honest group change. Each rule rests on an *envelope*, a bound on the honest mean that holds for every possible honest group.

3.1 Aligned calibration

For every calibration example $j \in [n]$, the hub sends X_j and its candidates, but not Y_j, to all K nodes; at a query x it does the same. This design is *alignment*. Let $\widetilde{s}^{\mathrm{cal}}_{ij}(y)$ and $\widetilde{s}^{\mathrm{qry}}_i(x,y)$ be node i's decoded calibration and query scores: $Q_{i,r} \circ s_i$ on channel r if $i \in \mathcal{H}$, and arbitrary finite values if $i \in \mathcal{A}$. The hub keeps

$$V = (v_{ij}) \in \mathbb{R}^{K \times n}, \quad v_{ij} = \widetilde{s}^{\mathrm{cal}}_{ij}(Y_j); \qquad w(x,y) \in \mathbb{R}^K, \quad w_i(x,y) = \widetilde{s}^{\mathrm{qry}}_i(x,y). \tag{4}$$

For $u \in \mathbb{R}^K$ and $H \subseteq [K]$, write $\bar{u}_H = |H|^{-1} \sum_{i \in H} u_i$.

Joint information is necessary: the law of the nodes' mean score is determined by their joint law, but not by their per-node marginal laws (Proposition B.1; cf. Makarov, 1982). Column j of V holds all nodes' scores on one example, so V retains the joint law. For $U \sim \mathrm{Unif}[0,1]$ and $T(u) = 1 - |2u-1|$, the pairs (U,U) and $(U,T(U))$ have the same uniform marginals, yet the cumulative distribution

functions (CDFs) of their means differ by up to $1/4$. At a 75% target, per-node calibration is exact for the first pair but covers 99.7% for the second; aligned calibration covers 75.1% for both (Table 8). Other routes to the joint law, such as sending cross-moments, need a correctly specified model of how the nodes' scores depend on each other; alignment needs no such model (Remark J.1).

3.2 THE FIXED-SET RULE

The possible honest groups are the *feasible* subsets

$$\mathfrak{H}_A = \{H \subseteq [K] : |H| \geq K - A\}, \qquad c_j(H; V) = \frac{1}{|H|} \sum_{i \in H} v_{ij}, \tag{5}$$

where $c_j(H; V)$ is the mean calibration report of group H on example j. Groups larger than $K - A$ are included because a may be smaller than A. For $k \leq n$, let os_k return the kth smallest of n values. Group H's calibration cutoff, and the largest such cutoff, are

$$T_k(H; V) = \mathrm{os}_k\{c_1(H; V), \ldots, c_n(H; V)\}, \qquad \tau_{A,k}(V) = \max_{H \in \mathfrak{H}_A} T_k(H; V). \tag{6}$$

With clean reports, $T_k(\mathcal{H}; V)$ is the oracle cutoff $R_{(k)}$. The padding $g_\rho = \rho^-_{\mathrm{cal}} + \rho^+_{\mathrm{qry}}$ allows for calibration reports rounded down and query reports rounded up. Group H *supports* y if $\bar{w}_H(x, y) \leq T_k(H; V) + g_\rho$, and fixed-set keeps exactly the supported candidates:

$$\Delta_{A,k}(V, w) = \min_{H \in \mathfrak{H}_A} \big\{\bar{w}_H - T_k(H; V)\big\}, \tag{7}$$

$$\mathcal{C}^{\mathrm{fs}}_\alpha(x) = \{y : \Delta_{A,k}\{V, w(x, y)\} \leq g_\rho\} = \big\{y : \exists H \in \mathfrak{H}_A,\ \bar{w}_H(x, y) \leq T_k(H; V) + g_\rho\big\}. \tag{8}$$

3.3 TWO RELAXATIONS

The two relaxations, *joint-threshold* and *deletion*, let the query use a different group from calibration, and deletion also lets each calibration example use its own group:

$$y \in \mathcal{C}^{\mathrm{jt}}_\alpha(x) \iff \exists H, H' \in \mathfrak{H}_A :\ \bar{w}_{H'}(x, y) \leq T_k(H; V) + g_\rho,$$

$$y \in \mathcal{C}^{\mathrm{del}}_\alpha(x) \iff \exists H', H_1, \ldots, H_n \in \mathfrak{H}_A :\ \bar{w}_{H'}(x, y) \leq \mathrm{os}_k\big\{c_1(H_1; V), \ldots, c_n(H_n; V)\big\} + g_\rho.$$

Each relaxation allows more choices of groups than the rule before it, so $\mathcal{C}^{\mathrm{fs}}_\alpha \subseteq \mathcal{C}^{\mathrm{jt}}_\alpha \subseteq \mathcal{C}^{\mathrm{del}}_\alpha$ (Theorem 4.3). Both have closed forms. With $v \in \mathbb{R}^K$ sorted as $v_{(1)} \leq \cdots \leq v_{(K)}$, the extreme feasible-subset means are the deletion means

$$\mathrm{U}_A(v) = \max_{H \in \mathfrak{H}_A} \bar{v}_H = \frac{1}{K - A} \sum_{\ell = A+1}^{K} v_{(\ell)}, \qquad \mathrm{L}_A(v) = \min_{H \in \mathfrak{H}_A} \bar{v}_H = \frac{1}{K - A} \sum_{\ell = 1}^{K-A} v_{(\ell)}, \tag{9}$$

which drop the A smallest and the A largest reports, respectively. Minimizing over H' and maximizing over H gives

$$\mathcal{C}^{\mathrm{jt}}_\alpha(x) = \{y : \mathrm{L}_A\{w(x, y)\} \leq \tau_{A,k}(V) + g_\rho\}, \tag{10}$$

and, because os_k is nondecreasing in each argument, maximizing over each H_j separately gives, with $v_{\cdot j}$ the jth column of V, $\widehat{R}_j = \mathrm{U}_A(v_{\cdot j})$, $\widehat{R}_{(k)}$ the kth smallest of $\widehat{R}_1, \ldots, \widehat{R}_n$, and $\widehat{s}^{\mathrm{qry}}_{\mathrm{del}}(x, y) = \mathrm{L}_A\{w(x, y)\}$,

$$\widehat{q}^{\mathrm{del}}_\alpha = \widehat{R}_{(k)} + g_\rho, \qquad \mathcal{C}^{\mathrm{del}}_\alpha(x) = \{y : \widehat{s}^{\mathrm{qry}}_{\mathrm{del}}(x, y) \leq \widehat{q}^{\mathrm{del}}_\alpha\}. \tag{11}$$

Joint-threshold thus compares every query with one scalar threshold $\tau_{A,k}(V)$ computed once from calibration.

Worked example. Fixed membership removes wood from the introductory question's answer set (Figure 1), although both relaxations keep it. Take $K = 16$, $A = 2$, $n = 9$, $\alpha = 0.1$, $k = 9$, and no rounding. Nodes 1–2 report $4/5$ on every calibration example and $9/10$ for every option; nodes 3–16 report 0 on calibration and $0,\ 1/10,\ 1/5,\ 3/10$ for copper, wood, glass, and rubber. Every column of V is the same, so $\tau_{A,k} = \widehat{R}_{(k)} = 4/35$ (the two $4/5$ reports averaged with twelve zeros). For wood, $\mathrm{L}_A(w) = 1/10 \leq 4/35$, so both relaxations keep it. But every node's wood report exceeds its own calibration report by $1/10$, so $\bar{w}_H - T_k(H; V) = 1/10 > 0$ for every H, and fixed-set drops wood. With nodes 3–16 honest, fixed-set and the oracle return {copper} and both relaxations

{copper, wood}: the relaxations' largest cutoff and smallest query mean come from different groups, and no single group supports wood.

Which rule to deploy. Tighter sets cost more hub computation but no extra communication. Fixed-set enumerates $|\mathfrak{H}_A| = \sum_{d \le A} \binom{K}{d}$ subsets: 137 at the default $K = 16$, $A = 2$; 697 at $A = 3$; 43,745 at $K = 64$, $A = 3$. Use fixed-set where that is affordable, as here (Section 5); joint-threshold when a single reusable threshold is preferable, since it enumerates only during calibration; and deletion, which needs no enumeration, at larger A (Appendices D and M.1).

4 Theoretical guarantees: coverage, minimality, and set size

Our guarantees answer the three questions a user of the rules would ask. Does the set still contain the correct answer whatever Byzantine nodes send (Theorem 4.1 and Corollary 4.2)? Could any safe rule return a smaller set (Theorem 4.3)? How many extra answers does robustness cost (Theorem 4.4)? All three concern one property: retaining every candidate the honest-mean oracle would include. A *transcript* is the full collection of decoded reports, V and every $w(x, y)$. It is *admissible* if at most A nodes are Byzantine and every honest report obeys Equation (2). A statement is *pathwise* if it holds for every calibration sample, query, and admissible transcript, before any averaging.

4.1 Coverage under arbitrary reports

The deletion means bracket the honest mean with no distributional model for Byzantine reports: being the extreme feasible-subset means (Section 3.3), they satisfy $\mathrm{L}_A(v) \le \bar{v}_{\mathcal{H}} \le \mathrm{U}_A(v)$ whenever at most A entries of v come from nodes that are Byzantine. Padding by ρ_r^- and ρ_r^+ keeps the bracket when honest reports are rounded, and neither bound can be tightened without more information (Theorem K.1).

Theorem 4.1 (Finite-sample validity of the deletion rule)**.** *Assume $Z_1, \ldots, Z_n, Z_{n+1}$ are exchangeable conditional on the frozen honest score functions and fixed honest set. Under $a \le A < K$, suppose the directional reconstruction bounds in Equation* (2) *hold. Then, uniformly over all finite Byzantine calibration and query reports, $\mathbb{P}\{Y_{n+1} \in \mathcal{C}_\alpha^{\mathrm{del}}(X_{n+1})\} \ge k/(n+1) \ge 1 - \alpha$. Validity needs no bounded scores, continuity, density, cross-node independence, or attack model.*

Thus arbitrary reporting in both phases cannot push coverage below the chosen level. The padded envelopes bracket the clean means on both channels, and order statistics preserve the inequality. Exchangeability is needed only for the clean rank event, as in certified robust conformal inference (Gendler et al., 2022; Yan et al., 2024; Appendix L).

When nodes fail at random, coverage degrades gracefully. A *failure law*, a probability distribution over which nodes are Byzantine, can replace the cap $a \le A$ without changing the rule (the proof also uses Theorem 4.3).

Corollary 4.2 (Coverage under a random failure law)**.** *Suppose the Byzantine set $\mathcal{A}$ is drawn once, before calibration, from any probability law under which the clean examples are exchangeable conditional on each realized value of $\mathcal{A}$ (for instance, independently of the data). Suppose also that every node's reports obey Equation* (2) *whenever that node is honest, and that $A < K$ is fixed in advance. Let $\varepsilon_A = \mathbb{P}\{|\mathcal{A}| > A\}$. Then for each of the three proposed rules, uniformly over finite Byzantine reports, which may depend on $\mathcal{A}$,*

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}_\alpha(X_{n+1})\} \;\ge\; \frac{k}{n+1}\,(1 - \varepsilon_A) \;\ge\; 1 - \alpha - \varepsilon_A. \tag{12}$$

The guarantee weakens only when more nodes fail than the budget declares, and by at most that event's probability. For independent failures, ε_A is the Poisson–binomial probability that more than A nodes fail (Appendix N). It is nondecreasing in every node's failure probability, so upper bounds on those give a valid coverage *floor*, the right side of Equation (12), to advertise. If the bounds are wrong, the corollary still holds at the true ε_A.

4.2 FIXED-SET IS THE SMALLEST RULE WITH CLEAN-SCORE INTEGRITY

Fixed-set improves on joint-threshold only by pairing each group's cutoff with that same group's query mean: joint-threshold already uses the sharpest scalar bound on the clean calibration order statistic, $\tau_{A,k}(V) + \rho^{-}_{\mathrm{cal}}$ (Lemma L.2). A *report-only* rule uses only the reports, the declared budget, and the error bounds.

Theorem 4.3 (Rule hierarchy and minimality)**.** *Assume the conditions of Theorem 4.1. For $k \le n$:*

1. *For every query and admissible transcript,*
$$\mathcal{C}^{\circ}_{\alpha}(x) \subseteq \mathcal{C}^{\mathrm{fs}}_{\alpha}(x) \subseteq \mathcal{C}^{\mathrm{jt}}_{\alpha}(x) \subseteq \mathcal{C}^{\mathrm{del}}_{\alpha}(x), \tag{13}$$
and each proposed set has coverage at least $k/(n+1) \ge 1-\alpha$.

2. *Fix decoded reports V and $w(x,y)$. By Equation* (2)*, an honest clean value lies at most ρ^{+}_{r} below and ρ^{-}_{r} above its report on channel r. An* entrywise compatible completion *is a set $H \in \mathfrak{H}_A$ with clean values in these intervals at every calibration and query entry of its nodes, and nothing else imposed. This is the* interval model*: it assumes no common score function or cross-example law. The clean rank event holds for y in a completion if H's mean clean query value for y is at most the kth smallest of H's n mean clean calibration values. Then*
$$y \in \mathcal{C}^{\mathrm{fs}}_{\alpha}(x) \iff \textit{the clean rank event holds for } y \textit{ in some such completion}, \tag{14}$$
so $\mathcal{C}^{\mathrm{fs}}$ is pointwise smallest among report-only rules that must keep every such y.

Fixed membership therefore removes every candidate that no single honest group can explain, and nothing more. A retained candidate must admit one coherent explanation: one possible honest group whose clean scores, within their error intervals, place it under that group's cutoff. When none exists, fixed-set drops the candidate even when separate optimizations would keep it, at no cost to oracle inclusion or coverage. Part 2 shows that this test is exact, with g_ρ tight, so no report-only rule with the clean-score integrity of Section 2 can discard more. This minimality is relative to the interval model: a hub modeling how honest scores relate across examples could rule out completions. Each inclusion in Equation (13) can be strict (Appendix L.1).

4.3 REPORT SPREAD LIMITS THE EXTRA ANSWERS

Theorem 4.4 (Pathwise proposed-set sandwich)**.** *Fix an honest set $\mathcal{H}$ with $a = K - |\mathcal{H}| \le A < K$, assume Equation* (2)*, and suppose all decoded reports on channel $r \in \{\mathrm{cal}, \mathrm{qry}\}$ lie in a common interval of width D_r. With the membership term $\mu_r = \min\{1, A/(K-A)\}D_r$ and $W_{\mathrm{del}} = \sum_r(\mu_r + \rho^{-}_{r} + \rho^{+}_{r})$, for $k \le n$, Equation* (13) *extends pathwise to $\mathcal{C}^{\mathrm{del}}_{\alpha}(x) \subseteq \{y : s_{\mathcal{H}}(x,y) \le R_{(k)} + W_{\mathrm{del}}\}$.*

Robustness therefore costs only candidates whose clean scores lie near the oracle cutoff: every extra candidate has a clean score within W_{del} above it, and no score smoothness is needed. The width combines uncertainty about which nodes are honest with reconstruction error, so extra bits shrink the second part but never the first. How many extra candidates appear depends on how many candidate scores lie near the cutoff (Corollary L.3 and Remark L.4).

The direction of an attack decides how many extra answers remain. Consider reports in a range $[\mathrm{lo}, \mathrm{hi}]$ ($[0,1]$ here). For any number of corrupt nodes, *high–low* (hi at calibration, lo at query) yields the largest proposed sets and *low–high* (lo then hi) the smallest, so they test worst-case inflation and coverage. High–low and low–high are switching attacks: they change direction between phases. Low–high targets coverage, since its small calibration reports lower the cutoff and its large query reports raise every candidate's score. *Stable maximum* sends hi in both phases. With $a = A$, low–high cancels the corrupt contribution: the honest group then gives both the largest calibration threshold and the smallest query mean, so all three rules coincide. Raising A with reports fixed never shrinks a set (Appendix E).

Two simpler baselines are also valid, but neither returns smaller sets than fixed-set while keeping clean-score integrity. A *guarded symmetric rule* discards the $m \ge A$ largest and m smallest reports, averages the rest, and adds a guard wide enough to restore validity. With a guard computed from the full range of values a channel can carry, the three proposed sets are contained in it pathwise (Proposition F.1). The partial-conjunction p-value merger ranks each node's query score among that

node's own calibration scores (a conformal p-value). It then discards the A smallest p-values and thresholds a scaled order statistic of the rest. It is valid when every honest node's p-value is, but it never estimates the honest mean, so it targets a different event (Proposition G.2). It can therefore return smaller sets than fixed-set (Section 5) without contradicting Theorem 4.3.

5 Simulations and experiments

We test the rules in three settings of increasing realism. The first uses simulated scores, where every factor can be varied. The second uses four multiple-choice question-answering benchmarks, three of them medical (Mihaylov et al., 2018; Jin et al., 2021; Pal et al., 2022; Hendrycks et al., 2021), scored by a small language model (GPT-2; Radford et al., 2019) over 16 document shards. The third uses 16 copies of a stronger model (Llama-3.2-3B; Grattafiori et al., 2024) as nodes, some hijacked by prompt injection. Unless stated otherwise, there are 16 nodes, at most two Byzantine, and the target coverage is 90%. Every method sees the same data and the same corrupted nodes, whose identities are hidden. We compare the three proposed rules with the honest-mean oracle, two other rules with guarantees, and seven common robust summaries without one, such as the median and the plain average. The full design, all comparators, and every number are in Appendices O, P, and R.

Fixed-set is closest to the oracle, and its advantage grows with the declared budget. Call the answers a rule keeps beyond the honest-mean oracle its *excess*. In simulations where the corrupt nodes send the largest possible score in both phases, fixed-set removes about three quarters of deletion's excess. Declaring a budget of six instead of two costs deletion about one extra answer per question, but fixed-set only about one fifth (Figure 2a). A larger budget admits more possible honest groups; deletion can take its cutoff from one group and its query score from another, while fixed-set must find a single group that supports the candidate. A cautious budget is therefore cheap. Sending scores with eight bits instead of exact values costs almost nothing (Figure 2b). Extra bits shrink only the rounding part of the width in Theorem 4.4, not the part due to not knowing which nodes are honest, so set sizes level off.

Coverage holds whenever the assumptions do. In every simulated setting that meets the assumptions, the proposed rules reach the target coverage (Table 11). They also reach it when nodes fail at random, as in the language-model experiment below, where some runs have more failures than declared (Table 19). Theorems 4.1 and 4.3 guarantee this within the budget; with random failures, Corollary 4.2 guarantees only a lower floor, which the rules exceed.

Summaries without a guarantee can fail. When corrupt nodes report low scores during calibration and high scores at the query, five of the seven common summaries miss the 90% target, one by more than 30 points, while the proposed rules do not (Figure 3a). The plain average fails the same way on the real benchmarks (Table 18). An attack that behaves the same in both phases cannot cause this, because a summary's calibration and query scores then stay exchangeable (Lemma G.1). The proposed rules hold because their envelopes bound the honest mean in each phase, whatever the reports (Theorem 4.1). Realistic failures are milder, but hijacked language models that mislead only the calibration step still pull the plain average below the target (Figure 3b). They lower the high calibration scores that set its cutoff, so it drops the correct answer more often (Appendix R).

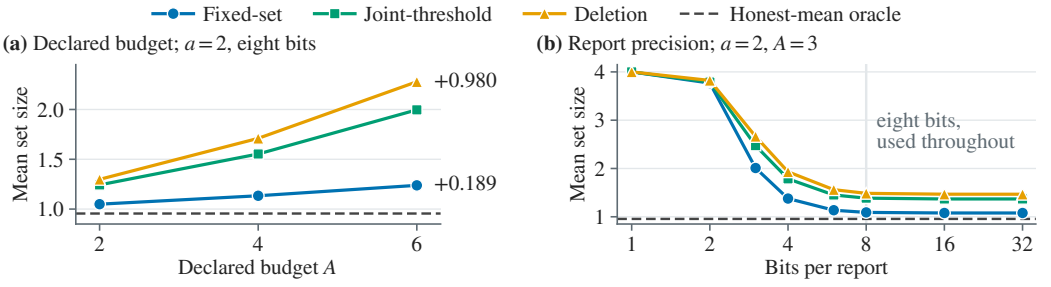


Figure 2: Fixed-set stays close to the oracle as the declared budget grows (a), and eight-bit reports lose almost nothing (b). Simulated scores; corrupt nodes send the largest score in both phases; dashed line: honest-mean oracle; labels in (a): rise from $A = 2$ to 6. Details in Tables 9 and 10.

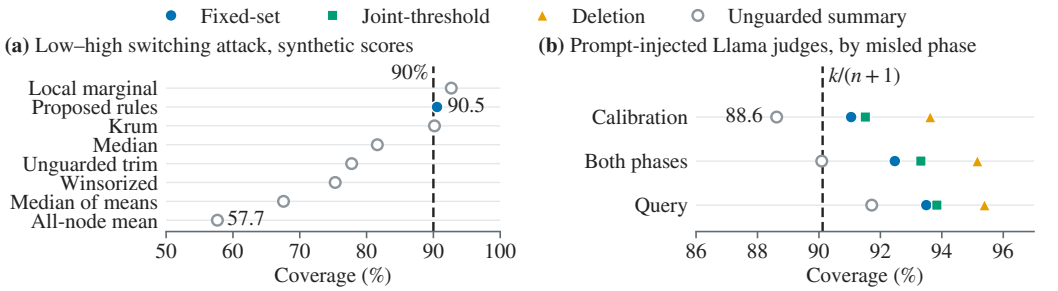


Figure 3: Summaries without a guarantee can miss the coverage target where the proposed rules hold. (a) Simulated scores; corrupt nodes report low scores at calibration and high scores at the query. (b) Llama-3.2-3B judges, some hijacked by prompt injection, by the phase they mislead. Comparators are defined in Appendix O.4.

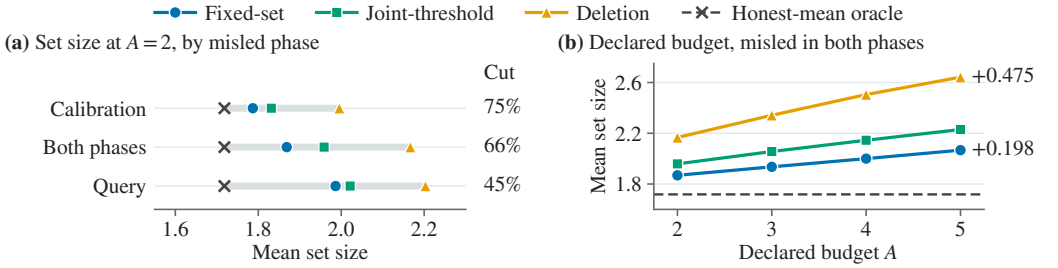


Figure 4: With 16 Llama-3.2-3B judges, some hijacked by prompt injection, fixed-set removes 45–75% of deletion's excess over the honest-mean oracle (a), and its advantage grows with the declared budget (b). (a) Rows: phases in which judges are misled; bars span the oracle to deletion, and Cut is the share of that excess fixed-set removes. (b) Judges misled in both phases (Table 20).

The savings carry over to real scorers. With the small GPT-2 scorer, fixed-set covers at least 90% in every real-data condition within the declared budget and, under stable maximum, removes 34–45% of deletion's excess on every benchmark (Table 13). Because this scorer is close to chance, these numbers measure the cost of robustness, not answer quality (Appendix P.7). With the stronger Llama judges, all proposed rules still reach the target, and fixed-set removes 45–75% of deletion's excess. Its saving roughly doubles as the declared budget grows from two to five (Figure 4), as in the simulations. Like wood in Section 3.3, the removed answers are ones that no possible honest group supports. A tempting shortcut, dropping the two judges that erred most on held-out questions, shrinks the sets further. But in two thirds of the runs it also drops answers the honest judges support, which fixed-set never does (Table 21).

The comparisons have limits, and fixed-set is cheap to run. One rule with a guarantee, the p-value merger, can return smaller sets under some attacks, because it protects a different target (Proposition G.2; Table 7). After calibration, fixed-set answers a batch of 667 questions in about 6 ms of hub computation (Table 1).

6 Discussion and Conclusion

Robustness becomes affordable once the hub remembers that the same nodes are honest throughout: fixed-set inference keeps every candidate for which the clean rank event can still hold under one possible honest group and, in the interval model, no others. Its payoff grows with the declared budget. Natural next steps are faster subset search for large federations, open-ended generation, and partial alignment, in which nodes share only some calibration questions.

REPRODUCIBILITY STATEMENT

Proofs and the full experimental design are in the appendix (see the roadmap). The supporting tables and figures reuse verified numerical assets from archived outputs; Appendix P describes reaggregation, implementation checks, and the disclosed correction to constant-column uncertainty calculations. The exact finite-alphabet verifier in the supplement passes all 15 of its checks. The accompanying supplementary archive contains implementation code, dependency locks, verified numerical summaries, and a reproduction `README.md`. It supports local verification and regeneration of the reported numerical assets; bulk model scores and original execution archives are not included, and the README lists what the archive omits. The archive also holds the registered manifests and every generated table and figure.

USE OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

We used generative AI tools (ChatGPT, Google Gemini, and Claude) for language editing and code formatting support. All data, results, and mathematical derivations are the authors' own work.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

James Bell, Adrià Gascón, Tancrède Lepoint, Baiyu Li, Sarah Meiklejohn, Mariana Raykova, and Cathie Yun. ACORN: Input validation for secure aggregation. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 4805–4822. USENIX Association, 2023. URL `https://www.usenix.org/conference/usenixsecurity23/presentation/bell`.

Yoav Benjamini and Ruth Heller. Screening for partial conjunction hypotheses. *Biometrics*, 64(4):1215–1222, 2008. doi: 10.1111/j.1541-0420.2007.00984.x.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, pp. 119–129. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/f4b9ec30ad9f68f89b29639786cb62ef-Abstract.html`.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191. ACM, 2017. doi: 10.1145/3133956.3133982.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516, 2024. doi: 10.1162/tacl_a_00715.

Abhijit Chakraborty, Chahana Dahal, and Vivek Gupta. Federated retrieval-augmented generation: A systematic mapping study. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 7362–7374. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.388.

Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017. doi: 10.1145/3154503.

Jason Clarkson, Wenkai Xu, Mihai Cucuringu, and Gesine Reinert. Split conformal prediction under data contamination. In *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 230 of *Proceedings of Machine Learning Research*, pp. 5–27. PMLR, 2024. URL `https://proceedings.mlr.press/v230/clarkson24a.html`.

Akash Dhasade, Rachid Guerraoui, Anne-Marie Kermarrec, Diana Petrescu, Rafael Pires, Mathis Randl, and Martijn de Vos. Efficient federated search for retrieval-augmented generation using lightweight routing. In *Distributed Applications and Interoperable Systems*, volume 16591 of *Lecture Notes in Computer Science*, pp. 3–20. Springer, 2026. doi: 10.1007/978-3-032-27358-1_1.

Prasanjit Dubey and Xiaoming Huo. Anytime-valid federated conformal RAG for LLM swarms, 2026a. URL https://arxiv.org/abs/2605.29139. arXiv:2605.29139v1.

Prasanjit Dubey and Xiaoming Huo. Federated language models under bandwidth budgets: Distillation rates and conformal coverage, 2026b. URL https://arxiv.org/abs/2605.09986. arXiv:2605.09986v2.

Naihe Feng, Yi Sui, Shiyi Hou, Jesse C. Cresswell, and Ga Wu. Response quality assessment for retrieval-augmented generation via conditional conformal factuality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2832–2836. ACM, 2025. doi: 10.1145/3726302.3730244.

Matteo Gasparin and Aaditya Ramdas. Merging uncertainty sets via majority vote, 2024. URL https://arxiv.org/abs/2401.09379.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9L1BsI4wP1H.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Rachid Guerraoui, Nirupam Gupta, and Rafael Pinot. Byzantine machine learning: A primer. *ACM Computing Surveys*, 56(7):1–39, 2024. doi: 10.1145/3616537.

Hangyu He, Xin Yuan, Kai Wu, Ren Ping Liu, and Wei Ni. pFedRAG: A personalized federated retrieval-augmented generation system with depth-adaptive tiered embedding tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14255–14268, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.769. URL https://aclanthology.org/2025.findings-emnlp.769/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Pierre Humbert, Batiste Le Bars, Aurélien Bellet, and Sylvain Arlot. One-shot federated conformal prediction. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14153–14177. PMLR, 2023. URL https://proceedings.mlr.press/v202/humbert23a.html.

Linus Jeary, Tom Kuipers, Mehran Hosseini, and Nicola Paoletti. Verifiably robust conformal prediction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 4295–4314. Curran Associates, Inc., 2024. doi: 10.52202/079017-0140. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0814a342597d65e0832fc7ec9b42c317-Paper-Conference.pdf.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421.

Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-RAG: Certified generation risks for retrieval-augmented language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22963–23000. PMLR, 2024a. URL https://proceedings.mlr.press/v235/kang24a.html.

Mintong Kang, Zhen Lin, Jimeng Sun, Cao Xiao, and Bo Li. Certifiably Byzantine-robust federated conformal prediction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23022–23057. PMLR, 2024b. URL https://proceedings.mlr.press/v235/kang24c.html.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=jXKKDEi5vJt.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering, 2023. URL https://arxiv.org/abs/2305.18404. ICML 2023 Workshop on Neural Conversational AI (TEACH).

Ehsan Lari, Reza Arablouei, and Stefan Werner. Partial model sharing improves Byzantine resilience in federated conformal prediction, 2026a. URL https://arxiv.org/abs/2605.11684. Accepted to the 34th European Signal Processing Conference (EUSIPCO 2026).

Ehsan Lari, Reza Arablouei, and Stefan Werner. Communication-efficient Byzantine-robust federated conformal prediction via partial sharing. *IEEE Transactions on Signal Processing*, pp. 1–16, 2026b. doi: 10.1109/TSP.2026.3737023.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33*, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. TRAQ: Trustworthy retrieval augmented question answering via conformal prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3799–3821, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.210. URL https://aclanthology.org/2024.naacl-long.210/.

Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael I. Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22942–22964. PMLR, 2023. URL https://proceedings.mlr.press/v202/lu23i.html.

G. D. Makarov. Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability and Its Applications*, 26(4):803–806, 1982. doi: 10.1137/1126086.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36029–36047. PMLR, 2024. URL https://proceedings.mlr.press/v235/mohri24a.html.

Junjie Mu and Qiongxiu Li. A wolf in sheep's clothing: Targeted routing hijacking in federated RAG, 2026. URL https://arxiv.org/abs/2605.28112. Accepted to the EMNLP 2026 Main Conference.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Sangdon Park, Osbert Bastani, and Taesoo Kim. ACon^2: Adaptive conformal consensus for provable blockchain oracles. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 3313–3330. USENIX Association, 2023. URL https://www.usenix.org/conference/usenixsecurity23/presentation/park.

Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022. doi: 10.1109/TSP.2022.3153135.

Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27907–27947. PMLR, 2023. URL https://proceedings.mlr.press/v202/plassier23a.html.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Ahmad Rammal, Kaja Gruntkowska, Nikita Fedin, Eduard Gorbunov, and Peter Richtárik. Communication compression for Byzantine robust learning: New efficient algorithms and improved rates. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1207–1215. PMLR, 2024. URL https://proceedings.mlr.press/v238/rammal24a.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

B. Rüger. Das maximale Signifikanzniveau des Tests: "Lehne H_0 ab, wenn k unter n gegebenen Tests zur Ablehnung führen". *Metrika*, 25(1):171–178, 1978. doi: 10.1007/BF02204362.

Yan Scholten and Stephan Günnemann. Provably reliable conformal prediction sets in the presence of data poisoning. In *International Conference on Learning Representations*, pp. 55873–55897, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8c4d21a4b33361c40c2142db9511c126-Abstract-Conference.html.

Yuanjie Shi, Peihong Li, Xuanyu Cao, and Yan Yan. Valid and efficient uncertainty quantification for federated joint shift. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 6232–6260. PMLR, 2026. URL https://proceedings.mlr.press/v337/shi26a.html.

Ananda Theertha Suresh, Felix X. Yu, Sanjiv Kumar, and H. Brendan McMahan. Distributed mean estimation with limited communication. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3329–3337. PMLR, 2017. URL https://proceedings.mlr.press/v70/suresh17a.html.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020. doi: 10.1093/biomet/asaa027.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715. URL https://doi.org/10.1007/b106715.

Jingshu Wang and Art B. Owen. Admissibility in partial conjunction testing. *Journal of the American Statistical Association*, 114(525):158–168, 2019. doi: 10.1080/01621459.2017.1385465.

Zhaoqi Wang, Daqing He, Zijian Zhang, Ye Liu, Jiamou Liu, Zhirui Zeng, Zhan Qin, Zhen Li, Xin Li, Hongwei Yao, Jincheng An, Yong Liu, Yi Li, Qi Sun, Xiulei Liu, and Liehuang Zhu. Combating knowledge corruption in agent systems: A Byzantine-tolerant secure collaborative RAG framework. In *Proceedings of the ACM Web Conference 2026*, pp. 2661–2672. ACM, 2026. doi: 10.1145/3774904.3792200.

Haifeng Wen, Osvaldo Simeone, and Hong Xing. Efficient federated conformal prediction with group-conditional guarantee. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 7290–7313. PMLR, 2026. URL https://proceedings.mlr.press/v337/wen26a.html.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust RAG against retrieval corruption. In *2026 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 392–410. IEEE, 2026. doi: 10.1109/SaTML68715.2026.00029.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.372. URL https://aclanthology.org/2024.findings-acl.372/.

Ge Yan, Yaniv Romano, and Tsui-Wei Weng. Provably robust conformal prediction with improved efficiency. In *International Conference on Learning Representations*, pp. 31974–32010, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8759c20675d67ef3b91c4607cecbb27e-Abstract-Conference.html.

Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5650–5659. PMLR, 2018. URL https://proceedings.mlr.press/v80/yin18a.html.

Soroush H. Zargarbashi, Mohammad Sadegh Akhondzadeh, and Aleksandar Bojchevski. Robust yet efficient conformal prediction sets. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17123–17147. PMLR, 2024. URL https://proceedings.mlr.press/v235/h-zargarbashi24a.html.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13764–13775. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.849.

Meiyi Zhu, Matteo Zecchin, Sangwoo Park, Caili Guo, Chunyan Feng, and Osvaldo Simeone. Federated inference with reliable uncertainty quantification over wireless channels via conformal prediction. *IEEE Transactions on Signal Processing*, 72:1235–1250, 2024. doi: 10.1109/TSP.2024.3358615.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844. USENIX Association, 2025. URL https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

APPENDIX ROADMAP

Appendices A to H give related work, the protocol, and additional formal results. Appendices I to O give notation, proofs, implementation details, and the experimental design. Appendix P gives the supporting empirical results. Appendix Q lists every assumption the coverage and minimality theorems use. Appendix R covers restricted subset families and the `Llama-3.2-3B-Instruct` judge panel. Throughout, *registered* means fixed and recorded before any evaluation data were scored, and the *operational* configuration is the default of Section 5 ($K = 16$, $a = A = 2$, eight-bit reports; Appendix O).

A EXTENDED RELATED WORK AND NOVELTY BOUNDARY

Conformal RAG. TRAQ composes retrieval and generation uncertainty into valid answer sets in a centralized pipeline (Li et al., 2024), and C-RAG certifies bounds on generation risk, including under distribution shift (Kang et al., 2024a). Our finite-label formulation is closest to conformal sets over a language model's multiple-choice answers (Kumar et al., 2023), with retrieval and scoring folded into the frozen score s_i of Section 2. Conformal factuality (Mohri & Hashimoto, 2024) and group-conditional subclaim assessment (Feng et al., 2025) are single-site; Campos et al. (2024) survey conformal prediction for language tasks. None treats nodes that can alter query-time messages.

Federated conformal prediction. Federated conformal prediction (FCP) either assumes honest clients or defends calibration only. Honest-client methods target a client mixture under partial exchangeability (Lu et al., 2023), correct label shift (Plassier et al., 2023), calibrate prescribed group mixtures with mergeable coresets (Wen et al., 2026), handle joint shift with a score-distribution sketch (Shi et al., 2026), calibrate in one round (Humbert et al., 2023), or correct wireless channel effects (Zhu et al., 2024). Rob-FCP (Kang et al., 2024b) and PRISM-FCP (Lari et al., 2026b;a) filter Byzantine calibration summaries but leave query-time scores unprotected. ACon^2 forms an online consensus over Byzantine data sources (Park et al., 2023) rather than calibrating an honest mean, and contaminated split conformal prediction studies corrupted examples rather than corrupted nodes (Clarkson et al., 2024). FC-RAG (Dubey & Huo, 2026b) calibrates from examples partitioned across nodes and assumes those summaries approximate the quantile of the same-example mean; by Proposition B.1, node-wise marginals alone do not imply this. Anytime-FC-RAG gives sequential validity for honest swarms (Dubey & Huo, 2026a).

Certified robust conformal prediction and p-value merging. Randomly smoothed conformal prediction (RSCP) and its improved variant RSCP+ inflate the threshold with a randomized-smoothing envelope (Gendler et al., 2022; Yan et al., 2024). Verifiably robust conformal prediction (VRCP) obtains score bounds by neural-network verification (Jeary et al., 2024). Reliable prediction sets (RPS) and CDF-aware sets (CAS) protect sets against training or calibration poisoning and test-time evasion (Scholten & Günnemann, 2025; Zargarbashi et al., 2024). Our coverage proof shares their mechanism, certified score bounds that preserve the clean rank event. Here, however, the perturbations are coordinated, candidate-specific messages from unknown nodes in both phases, and coupling one honest set across all of them yields the fixed-set rule and its minimality (Theorem 4.3). Our p-merger baseline is the partial-conjunction construction (Benjamini & Heller, 2008; Wang & Owen, 2019) built on order-statistic merging (Rüger, 1978; Vovk & Wang, 2020), specialized to at most A arbitrary reports; majority vote similarly merges dependent conformal sets (Gasparin & Ramdas, 2024). We give a self-contained proof and sharp constant, not a new merger, and the merger does not target the honest mean.

Robust aggregation and communication. Coordinatewise medians, trimmed means, the median of means and geometric medians are standard Byzantine-robust tools (Yin et al., 2018; Chen et al., 2017; Pillutla et al., 2022; Guerraoui et al., 2024), and estimating an honest mean under heterogeneity costs an explicit spread term (Karimireddy et al., 2022). Work on quantized and compressed robust learning studies related communication tradeoffs (Suresh et al., 2017; Rammal et al., 2024). Krum's guarantees rest on SGD assumptions that do not transfer to candidate scores (Blanchard et al., 2017), so it is only a secondary baseline. Our one-sided deletion means solve an extremal feasible-subset problem, and the envelopes use deterministic error bounds (Lemma I.1) rather than assuming quantization stays unbiased after sorting.

Federated RAG systems and attacks. Chakraborty et al. (2025) map federated RAG; routing and personalized retrieval reduce cost (Dhasade et al., 2026; He et al., 2025). SecureCollaRAG filters documents under an independent malicious-source model (Wang et al., 2026), routing hijacking attacks a different surface (Mu & Li, 2026), corpus poisoning corrupts the honest score functions themselves (Zhong et al., 2023; Zou et al., 2025), and RobustRAG certifies answers against malicious passages within one pipeline (Xiang et al., 2026). None provides conformal coverage against arbitrary node messages.

B WHY LOCAL SCORE DISTRIBUTIONS CANNOT CALIBRATE THE HONEST MEAN

Exact local score distributions omit the dependence needed to calibrate a same-example mean. A cumulative distribution function (CDF) gives the probability that a score is at or below each threshold; write $\|F-G\|_\infty = \sup_t |F(t)-G(t)|$ for the largest CDF discrepancy. The following specialization of classical fixed-marginal dependence facts (Makarov, 1982) gives an explicit calibration obstruction.

Proposition B.1 (Marginals do not identify the honest-average law)**.** *Even with two honest nodes, no Byzantine reports, infinite calibration data, and continuous scores in* $[0,1]$*, the two node-wise marginal score laws do not identify the law of their average. There exist two worlds with identical* $\mathrm{Unif}[0,1]$ *node marginals whose average-score CDFs* F_0, F_1 *satisfy*

$$\|F_0 - F_1\|_\infty = \frac{1}{4}.$$

Consequently, every possibly randomized estimator $\widehat{F}$ *whose population input consists only of the local marginal CDFs obeys the following lower bound, where* $\mathbb{E}_r$ *averages over the estimator's randomness in world* r*:*

$$\max_{r\in\{0,1\}} \mathbb{E}_r\|\widehat{F} - F_r\|_\infty \geq \frac{1}{8}.$$

At target coverage $3/4$*, any common deterministic threshold valid in both worlds has at least* $1/4$ *overcoverage in one of them.*

With $U \sim \mathrm{Unif}[0,1]$, the score pairs are (U,U) in World 0, with average-score CDF F_0, and $(U,T(U))$ in World 1, with CDF F_1, where $T(u) = 2u$ for $u \leq 1/2$ and $T(u) = 2-2u$ otherwise; both node marginals are uniform. More precise local summaries cannot recover the missing dependence; alignment supplies it by keeping together the reports produced for the same example. The conclusion concerns calibration of the honest arithmetic mean, not other targets. Appendix J proves the result.

C DETAILED PROTOCOL AND COMMUNICATION MODEL

This protocol keeps three sources of uncertainty separate: unknown honest identities, reconstruction error from finite-bit reports, and missing honest responses. Bounded scores make communication costs and efficiency bounds explicit. At a fixed input and candidate, disagreement among honest nodes is measured by the clean pointwise range

$$R_{\mathcal{H}}(x,y) = \max_{i\in\mathcal{H}} s_i(x,y) - \min_{i\in\mathcal{H}} s_i(x,y).$$

For the explicit finite-bit protocol and all experiments, fix $0 < S_{\max} < \infty$ and include deterministic clipping in the frozen score so that $s_i \in [0, S_{\max}]$. Then $R_{\mathcal{H}} \leq S_{\max}$.

C.1 THREAT AND COMMUNICATION MODELS

Adversarial reports and missing messages. Authentication gives each registered node one report slot; the primary guarantee requires

$$0 \leq a \leq A < K.$$

Byzantine reports are unrestricted as in Section 2. The hub's input gateway authenticates one slot per registered node, rejects malformed contents, and fills every rejected or missing authenticated-node slot with a registered finite sentinel, a preassigned replacement report. Such a slot-level absence

or malformed signed packet is handled by the protocol's adversarial/dropout accounting, not as an infrastructure failure. Transport, scheduler, storage, or service failures are retried under the registered retry policy, fixed before evaluation. An unresolved infrastructure failure aborts the run rather than being relabeled as a sentinel report or deterministic model fallback (the frozen score map's registered output for a retrieval or model failure; Appendix M). An aborted run carries no coverage guarantee. The honest-dropout budget D is the number of honest nodes allowed to miss a report (unrelated to the report width D_r); the primary protocol sets $D = 0$: honest nodes respond, while a Byzantine nonresponse is simply one of the at most A arbitrary reports.

Information available to the hub and nodes. The default protocol broadcasts shared calibration inputs and candidates, so it is not a private-input protocol. It does not pool private node corpora. The vector form, in which every node scores all M candidates of each calibration example, does not transmit the calibration label to nodes, but it assumes the calibration inputs themselves may be shared. Sensitive inputs require consent, secure execution, or a separate privacy mechanism, none of which is analyzed here. Secure aggregation can hide individual submissions but cannot make their contents truthful; input validation is complementary (Bonawitz et al., 2017; Bell et al., 2023). Moreover, an ordinary sum-only secure-aggregation interface does not provide the identity-resolved report array needed to enumerate feasible honest sets, delete selected identities, or validate per-node packets. Implementing those rules while hiding reports would require a richer secure-computation and authenticated-input-validation protocol, which we do not analyze. Not transmitting the recorded label is an interface restriction, not a confidentiality guarantee: a capable node may infer the answer from the input.

Reconstructing finite-bit scores. Quantization replaces each real score by one of finitely many values that the hub can decode. For integer $b \geq 1$, let Q_b be the nearest-neighbor scalar quantizer with 2^b equally spaced reconstruction levels including 0 and $S_{\max}$:

$$Q_b(u) = \frac{S_{\max}}{2^b - 1} \operatorname{round}\left(\frac{(2^b - 1)u}{S_{\max}} \right), \qquad \rho(b) = \frac{S_{\max}}{2(2^b - 1)}. \tag{15}$$

For $u \in [0, S_{\max}]$, at an exact cell midpoint, round selects the smaller reconstruction level, making Q_b deterministic. Thus $|Q_b(u) - u| \leq \rho(b)$ on the declared score domain. Heterogeneous registered maps $Q_{i,r}$ satisfying Equation (2) are allowed; when $Q_{i,r} = Q_{b_i^r}$ with depths $b_i^r \geq 1$, set $b_r = \min_i b_i^r$ and $\rho_r^- = \rho_r^+ = \rho(b_r)$.

Aggregation notation. For real reports $v_1, \ldots, v_K$, let $v_{(1)} \leq \cdots \leq v_{(K)}$ be their order statistics. For a symmetric comparator, choose a registered trim m satisfying $A \leq m$ and $2m < K$, and define

$$\mathrm{TM}_m(v_1, \ldots, v_K) = \frac{1}{K - 2m} \sum_{\ell=m+1}^{K-m} v_{(\ell)}.$$

The primary symmetric comparator sets $m = A$; over-trimming is an efficiency diagnostic.

Label-aware scalar variant. The vector protocol does not transmit the label to nodes and makes calibration information operationally match deployment. If label disclosure is acceptable, each honest node may instead transmit only the true-label score. The guarded arbitrary-message results remain unchanged and calibration communication falls by a factor of M. The unguarded matched-pipeline result, which relies on identical scoring behavior in calibration and prediction (Lemma G.1), then requires a separate exchangeability justification, because Byzantine nodes see the label during calibration but not deployment.

D Computational cost and exact decisions

Fixed-set enumerates every feasible honest set, obtained by deleting at most A identities; their number is

$$N_A(K) = \sum_{d=0}^{A} \binom{K}{d}.$$

Thus $N_2(16) = 137$, $N_3(16) = 697$, and $N_3(64) = 43{,}745$. When $k \le n$, complement sums (subtracting deleted-node sums from the total) and linear-time order-statistic selection give calibration cost $O\{nK + n\sum_{d=0}^{A}\binom{K}{d}(d+1)\}$ and, after storing the $T_k(H;V)$ values, fixed-set query cost $O\{MK + M\sum_{d=0}^{A}\binom{K}{d}(d+1)\}$. Ordinary sorting adds the usual logarithmic factor. The registered quantized implementation uses integer code sums and exact rational comparisons to preserve boundary ties and the set hierarchy; Appendix M.1 gives the arithmetic contract.

Measured CPU cost. Exact fixed-set decisions complete in milliseconds for a batch of queries at the operational budget in a controlled local benchmark. We use the retained OpenBookQA score cache from both corpus partitions, with $K = 16$, $M = 4$, eight-bit clean reports, nominal 90% coverage, and declared budgets $A \in \{2, 3\}$. The first 333 cache-order examples calibrate the rule and the remaining 667 form one query batch. This fixed split measures computation; it adds no coverage estimate to the registered study. All methods receive the same quantized arrays. Measurements use an Intel Core i7-1068NG7 CPU at a nominal 2.30 GHz, 32 GiB RAM, Python 3.12.7, and NumPy 1.26.4, with one process and native thread pools limited to one. Each stage receives three warm-ups and 21 timed repetitions with rotating execution order; the code archive retains every timing and its quartiles.

The benchmark separates reusable calibration state from query work while preserving the production arithmetic and diagnostic outputs. Here "production API" is the released library function, which performs calibration and query in one call; the benchmark's "staged" code splits the two phases so that cached calibration state can be timed. Across both partitions, budgets, and three methods, all 12 staged-versus-unstaged output comparisons match the unchanged grid implementation (the exact integer-arithmetic implementation of Appendix M.1) exactly, and all prediction sets match the public production API. Table 1 also times the unchanged combined grid call. Timings exclude retrieval, model scoring, communication, input loading, quantization, grid validation, and true-label score extraction.

Table 1: Cached fixed-set inference processes 667-query batches in milliseconds. Times are medians of 21 measurements, with ranges across the two retained partitions; $K = 16$, $n = 333$, $M = 4$, eight-bit reports, and nominal 90% coverage. Query uses staged cached kernels; Combined uses the production function.

A	Rule	Cal. ms	Query batch ms	Combined ms	Cache KiB	Query peak MiB
2	Fixed-set	2.96–3.19	6.27–6.47	10.04–10.92	21.16	6.33
2	Joint-threshold	3.22–3.28	0.68–0.76	4.13–4.14	0.33	0.41
2	Deletion	0.35–0.40	0.57–0.60	0.92–1.00	0.33	0.41
3	Fixed-set	13.44–14.42	21.13–22.54	35.18–38.15	104.28	30.55
3	Joint-threshold	13.82–14.58	0.77–0.82	15.54–15.70	0.33	0.41
3	Deletion	0.33–0.82	0.59–0.69	1.04–1.14	0.33	0.41

For fixed-set at $A = 2$, dividing the batch time by 667 gives 9.39–9.70 μs per query (amortized, not isolated latency).

The memory columns distinguish retained state from allocations during query processing. Cached state counts reachable Python objects and owned NumPy storage; query peak is the maximum of three separate `tracemalloc` measurements per partition, excluding preallocated inputs and cache. It includes traced NumPy buffers but is not total process memory. Table 1 reports ranges of partition-specific medians for timing, not confidence intervals. Processor scheduling and frequency are uncontrolled; the benchmark is a local measurement of these kernels, not a claim about end-to-end deployment latency or scaling to larger budgets.

E Endpoint attacks and conservative budgets

Table 2 lists the three endpoint attacks of Section 4; the corollary below makes their roles precise.

Table 2: Each endpoint attack sends one end of the report range in each phase: high–low yields the largest proposed sets, and low–high at $a = A$ cancels the corrupt contribution. All three rules retain their coverage guarantee under the stated assumptions; efficiency depends on the direction of the reports.

Attack	Calibration	Query	Role in the comparison
Stable maximum	hi	hi	Measures gains from fixed membership across phases
High–low	hi	lo	Maximizes proposed-set inclusion for a fixed honest transcript
Low–high	lo	hi	At $a = A$, recovers the padded honest decoded comparison

Corollary E.1 (Endpoint attacks and conservative budgets)**.** *If $k = n + 1$, all set conclusions below are immediate, since every rule then returns $\mathcal{Y}$. Suppose $k \le n$. Fix the honest decoded reports and suppose Byzantine reports must lie in* $[\mathrm{lo}, \mathrm{hi}]$*. For each of $\mathcal{C}^{\mathrm{fs}}$, $\mathcal{C}^{\mathrm{jt}}$, and $\mathcal{C}^{\mathrm{del}}$, setting every Byzantine calibration report to* hi *and every Byzantine query report to* lo *produces a prediction set containing the set under every other Byzantine transcript. Thus the high–low endpoint attack is pathwise binding for set inclusion, and hence for set size and full-set status, within this bounded-message class. It also maximizes proposed-set coverage for fixed clean data and is therefore not an undercoverage stress. Symmetrically, setting every Byzantine calibration report to* lo *and every Byzantine query report to* hi *produces a set contained in the set under every other Byzantine transcript in* $[\mathrm{lo}, \mathrm{hi}]$*, whatever the number a of Byzantine nodes. Low–high therefore minimizes proposed-set coverage for fixed clean data within this class. Neither extremal result extends to the p-value merger or to point accuracy (whether the top-ranked answer is correct).*

Conversely, suppose $a = A$ and low–high sends calibration values no larger than every honest decoded value and query values no smaller than every honest decoded value. Define the honest-decoded padded set

$$\mathcal{C}_\alpha^{\mathrm{dec},\mathcal{H}}(x) = \left\{ y : \frac{1}{h} \sum_{i \in \mathcal{H}} w_i(x, y) \le T_k(\mathcal{H}; V) + g_\rho \right\}.$$

Then all three proposed rules equal $\mathcal{C}_\alpha^{\mathrm{dec},\mathcal{H}}(x)$ pathwise. If, in addition, the registered reconstruction errors are zero, then

$$\mathcal{C}_\alpha^{\mathrm{fs}}(x) = \mathcal{C}_\alpha^{\mathrm{jt}}(x) = \mathcal{C}_\alpha^{\mathrm{del}}(x) = \mathcal{C}_\alpha^{\circ}(x)$$

pathwise. If the $n + 1$ clean true-label scores are exchangeable and almost surely distinct, their common coverage is exactly $k/(n + 1)$ for $k \le n$.

Finally, for a fixed transcript, fixed directional padding, and $0 \le A_1 \le A_2 < K$, each of the three proposed sets at budget A_1 is contained in its counterpart at budget A_2.

Interpretation. High coverage under high–low can coexist with an uninformative all-answer set: this attack tests efficiency, not resistance to undercoverage.

The final monotonicity makes the tradeoff in choosing A explicit. In the primary no-dropout protocol the additive padding g_ρ is independent of A, but conservative budgets still enlarge the feasible-set family and widen the deletion aggregates. Underestimating A invalidates all three guarantees when $a > A$; no universal fragility ordering against the symmetric rule or p-merger follows without a distribution and attack model.

F Symmetric comparison and its limitations

We first define the guarded symmetric comparator, whose guard is an additive threshold adjustment, then state precisely when all proposed sets are contained in it. Take $A \le m$, $2m < K$, symmetric reconstruction error $\rho_r = \rho_r^- = \rho_r^+$, and define

$$\widehat{R}_j^{\mathrm{sym,m}} = \mathrm{TM}_m(\widetilde{s}_{1j}^{\mathrm{cal}}(Y_j), \ldots, \widetilde{s}_{Kj}^{\mathrm{cal}}(Y_j)),$$
$$\widehat{s}^{\mathrm{sym,m}}(x, y) = \mathrm{TM}_m(\widetilde{s}_1^{\mathrm{qry}}(x, y), \ldots, \widetilde{s}_K^{\mathrm{qry}}(x, y)).$$

For $k \le n$, let $\widehat{R}_{(k)}^{\mathrm{sym,m}}$ be the kth order statistic of $\widehat{R}_1^{\mathrm{sym,m}}, \ldots, \widehat{R}_n^{\mathrm{sym,m}}$. For this comparison, take D_r to be the full decoded-alphabet width, the width of the range of values a report can decode to, and set $g_r(m) = mD_r/(K - A)$. For $k \le n$, the comparator uses threshold $\widehat{R}_{(k)}^{\mathrm{sym,m}} + g_{\mathrm{cal}}(m) +$

$g_{\mathrm{qry}}(m) + \rho_{\mathrm{cal}} + \rho_{\mathrm{qry}}$ and prediction set

$$\mathcal{C}_\alpha^{\mathrm{sym,m}}(x) = \left\{ y : \widehat{s}^{\mathrm{sym,m}}(x,y) \le \widehat{R}_{(k)}^{\mathrm{sym,m}} + g_{\mathrm{cal}}(m) + g_{\mathrm{qry}}(m) + \rho_{\mathrm{cal}} + \rho_{\mathrm{qry}} \right\}.$$

When $k = n+1$, the comparator, like the proposed rules, returns $\mathcal{Y}$. Write $\mathcal{C}^{\mathrm{sym}} = \mathcal{C}^{\mathrm{sym,A}}$ for the primary comparator.

Proposition F.1 (Dominance over every eligible full-alphabet symmetric guard)**.** *Suppose $A \le m$, $2m < K$, all methods use the same reports and symmetric quantization bounds, and $g_r(m) = mD_r/(K-A)$ is computed from the full decoded-alphabet width. Then*

$$\mathcal{C}_\alpha^{\mathrm{fs}}(x) \subseteq \mathcal{C}_\alpha^{\mathrm{jt}}(x) \subseteq \mathcal{C}_\alpha^{\mathrm{del}}(x) \subseteq \mathcal{C}_\alpha^{\mathrm{sym,m}}(x) \quad \textit{for every } x$$

pathwise. Over-trimming itself does not break the result when its guard grows with m. The claim can fail if an $m > A$ comparator retains the undersized $AD_r/(K-A)$ guard or uses genuinely tighter certified honest-range information; it is not an unconditional dominance statement. For example, with $K = 5$, $A = 1$, $m = 2$, $n = k = 1$ (so $\alpha \ge 1/2$), zero reconstruction error, reports in $[0,1]$, calibration reports $v = (0,0,0,1,1)$ and query reports $w = (0,0,1,1,1)$, deletion keeps the candidate because $\mathrm{L}_A(w) = \mathrm{U}_A(v) = 1/2$, whereas the comparator with the undersized guard $2AD_r/(K-A) = 1/2$ drops it, since $\mathrm{TM}_2(w) = 1 > \mathrm{TM}_2(v) + 1/2 = 1/2$.

The guard is needed because trimming both tails moves the aggregate away from the honest mean when honest nodes disagree, even without an attack. For exact honest score values x_i, write $R_{\mathcal{H}} = \max_{i\in\mathcal{H}} x_i - \min_{i\in\mathcal{H}} x_i$ and $\bar{x}_{\mathcal{H}} = h^{-1}\sum_{i\in\mathcal{H}} x_i$, and write $(x_{\mathcal{H}}, v_{\mathcal{A}})$ for the report vector with these honest entries and Byzantine entries $v_{\mathcal{A}}$.

Lemma F.2 (Sharp contaminated symmetric trimmed mean)**.** *If $a \le A \le m$ and $2m < K$, then for arbitrary finite Byzantine reports*

$$|\mathrm{TM}_m(x_{\mathcal{H}}, v_{\mathcal{A}}) - \bar{x}_{\mathcal{H}}| \le \frac{m}{h} R_{\mathcal{H}}.$$

The attacked trimmed mean lies between the means of the smallest and largest $h-m$ honest values, and the coefficient m/h is worst-case sharp, even when $a = 0$. Deterministic symmetric quantization adds at most ρ_r.

Thus, with a registered bound $\bar{R}_r$ on the honest score range, the symmetric comparator is valid with per-channel guard $m\bar{R}_r/(K-A) + \rho_r$, which $m = A$ minimizes; the $m > A$ sweep is a conservatism diagnostic. Phase switching makes some padding unavoidable for this rule:

Proposition F.3 (Low–high coverage lower bound)**.** *Let $a = A$, $2A < K$, $h = K - A$, and suppose $k \le n$. Let honest scores have the location form $s_i(Z_j) = S_j + \xi_i$: all honest nodes share the example-specific score S_j, and ξ_i is node i's fixed offset. The offsets have range R and the S_j are i.i.d. continuous. Let $S_{(k)}$ be the kth order statistic of $S_1, \ldots, S_n$. Write $\xi_{(1)} \le \cdots \le \xi_{(h)}$ for the ordered honest offsets. Assume exact honest reports and symmetric trim $m = A$. Let Byzantine reports lie below every honest report in calibration and above every honest report at prediction. Define*

$$L_\xi = \frac{1}{h-A}\sum_{i=1}^{h-A} \xi_{(i)}, \qquad U_\xi = \frac{1}{h-A}\sum_{i=A+1}^{h} \xi_{(i)}, \qquad \Delta_\xi = U_\xi - L_\xi.$$

For a symmetric-trimmed threshold with total padding g, coverage equals

$$\mathbb{P}\{S_{n+1} \le S_{(k)} - (\Delta_\xi - g)\}. \tag{16}$$

For every $g < \Delta_\xi$, a bounded uniform law for S with support width less than $\Delta_\xi - g$ makes this probability zero. Moreover,

$$\sup_{\mathrm{range}(\xi)\le R} \Delta_\xi = R \min\left\{1, \frac{A}{K-2A}\right\}. \tag{17}$$

Hence any fixed total-padding constant uniformly valid over this subclass must be at least the right-hand side of Equation (17)*. This lower bound is generally smaller than the sufficient $2AR/(K-A)$ guard.*

Switching from low calibration reports to high query reports changes which honest offsets survive trimming, and a narrow clean-score law cannot absorb the gap. The bound concerns this location family and this symmetric rule; it imposes no padding on fixed-set inference.

G MATCHED PIPELINES AND THE ROBUST p-VALUE ALTERNATIVE

Lemma G.1 (A static phase-blind map preserves exchangeability)**.** *Let $\mathcal{F}$ contain all state frozen before calibration, and let Ξ_j represent any additional randomness used on example j. Conditional on $\mathcal{F}$, suppose $Z_1, \dots, Z_{n+1}$ are exchangeable, $\Xi_1, \dots, \Xi_{n+1}$ are i.i.d. and independent of the examples, and one fixed measurable map satisfies*

$$T_j = \phi(Z_j, \Xi_j; \mathcal{F}), \qquad j \in [n+1].$$

Then $T_1, \dots, T_{n+1}$ are exchangeable conditional on $\mathcal{F}$. Consequently, an unpadded split-conformal rule applied to the same attacked aggregate in both phases has coverage at least $k/(n+1)$.

A sufficient condition for the lemma's single fixed map is static, example-independent node state; matched candidates, information exposure, bit depths and reconstruction maps; identical packet handling, tie rules and aggregation; and a fixed phase-blind Byzantine map with only i.i.d. per-example randomness. The lemma explains unguarded coverage under stable attacks but says nothing about the honest-mean score and does not cover switching attacks; deletion's validity comes from Theorem 4.1.

An honest-mean-free alternative uses node-wise conformal p-values, which the hub computes from the same aligned calibration reports:

$$p_i(x, y) = \frac{1 + \sum_{j=1}^{n} \mathbf{1}\{\widetilde{s}_{ij}^{\mathrm{cal}}(Y_j) \geq \widetilde{s}_i^{\mathrm{qry}}(x, y)\}}{n+1} \tag{18}$$

This is the usual unsmoothed conformal p-value. A p-value is *superuniform* if $\mathbb{P}\{p_i(X, Y) \leq t\} \leq t$ for every $t \in [0, 1]$: an honest node reports a small p-value for the true answer no more often than its nominal level. For honest nodes whose scoring transformation and quantizer are matched across phases and whose calibration/future pairs are exchangeable for the same target law, applying the exchangeable-rank argument to the negated scores proves marginal superuniformity, with the weak comparison making ties conservative. Matching the transformation alone does not establish this common-target condition. Shared calibration induces dependence among these p-values, which the next result permits.

Proposition G.2 (Partial-conjunction/Rüger merger with up to A arbitrary reports)**.** *Let at most A of K reported p-values be arbitrary, and suppose every honest true-label value $p_i(X, Y)$ is marginally superuniform; no dependence assumption is made. For any prespecified $\ell \in \{1, \dots, K - A\}$, sort all reports increasingly, writing $p_{(r)}$ for the rth smallest reported value, and define*

$$p_\ell^{\mathrm{merge}} = \min\left\{1, \frac{K - A}{\ell} p_{(A+\ell)}\right\}.$$

Then p_ℓ^{merge} is superuniform, and the factor $(K - A)/\ell$ is sharp under arbitrary honest dependence. In particular, if $2A < K$, the prespecified majority-style choice $\ell = A + 1$ includes candidate y exactly when

$$p_{(2A+1)}(x, y) > \frac{\alpha(A+1)}{K - A}. \tag{19}$$

The denominator $K - A$ is sharp; replacing it by K, as in a direct conservative counting bound, remains valid but lowers the inclusion cutoff and can produce larger prediction sets.

Each honest p-value must be valid for the common target, which federated conformal theory makes explicit through exchangeability or shift assumptions (Lu et al., 2023; Plassier et al., 2023; Barber et al., 2023).

H BIT BUDGETS AND CERTIFICATE-WIDTH AUDIT

Extra bits shrink the reconstruction terms of the padding and of the set-size width geometrically, but never the term due to unknown membership. Once divided among examples and candidates, the bit budget controls reconstruction error. This audit distinguishes the padding added by the algorithm, denoted G, from the extra clean-score threshold width in the set-size bound, denoted W; a large W does not itself imply a large realized prediction set. Suppose node i has B_i^{qry} bits per query and B_i^{cal}

bits for its complete n-example calibration transcript. With

$$b_i^{\mathrm{qry}} = \left\lfloor \frac{B_i^{\mathrm{qry}}}{M} \right\rfloor, \quad b_i^{\mathrm{cal}} = \left\lfloor \frac{B_i^{\mathrm{cal}}}{nM} \right\rfloor, \quad b_r = \min_i b_i^r,$$

and $b_r \geq 1$, the uniform quantizer yields

$$\rho_r = \frac{S_{\max}}{2(2^{b_r} - 1)}.$$

Across n examples, the vector protocol sends $nM \sum_i b_i^{\mathrm{cal}}$ score bits, while one test query sends $M \sum_i b_i^{\mathrm{qry}}$ score bits, excluding packet metadata; unused remainder bits are allowed, so these payloads fit within the registered budgets. The deletion threshold padding is $G_{\mathrm{del}} = \rho_{\mathrm{cal}} + \rho_{\mathrm{qry}}$, the padding g_ρ of Section 3.2 with symmetric errors. For a common decoded alphabet $[0, S_{\max}]$, its sandwich width, the outer-bound width W_{del} of Theorem 4.4 with $D_r = S_{\max}$, is

$$W_{\mathrm{del}} = 2 \min\left\{1, \frac{A}{K - A}\right\} S_{\max} + 2\rho_{\mathrm{cal}} + 2\rho_{\mathrm{qry}}. \tag{20}$$

The primary $m = A$ symmetric comparator instead has threshold padding $G_{\mathrm{sym}} = 2AS_{\max}/(K - A) + \rho_{\mathrm{cal}} + \rho_{\mathrm{qry}}$ and sandwich width $W_{\mathrm{sym}} = 2G_{\mathrm{sym}}$.

At the operational cell $(K, A, b_{\mathrm{cal}}, b_{\mathrm{qry}}) = (16, 2, 8, 8)$ with $S_{\max} = 1$, the prespecified certificate-width screen, which compares each G and W with the score range $S_{\max}$, is

$$G_{\mathrm{del}} = 0.00392, \qquad W_{\mathrm{del}} = 0.29356,$$
$$G_{\mathrm{sym}} = 0.28964, \qquad W_{\mathrm{sym}} = 0.57927,$$

and at $\alpha = 0.1$ the p-merger cutoff is $0.1 \times 3/14 = 0.02143$.

At the prespecified four-bit stress cell $(K, A, b_{\mathrm{cal}}, b_{\mathrm{qry}}) = (16, 3, 4, 4)$ with $S_{\max} = 1$ (defined in Appendix O),

$$G_{\mathrm{del}} = 0.06667, \qquad W_{\mathrm{del}} = 0.59487,$$
$$G_{\mathrm{sym}} = 0.52821, \qquad W_{\mathrm{sym}} = 1.05641.$$

The corresponding stress-cell p-merger cutoff is $0.1 \times 4/13 = 0.03077$. There the symmetric width exceeds the whole score range and its padding consumes more than half of it, while the deletion certificate stays narrower. These are design diagnostics, neither sufficient nor necessary for useful realized sets, so every cell also reports realized set size and full-set rate (the share of queries whose set contains all M candidates). The decoded-alphabet width in Equation (20) is the price of an upper efficiency certificate, not a validity guard.

I Notation and elementary order-statistic facts

This appendix fixes notation and elementary facts, and together with Appendices J to L and N it supplies complete proofs of every formal statement in the main paper; the restricted-family results are proved in Appendix R. It also fixes the tie and quantile conventions the proofs use. The proof sequence follows the argument of the paper: Appendix J explains why node-wise marginals are insufficient, Appendix K derives the score envelopes, and Appendix L turns those envelopes into coverage and efficiency guarantees. Readers interested in implementation can begin with Appendix M; the completed study's design and results appear in Appendices O and P. Appendix Q collects the assumptions needed for each guarantee, and Table 3 lists the principal symbols.

Table 3: Principal notation, roughly in the order the main text introduces it, followed by symbols used only in the appendix.

Symbol	Meaning	Symbol	Meaning
K, $[K]$	number of nodes; $\{1, \dots, K\}$	$V = (v_{ij})$	decoded calibration reports on the correct answer
$\mathcal{A}$, $\mathcal{H}$	Byzantine and honest node sets	$w(x, y)$	decoded query reports for candidate y
$a = \lvert\mathcal{A}\rvert$, A	realized and declared corruption	$\mathfrak{H}_A$	feasible honest subsets, $\lvert H\rvert \geq K - A$
$h = \lvert\mathcal{H}\rvert$	number of honest nodes	$c_j(H; V)$	mean report of H on example j
$\mathcal{Y}$, M	answer space and its size	$T_k(H; V)$	kth smallest of $c_1, \dots, c_n$
s_i, $s_{\mathcal{H}}$	node score; honest-mean score	$\tau_{A,k}$	largest $T_k(H; V)$ over $\mathfrak{H}_A$
$Q_{i,r}$	quantizer on channel r	$\Delta_{A,k}$	fixed-set margin (7)
ρ_r^-, ρ_r^+	directional error bounds	U_A, L_A	means of the largest / smallest $K - A$
g_ρ	padding $\rho_{\mathrm{cal}}^- + \rho_{\mathrm{qry}}^+$	$\mathcal{C}^\circ, \mathcal{C}^{\mathrm{fs}}, \mathcal{C}^{\mathrm{jt}}, \mathcal{C}^{\mathrm{del}}$	oracle and proposed sets
n, α, k	calibration size, level, $\lceil (n+1)(1-\alpha) \rceil$	D_r, W_{del}	report width; outer bound width
R_j, $R_{(k)}$	clean calibration scores; kth smallest	ε_A	$\mathbb{P}\{a > A\}$ under a failure law
$Z_j = (X_j, Y_j)$	example j; $j = n+1$ is the future question	os_k	kth smallest of n values
$\mathcal{X}$	input space	$\widehat{R}_j$, $\widehat{q}_\alpha^{\mathrm{del}}$	$\mathrm{U}_A(v_{\cdot j})$; deletion threshold (11)
$\bar{v}_H$	mean of v over the identities in H	μ_r	membership term $\min\{1, A/(K-A)\} D_r$
lo, hi	ends of the decodable report range	m, TM_m	symmetric trim count; mean after dropping m per tail
Appendix only			
$S_{\max}$	score bound, $s_i \in [0, S_{\max}]$	$R_{\mathcal{H}}$	honest range, $\max_{i \in \mathcal{H}} s_i - \min_{i \in \mathcal{H}} s_i$
D	honest-dropout budget (not D_r)	ρ_r	symmetric error bound, $\rho_r^- = \rho_r^+$

Terms used throughout. Italicized terms are defined where first used in Sections 1 to 5.

Throughout, repeated values are counted separately in every multiset. Thus ties do not make the trimmed mean ambiguous: any permutation of equal values has the same sum over the retained middle block.

For a vector $u = (u_1, \dots, u_N)$, write $u_{(1)} \leq \cdots \leq u_{(N)}$ for its order statistics. We repeatedly use the following elementary fact: moving every coordinate by at most η moves every ordered value, and hence every trimmed mean, by at most η.

Lemma I.1 (Sup-norm stability of order statistics). *If $u, v \in \mathbb{R}^N$ and $\max_i \lvert u_i - v_i\rvert \leq \eta$, then*

$$\lvert u_{(j)} - v_{(j)}\rvert \leq \eta, \qquad j = 1, \dots, N.$$

Consequently, for $2m < N$,

$$\lvert \mathrm{TM}_m(u) - \mathrm{TM}_m(v)\rvert \leq \eta.$$

Proof. At least j coordinates of u are no larger than $u_{(j)}$. The corresponding coordinates of v are no larger than $u_{(j)} + \eta$, so $v_{(j)} \leq u_{(j)} + \eta$. Interchanging u and v gives $u_{(j)} \leq v_{(j)} + \eta$. Averaging the inequalities over the retained indices $j = m+1, \dots, N-m$ proves the trimmed-mean statement. □

Quantiles remain well-defined when score distributions have ties. We use the generalized inverse $F^{-1}(p) = \inf\{t : F(t) \geq p\}$, the smallest threshold at which the CDF reaches p. All score functions, corpora, model weights, prompts, and the honest set may be random before calibration. Every probability statement is conditional on the information (formally, the sigma-field) that fixes them. We suppress that conditioning to keep notation readable.

J PROOF OF THE NONIDENTIFICATION RESULT

The proof constructs two systems with identical local score distributions but different distributions for their mean. A procedure that sees only the local distributions cannot distinguish the systems, which yields both the distribution-estimation lower bound and the threshold example.

Proof of Proposition B.1. Let $U \sim \mathrm{Unif}[0,1]$ and define the continuous tent map

$$T(u) = \begin{cases} 2u, & 0 \le u \le 1/2, \\ 2 - 2u, & 1/2 < u \le 1. \end{cases}$$

For $t \in [0,1]$,

$$\mathbb{P}\{T(U) \le t\} = \mathbb{P}\{U \le t/2\} + \mathbb{P}\{U \ge 1 - t/2\} = t.$$

Thus $T(U) \sim \mathrm{Unif}[0,1]$.

In World 0 take $(s_1(U), s_2(U)) = (U, U)$. Its average $M_0 = U$ has CDF $F_0(t) = t$ on $[0,1]$. In World 1 take $(s_1(U), s_2(U)) = (U, T(U))$. Both local marginals are again uniform, while

$$M_1 = \frac{U + T(U)}{2} = \begin{cases} 3U/2, & U \le 1/2, \\ 1 - U/2, & U > 1/2. \end{cases}$$

For $0 \le t < 1/2$, only the first branch can be no larger than t, so

$$F_1(t) = \mathbb{P}\{U \le 2t/3\} = \frac{2t}{3}.$$

For $1/2 \le t < 3/4$, both branches contribute and their preimages are disjoint:

$$\begin{aligned} F_1(t) &= \mathbb{P}\{U \le 2t/3\} + \mathbb{P}\{U > 1/2,\ 1 - U/2 \le t\} \\ &= \frac{2t}{3} + \mathbb{P}\{U \ge 2(1-t)\} \\ &= \frac{8t}{3} - 1. \end{aligned}$$

Therefore

$$F_1(t) = \begin{cases} 0, & t < 0, \\ 2t/3, & 0 \le t < 1/2, \\ 8t/3 - 1, & 1/2 \le t < 3/4, \\ 1, & t \ge 3/4. \end{cases}$$

On $[0, 1/2)$ the absolute gap from F_0 is $t/3$, whose supremum is $1/6$. On $[1/2, 3/4)$ the signed gap is $F_1(t) - F_0(t) = 5t/3 - 1$, increasing from $-1/6$ to $1/4$. On $[3/4, 1]$ it is $1 - t$, maximized at $1/4$. Hence $\|F_0 - F_1\|_\infty = 1/4$.

The population input to a marginal-only estimator is the same pair of uniform CDFs in both worlds. Consequently $\widehat{F}$ has the same distribution in both. For every realization, the triangle inequality gives

$$\frac{1}{4} = \|F_0 - F_1\|_\infty \le \|\widehat{F} - F_0\|_\infty + \|\widehat{F} - F_1\|_\infty.$$

Take expectation with respect to the estimator's randomization and use $\max(x, y) \ge (x + y)/2$ to obtain the lower bound $1/8$.

Finally, $F_0^{-1}(3/4) = 3/4$. Solving $8t/3 - 1 = 3/4$ gives $F_1^{-1}(3/4) = 21/32$. Any single deterministic threshold with coverage at least $3/4$ in World 0 must be at least $3/4$. Since $M_1 \le 3/4$ almost surely, that threshold has coverage one in World 1, which is overcoverage $1/4$ relative to the target. This completes the proof. □

Remark J.1 (What the proposition does and does not rule out). The proposition applies to protocols whose population information consists only of the separate local CDFs, including arbitrarily fine histograms and exact local quantiles. It does not rule out a protocol that retains paired score vectors on shared examples, transmits cross-moments under a correctly specified copula model, or defines its deployment score to be a different functional whose law is identified from the transmitted summaries. The aligned protocol takes the first route.

K PROOFS FOR DELETION AND SYMMETRIC AGGREGATION

The first proof finds the largest and smallest honest means compatible with a reported score vector and the corruption budget. These envelopes support coverage even when Byzantine reports are unbounded. The subsequent bounded-report and symmetric-trimming results quantify how far an aggregate can lie from the clean honest mean; their additional assumptions are stated separately.

Theorem K.1 (Exact deletion envelopes and pointwise optimality)**.** *For every $v \in \mathbb{R}^K$, write $\bar{v}_H = |H|^{-1} \sum_{i \in H} v_i$. Then*

$$\mathrm{U}_A(v) = \max_{H \in \mathfrak{H}_A} \bar{v}_H, \qquad \mathrm{L}_A(v) = \min_{H \in \mathfrak{H}_A} \bar{v}_H, \tag{21}$$

so that, if at most A coordinates are Byzantine, exact honest reports satisfy

$$\mathrm{L}_A(v) \le \bar{v}_{\mathcal{H}} \le \mathrm{U}_A(v). \tag{22}$$

Moreover, U_A is the pointwise smallest function that upper-bounds every feasible subset mean in Equation (21)*, and L_A is the pointwise largest corresponding lower bound. Writing $s_i(z) = s_i(x, y)$ and $\widetilde{s}^r(z) = (\widetilde{s}^r_1(z), \dots, \widetilde{s}^r_K(z))$ for $z = (x, y)$, if honest reports on channel r obey $-\rho^-_r \le \widetilde{s}^r_i(z) - s_i(z) \le \rho^+_r$, then, uniformly over Byzantine reports,*

$$s_{\mathcal{H}}(z) \le \mathrm{U}_A(\widetilde{s}^r(z)) + \rho^-_r, \qquad \mathrm{L}_A(\widetilde{s}^r(z)) \le s_{\mathcal{H}}(z) + \rho^+_r. \tag{23}$$

The two bounds are separately pointwise sharp when honest reconstructions are constrained only by the common directional-error intervals.

Equality can fail for a particular fixed reconstruction map, which also encodes the clipping range and half-open quantizer cells; at eight bits with $S_{\max} = 1$ the bound exceeds the value attainable under that map by at most about $0.002/|H|$ per clipped entry.

Proof. We first optimize over honest subset size, then account for reconstruction error. The equality constructions also establish pointwise sharpness. Fix a subset size s. The largest mean over all s-subsets is the mean of the largest s coordinates, and the smallest is the mean of the smallest s. The top-s mean is nonincreasing in s: adding a coordinate no larger than the current top-s mean cannot increase it. The bottom-s mean is nondecreasing by the symmetric argument. Hence both extrema over $s \ge K - A$ occur at $s = K - A$, proving Equation (21).

The true honest set has $h = K - a \ge K - A$ coordinates, so its reported mean is one of the feasible subset means. This proves Equation (22). If a function $U(v)$ upper-bounds every feasible subset mean, choose the $K - A$ indices of the largest reports to get $U(v) \ge \mathrm{U}_A(v)$. Thus U_A is pointwise smallest. Choosing the bottom $K - A$ indices proves the lower statement.

Let q_i be the decoded honest report and x_i its clean score, and write $(q_{\mathcal{H}}, v_{\mathcal{A}})$ for the full report vector. From $q_i - x_i \ge -\rho^-_r$,

$$\bar{x}_{\mathcal{H}} \le \bar{q}_{\mathcal{H}} + \rho^-_r \le \mathrm{U}_A(q_{\mathcal{H}}, v_{\mathcal{A}}) + \rho^-_r.$$

Similarly, $q_i - x_i \le \rho^+_r$ gives

$$\mathrm{L}_A(q_{\mathcal{H}}, v_{\mathcal{A}}) \le \bar{q}_{\mathcal{H}} \le \bar{x}_{\mathcal{H}} + \rho^+_r.$$

These are Equation (23). In the interval model, where each honest clean score is known only to lie in its common directional-error interval, equality in the upper robust envelope is obtained by choosing the top $K - A$ coordinates as honest and taking their clean values to be $q_i + \rho^-_r$. The lower construction is symmetric. This establishes the stated pointwise sharpness in the interval model. The construction need not be attainable by a particular fixed reconstruction map, whose clipping range and half-open cells the interval model ignores. □

The next elementary bound is used only for efficiency, not validity.

Lemma K.2 (Two-sided deviations under bounded reports)**.** *Let x_i be honest node i's clean score and q_i its decoded report, write $\bar{x}_{\mathcal{H}}$ for the clean honest mean and $(q_{\mathcal{H}}, v_{\mathcal{A}})$ for the report vector with honest coordinates q_i and arbitrary Byzantine coordinates v_i. Suppose all decoded reports lie in an interval of width B (the D_r of Theorem 4.4 on one channel), $a \le A$, and honest reconstruction errors obey $-\rho^- \le q_i - x_i \le \rho^+$. Then*

$$-\rho^- \le \mathrm{U}_A(q_{\mathcal{H}}, v_{\mathcal{A}}) - \bar{x}_{\mathcal{H}} \le \min\left\{1, \frac{A}{h}\right\} B + \rho^+, \tag{24}$$

$$-\min\left\{1, \frac{A}{h}\right\} B - \rho^- \le \mathrm{L}_A(q_{\mathcal{H}}, v_{\mathcal{A}}) - \bar{x}_{\mathcal{H}} \le \rho^+. \tag{25}$$

The coefficient $\min\{1, A/h\}$ *is worst-case sharp.*

Proof. The inner inequalities $\mathrm{U}_A \ge \bar{q}_{\mathcal{H}}$ and $\mathrm{L}_A \le \bar{q}_{\mathcal{H}}$ follow from Theorem K.1, proving the left side of Equation (24) and the right side of Equation (25). It remains to show $\mathrm{U}_A - \bar{q}_{\mathcal{H}} \le \min\{1, A/h\}B$; the lower claim follows after negating all reports.

If $B = 0$, the report-deviation claim is immediate. Assume $B > 0$ and by translation and scaling take the report interval to be $[0, 1]$. For any fixed honest set, the function $v \mapsto \mathrm{U}_A(v) - \bar{v}_{\mathcal{H}}$ is the maximum of finitely many linear functions, so its maximum over the cube is attained at a vertex. Suppose a binary vertex has u honest ones and $b \le a$ Byzantine ones. Put $\bar{h} = K - A$, the smallest possible honest count. Then

$$\mathrm{U}_A(v) - \bar{v}_{\mathcal{H}} = \frac{\min\{u + b, \bar{h}\}}{\bar{h}} - \frac{u}{h}.$$

If $u + b \ge \bar{h}$, the first term is one and $u \ge \bar{h} - b$, so the difference is at most $(h - \bar{h} + b)/h = (A - a + b)/h \le A/h$. If $u + b < \bar{h}$, the expression is $(u + b)/\bar{h} - u/h$. It is nondecreasing in u because $h \ge \bar{h}$, so it is bounded above by its value at the adjacent boundary $u = \bar{h} - b$, which obeys the same bound. The difference between two means in $[0, 1]$ is also at most one, proving the minimum of the two bounds. Rescaling restores B. Finally, quantization can move the honest reported mean above the clean mean by at most ρ^+, proving the upper inequality. If $h \ge A$, put A honest reports at the lower endpoint, all other honest reports at the upper endpoint, and all Byzantine reports at the upper endpoint. If $h < A$, put all honest reports at the lower endpoint and at least $K - A$ Byzantine reports at the upper endpoint. These arrays attain the displayed coefficient; negation gives the lower equality. □

The symmetric comparator is controlled by how far extra reports can shift the ranks of honest values. Let $x^{\mathcal{H}}_{(1)} \le \cdots \le x^{\mathcal{H}}_{(h)}$ and define

$$L_{\mathcal{H},m} = \frac{1}{h - m} \sum_{j=1}^{h-m} x^{\mathcal{H}}_{(j)}, \qquad U_{\mathcal{H},m} = \frac{1}{h - m} \sum_{j=m+1}^{h} x^{\mathcal{H}}_{(j)}.$$

Lemma K.3 (Symmetric interlacing sandwich)**.** *If* $a \le A \le m$ *and* $2m < K$*, then*

$$L_{\mathcal{H},m} \le \mathrm{TM}_m(x_{\mathcal{H}}, v_{\mathcal{A}}) \le U_{\mathcal{H},m}.$$

Proof. Let $w_{(1)} \le \cdots \le w_{(K)}$ order all reports. Adding a values to the honest multiset gives $x^{\mathcal{H}}_{(j-a)} \le w_{(j)} \le x^{\mathcal{H}}_{(j)}$ whenever the displayed honest indices exist. Put $L = K - 2m$. For retained index $j = m + r$, $r = 1, \ldots, L$, averaging the interlacing inequalities yields

$$\frac{1}{L} \sum_{j=m+1-a}^{h-m} x^{\mathcal{H}}_{(j)} \le \mathrm{TM}_m(w) \le \frac{1}{L} \sum_{j=m+1}^{K-m} x^{\mathcal{H}}_{(j)}.$$

Since $h - m = L + (m - a)$, the left sum contains the largest L values among the first $h - m$ honest order statistics, so its mean is at least $L_{\mathcal{H},m}$. The upper sum contains the smallest L values among the last $h - m$ honest order statistics, so its mean is at most $U_{\mathcal{H},m}$. □

Proof of Lemma F.2. By Lemma K.3, it suffices to bound the two honest deletion means. If $m = 0$ the claim is immediate. Let B_m be the mean of the m smallest honest values. Partitioning the honest sum gives

$$\bar{x}_{\mathcal{H}} = \frac{h - m}{h} U_{\mathcal{H},m} + \frac{m}{h} B_m,$$

and hence $0 \le U_{\mathcal{H},m} - \bar{x}_{\mathcal{H}} \le (m/h)R_{\mathcal{H}}$. Negating the values gives $0 \le \bar{x}_{\mathcal{H}} - L_{\mathcal{H},m} \le (m/h)R_{\mathcal{H}}$.

For sharpness at any admissible a, take exactly m honest values equal to zero, all other honest values equal to one, and all Byzantine reports equal to one. The retained mean is one while the honest mean is $1 - m/h$. Thus the coefficient is attained for every admissible a, including $a = 0$. If honest reports are perturbed by at most ρ, keep Byzantine values fixed, invoke Lemma I.1, and add ρ by the triangle inequality. □

L PROOFS OF VALIDITY, EFFICIENCY, AND LOWER BOUNDS

Exchangeability lower-bounds the probability of the clean rank event, and the score envelopes show pathwise that this event implies inclusion.

Lemma L.1 (Exchangeable rank lemma)**.** *Let $T_1, \dots, T_{n+1}$ be exchangeable real random variables and $k \in \{1, \dots, n\}$. If $T^{\mathrm{cal}}_{(k)}$ is the kth order statistic of $T_1, \dots, T_n$, then*

$$\mathbb{P}\{T_{n+1} \le T^{\mathrm{cal}}_{(k)}\} \ge \frac{k}{n+1}.$$

Proof. Let $\zeta_1, \dots, \zeta_{n+1}$ be i.i.d. $\mathrm{Unif}[0,1]$ tie-breakers independent of the T_j and order (T_j, ζ_j) lexicographically. The rank of pair $n+1$ is uniform. A rank at most k implies $T_{n+1} \le T^{\mathrm{cal}}_{(k)}$ after discarding the tie-breakers, proving the inequality. □

Proof of Theorem 4.1. If $k = n+1$, the returned set is $\mathcal{Y}$. Suppose $k \le n$. The calibration part of Equation (23) gives $R_j \le \widehat{R}_j + \rho^-_{\mathrm{cal}}$ and therefore

$$R_{(k)} \le \widehat{R}_{(k)} + \rho^-_{\mathrm{cal}}.$$

On $\{R_{n+1} \le R_{(k)}\}$, the query envelope gives

$$\begin{aligned} \widehat{s}^{\mathrm{qry}}_{\mathrm{del}}(X_{n+1}, Y_{n+1}) &\le R_{n+1} + \rho^+_{\mathrm{qry}} \\ &\le R_{(k)} + \rho^+_{\mathrm{qry}} \\ &\le \widehat{R}_{(k)} + \rho^-_{\mathrm{cal}} + \rho^+_{\mathrm{qry}} = \widehat{q}^{\mathrm{del}}_{\alpha}. \end{aligned}$$

Thus the clean rank event implies coverage pathwise. Apply Lemma L.1 to $R_1, \dots, R_{n+1}$. No property of the Byzantine transcript beyond finite decoding was used. □

L.1 FIXED-MEMBERSHIP IDENTIFIED RULES

We identify the largest compatible calibration threshold, then reuse one feasible honest subset across calibration and query entries to obtain the fixed-set rule, its nesting, coverage, and minimality among rules that preserve the clean rank event in every entrywise compatible completion.

Lemma L.2 (Sharp joint calibration envelope)**.** *Assume the conditions of Theorem 4.1 and let $k \le n$. In the interval model of Theorem 4.3,*

$$\sup_{\substack{H \in \mathfrak{H}_A,\ (x_{ij}) \\ -\rho^-_{\mathrm{cal}} \le v_{ij} - x_{ij} \le \rho^+_{\mathrm{cal}}\ (i \in H,\, j \in [n])}} \mathrm{os}_k \left\{ \frac{1}{|H|} \sum_{i \in H} x_{ij} : j \in [n] \right\} = \tau_{A,k}(V) + \rho^-_{\mathrm{cal}}. \tag{26}$$

Hence the right side upper-bounds the clean calibration kth order statistic for every feasible fixed honest set and compatible clean-score array, and it is the smallest such scalar bound based only on V, the declared budget, and the reconstruction bounds.

Proof. Fix $k \le n$. For any $H \in \mathfrak{H}_A$ and compatible clean calibration array (x_{ij}), the directional interval gives

$$\frac{1}{|H|} \sum_{i \in H} x_{ij} \le c_j(H; V) + \rho^-_{\mathrm{cal}} \quad \text{for every } j.$$

Order-statistic monotonicity therefore bounds the left side of Equation (26) by $T_k(H; V) + \rho^-_{\mathrm{cal}} \le \tau_{A,k}(V) + \rho^-_{\mathrm{cal}}$. Because $\mathfrak{H}_A$ is finite, choose a maximizer $H^\star$ and set $x_{ij} = v_{ij} + \rho^-_{\mathrm{cal}}$ for $i \in H^\star$. These values satisfy the abstract directional intervals and attain the upper bound. This proves the identity. Any scalar upper bound on the clean order statistic that is valid in every entrywise compatible completion must also bound this maximizing completion. Thus no smaller bound based only on V, the declared budget, and the reconstruction bounds can have the required property. □

Proof of Theorem 4.3. Fix $k \le n$ and let $\mathcal{H}$ be the actual honest set. Its calibration values obey

$$T_k(\mathcal{H}; V) \ge R_{(k)} - \rho^-_{\mathrm{cal}}, \qquad \frac{1}{h} \sum_{i \in \mathcal{H}} w_i(x, y) \le s_{\mathcal{H}}(x, y) + \rho^+_{\mathrm{qry}}.$$

Thus $s_{\mathcal{H}}(x, y) \le R_{(k)}$ implies that the margin in Equation (7) at $H = \mathcal{H}$ is at most g_ρ, proving $\mathcal{C}^\circ \subseteq \mathcal{C}^{\mathrm{fs}}$.

For every feasible H,

$$\frac{1}{|H|}\sum_{i\in H} w_i \ge \mathrm{L}_A(w), \qquad T_k(H;V) \le \tau_{A,k}(V).$$

Consequently

$$\Delta_{A,k}(V,w) \ge \mathrm{L}_A(w) - \tau_{A,k}(V),$$

and membership in $\mathcal{C}^{\mathrm{fs}}$ implies membership in $\mathcal{C}^{\mathrm{jt}}$. For every H and j, Theorem K.1 gives $c_j(H;V) \le \mathrm{U}_A(v_{\cdot j})$. Applying order-statistic monotonicity and then maximizing over H yields

$$\tau_{A,k}(V) \le \mathrm{os}_k\{\mathrm{U}_A(v_{\cdot 1}), \dots, \mathrm{U}_A(v_{\cdot n})\} = \widehat{R}_{(k)},$$

so $\mathcal{C}^{\mathrm{jt}} \subseteq \mathcal{C}^{\mathrm{del}}$. This proves the hierarchy. On the clean true-label rank event, its leftmost inclusion gives membership in all three deployable sets. Applying Lemma L.1 proves their coverage.

It remains to prove the entrywise compatibility characterization. Fix H. Every compatible clean query array (z_i) satisfies

$$\frac{1}{|H|}\sum_{i\in H} z_i \ge \frac{1}{|H|}\sum_{i\in H} w_i - \rho^{+}_{\mathrm{qry}},$$

whereas the compatible clean calibration order statistic is at most $T_k(H;V) + \rho^{-}_{\mathrm{cal}}$. Hence the clean rank event can hold only if

$$\frac{1}{|H|}\sum_{i\in H} w_i - T_k(H;V) \le g_\rho. \tag{27}$$

Conversely, when Equation (27) holds, set $x_{ij} = v_{ij} + \rho^{-}_{\mathrm{cal}}$ and $z_i = w_i - \rho^{+}_{\mathrm{qry}}$ for every $i \in H$. These values are compatible and make the clean rank inequality exactly equivalent to Equation (27). Taking the union over $H \in \mathfrak{H}_A$ proves Equation (14). A (V, w)-measurable report-only rule that must preserve the clean rank event in every entrywise compatible completion must include a candidate whenever any such completion exists; the fixed-set rule includes exactly those candidates. This proves the stated pointwise minimality. The $k = n + 1$ case follows from the common full-set convention. □

Why the inclusions can be strict. The following examples isolate the information lost at each relaxation. At zero padding, let $K = 3, A = 1, n = 2, k = 1$, which needs $\alpha \ge 2/3$; these examples isolate mechanics, not a practical level. With rows indexed by nodes, the matrix

$$V = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$$

has $\tau_{A,k} = 2/3$, whereas the coordinatewise deletion order statistic is 1; a constant query vector with entries $3/4$ makes $\mathcal{C}^{\mathrm{jt}} \subsetneq \mathcal{C}^{\mathrm{del}}$. For

$$V = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad w = (0, 1, 1),$$

one has $\tau_{A,k} = \mathrm{L}_A(w) = 1/2$, but every feasible same-set margin is positive (its minimum is $1/3$), so $\mathcal{C}^{\mathrm{fs}} \subsetneq \mathcal{C}^{\mathrm{jt}}$. Finally, with $K = 2, A = 1, n = k = 1$ (so $\alpha \ge 1/2$), actual honest set $\{1\}$, calibration reports $(0, 1)$, and query reports $(1, 0)$, the candidate is excluded by $\mathcal{C}^\circ$ but included by $\mathcal{C}^{\mathrm{fs}}$ using feasible set $\{2\}$. Thus the oracle inclusion can also be strict.

Why every feasible subset size is retained. Enumerating more than the smallest feasible subsets protects the full honest mean when the declared budget is conservative. Sets larger than $K - A$ in Equation (5) can be necessary for the stated honest-mean identification and oracle containment. For $K = 3, A = 1, n = 2, k = 1$ and

$$V = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix},$$

every two-node subset has calibration order statistic $1/2$, while $[3]$ has $2/3$. With no Byzantine nodes and $w = (3/5, 3/5, 3/5)$, the full honest-mean oracle therefore includes the candidate, whereas fixed-set and joint-threshold variants restricted to size-two subsets exclude it. This is not a marginal-coverage counterexample. Any fixed $H_0 \subseteq \mathcal{H}$ of size $K - A$, chosen independently of the examples, defines an exchangeable clean subset-mean score. The exact-size fixed-set union contains its directionally padded conformal set, and the exact-size joint-threshold rule contains that union. Both retain marginal coverage at least $k/(n+1)$, but need not preserve the rank event of the full honest mean when $a < A$. Enumerating all feasible sizes supplies the stronger identification and containment guarantees proved here.

Proof of Corollary E.1. Each subset calibration mean, its order statistic, $\tau_{A,k}$, and U_A are coordinatewise nondecreasing in calibration reports. Each subset query mean and L_A are coordinatewise nondecreasing in query reports. Therefore increasing Byzantine calibration coordinates and decreasing their query coordinates can only decrease the fixed-set margin, increase the joint threshold, increase the deletion threshold, or decrease either query aggregate. It can only add candidates to each proposed set. The endpoints $(\mathrm{hi}, \mathrm{lo})$ attain the coordinatewise extremum. Reversing both directions, the endpoints $(\mathrm{lo}, \mathrm{hi})$ attain the opposite extremum and can only remove candidates relative to any other transcript in $[\mathrm{lo}, \mathrm{hi}]$. Neither argument uses $a \le A$.

For low–high with $a = A$, the number of honest identities is $K - A$. Because the Byzantine calibration reports lie below the honest reports, the largest $K - A$ reports are exactly the honest multiset; because the Byzantine query reports lie above the honest reports, the smallest $K - A$ reports are exactly the honest multiset. Thus, for each calibration column j, $\mathrm{U}_A(v_{\cdot j}) = c_j(\mathcal{H}; V)$ and $\mathrm{L}_A(w) = h^{-1} \sum_{i \in \mathcal{H}} w_i$. More explicitly, any $G \in \mathfrak{H}_A$ has size at least h. The extremal identity in Equation (21) therefore gives $c_j(G; V) \le c_j(\mathcal{H}; V)$ for every j and, under the reversed endpoint ordering at query time, $|G|^{-1} \sum_{i \in G} w_i \ge h^{-1} \sum_{i \in \mathcal{H}} w_i$. Consequently,

$$\tau_{A,k}(V) = T_k(\mathcal{H}; V), \qquad \Delta_{A,k}(V, w) = \frac{1}{h} \sum_{i \in \mathcal{H}} w_i - T_k(\mathcal{H}; V),$$

and the deletion, joint-threshold, and fixed-set membership criteria all equal the criterion defining $\mathcal{C}_\alpha^{\mathrm{dec},\mathcal{H}}(x)$, for arbitrary registered directional padding. With zero registered reconstruction error, decoded honest reports equal their clean values, so $\mathcal{C}^{\mathrm{dec},\mathcal{H}} = \mathcal{C}^\circ$. Under exchangeability and no ties, the future clean score has a uniform rank among the $n+1$ scores, so coverage is exactly $k/(n+1)$.

Finally, $\mathfrak{H}_{A_1} \subseteq \mathfrak{H}_{A_2}$ when $A_1 \le A_2$. Thus the fixed-set minimum can only decrease and the joint-threshold maximum can only increase. The mean of the largest $K - A$ reports is nondecreasing in A, while the mean of the smallest $K - A$ reports is nonincreasing. With fixed padding, each change can only enlarge its corresponding prediction set. □

Proof of Theorem 4.4. The $k = n + 1$ case is immediate. Let $k \le n$. The first three inclusions are Equation (13); it remains to prove the outer inclusion for $\mathcal{C}^{\mathrm{del}}$.

By Lemma K.2 with $B = D_r$ on each channel and $h \ge K - A$, calibration scores obey

$$\widehat{R}_j \le R_j + \mu_{\mathrm{cal}} + \rho^+_{\mathrm{cal}},$$

so the same inequality holds for their kth order statistics. The query lower bound is

$$\widehat{s}^{\mathrm{qry}}_{\mathrm{del}}(x, y) \ge s_{\mathcal{H}}(x, y) - \mu_{\mathrm{qry}} - \rho^-_{\mathrm{qry}}.$$

If $y \in \mathcal{C}^{\mathrm{del}}_\alpha(x)$, combine these inequalities with the definition of its threshold to obtain

$$s_{\mathcal{H}}(x, y) \le R_{(k)} + \mu_{\mathrm{cal}} + \mu_{\mathrm{qry}} + \rho^-_{\mathrm{cal}} + \rho^+_{\mathrm{cal}} + \rho^-_{\mathrm{qry}} + \rho^+_{\mathrm{qry}} = R_{(k)} + W_{\mathrm{del}}.$$

This proves the pathwise sandwich. □

The outer width also controls the *average* number of extra answers, once the future question is assumed independent of the calibration split.

Corollary L.3 (Expected excess set size)**.** *Assume the conditions of Theorem 4.4 and let* $N(t) = \mathbb{E}_X \sum_{y \in \mathcal{Y}} \mathbf{1}\{s_{\mathcal{H}}(X, y) \le t\}$ *be the expected number of candidates with clean score at most* t*, where* $\mathbb{E}_X$ *averages over a future input and* $\mathbf{1}$ *equals one when its condition holds and zero otherwise. Write* $\omega_{\mathrm{size}}(u) = \sup_t \{N(t+u) - N(t)\}$ *for* $u \ge 0$ *and take the widths* D_r *and the reconstruction bounds*

to be fixed conditional on the pre-calibration system. If the future X is independent of the calibration split conditional on that system, then, for every $\mathcal{C} \in \{\mathcal{C}^{\mathrm{fs}}, \mathcal{C}^{\mathrm{jt}}, \mathcal{C}^{\mathrm{del}}\}$,

$$0 \le \mathbb{E}\big[|\mathcal{C}_\alpha(X)| - |\mathcal{C}^\circ_\alpha(X)|\big] \le \omega_{\mathrm{size}}(W_{\mathrm{del}}). \tag{28}$$

If each candidate-score cumulative distribution function (CDF) $G_y(t) = \mathbb{P}_X\{s_{\mathcal{H}}(X, y) \le t\}$ satisfies $|G_y(t) - G_y(t')| \le L_y|t - t'|$ for a nonnegative constant L_y (the Lipschitz condition), the upper bound is at most $\min\{M, W_{\mathrm{del}} \sum_y L_y\}$.

Proof. Conditional on the clean calibration examples, write $q = R_{(k)}$; no conditioning on the adversarial transcript is needed. The expectation below averages over the future input and any randomness in Byzantine reporting. For any $\mathcal{C} \in \{\mathcal{C}^{\mathrm{fs}}, \mathcal{C}^{\mathrm{jt}}, \mathcal{C}^{\mathrm{del}}\}$, independence of the future X from calibration and the pathwise inclusions give

$$\begin{aligned} 0 &\le \mathbb{E}\big[|\mathcal{C}_\alpha(X)| - |\mathcal{C}^\circ_\alpha(X)| \mid Z_1, \dots, Z_n\big] \\ &\le N(q + W_{\mathrm{del}}) - N(q) \le \omega_{\mathrm{size}}(W_{\mathrm{del}}). \end{aligned}$$

Averaging over calibration proves Equation (28). If each G_y is L_y-Lipschitz, sum $G_y(q + W_{\mathrm{del}}) - G_y(q) \le L_y W_{\mathrm{del}}$ and cap the result by M. □

Proof of Proposition F.1. If $k = n + 1$, all rules return $\mathcal{Y}$. Suppose $k \le n$. Put $\bar{h} = K - A$ and $T = \mathrm{TM}_m(v)$. For ordered reports in an interval of width B, direct decomposition gives

$$\mathrm{U}_A(v) - \mathrm{TM}_m(v) \le \frac{m}{K - A} B, \qquad \mathrm{TM}_m(v) - \mathrm{L}_A(v) \le \frac{m}{K - A} B.$$

Indeed,

$$\begin{aligned} \bar{h}\{\mathrm{U}_A(v) - T\} &= \sum_{\ell=A+1}^{m} \{v_{(\ell)} - T\} + \sum_{\ell=K-m+1}^{K} \{v_{(\ell)} - T\} \le mB, \\ \bar{h}\{T - \mathrm{L}_A(v)\} &= \sum_{\ell=1}^{m} \{T - v_{(\ell)}\} + \sum_{\ell=K-m+1}^{K-A} \{T - v_{(\ell)}\} \le mB. \end{aligned}$$

In the first line the first sum is nonpositive and the second has m terms; in the second line the second sum is nonpositive and the first has m terms. Empty sums are zero. The arrays with $K - m$ zeros followed by m copies of B, and with m zeros followed by $K - m$ copies of B, attain the two respective constants.

Order-statistic monotonicity transfers the first inequality to calibration kth order statistics. If $y \in \mathcal{C}^{\mathrm{del}}_\alpha(x)$, then

$$\begin{aligned} \widehat{s}^{\mathrm{sym,m}}(x, y) &\le \widehat{s}^{\mathrm{qry}}_{\mathrm{del}}(x, y) + g_{\mathrm{qry}}(m) \\ &\le \widehat{R}_{(k)} + \rho_{\mathrm{cal}} + \rho_{\mathrm{qry}} + g_{\mathrm{qry}}(m) \\ &\le \widehat{R}^{\mathrm{sym,m}}_{(k)} + g_{\mathrm{cal}}(m) + g_{\mathrm{qry}}(m) + \rho_{\mathrm{cal}} + \rho_{\mathrm{qry}}, \end{aligned}$$

which is exactly membership in $\mathcal{C}^{\mathrm{sym,m}}$. Prepending the hierarchy from Theorem 4.3 proves the full claim. □

Proof of Proposition F.3. With A Byzantine reports below the honest range, lower trimming removes the Byzantine values and upper trimming removes the A largest honest values. Thus the calibration aggregate is $S_j + L_\xi$. With Byzantine reports above the honest range at prediction, it is $S_{n+1} + U_\xi$. The threshold event is therefore exactly Equation (16). If $d = \Delta_\xi - g > 0$ and S is uniform on an interval of width less than d, then $S_{(k)} - d$ is below the lower endpoint for every calibration realization; coverage is zero.

Put $r = h - A = K - 2A$. If $h \ge 2A$, cancellation of the overlapping terms gives

$$\Delta_\xi = \frac{\sum_{i=h-A+1}^{h} \xi_{(i)} - \sum_{i=1}^{A} \xi_{(i)}}{r} \le \frac{A}{r} R.$$

When $h < 2A$, the lower and upper r-blocks are disjoint and their mean gap is at most R. Arrays whose offsets take only the values $\{0, R\}$ attain both bounds, proving Equation (17). □

Proof of Lemma G.1. Conditional on $\mathcal{F}$, the pairs (Z_j, Ξ_j) are exchangeable. Applying the same measurable map ϕ to every pair preserves joint permutation invariance, so the T_j are exchangeable. The final claim follows from Lemma L.1, with the full-set convention when $k = n + 1$. □

Proof of Proposition G.2. The proof counts how many honest nodes must report small p-values before the merger can reject, then constructs dependent honest values attaining that probability bound. Let the actual Byzantine count be $a \le A$ and, for $t \in [0, 1]$, define $N_{\mathcal{H}}(t) = \sum_{i \in \mathcal{H}} \mathbf{1}\{p_i \le t\}$. If $p_{(A+\ell)} \le t$, at least $A + \ell - a$ honest reports are at most t. Marginal superuniformity, linearity of expectation, and Markov's inequality give

$$\mathbb{P}\{p_{(A+\ell)} \le t\} \le \frac{(K-a)t}{A+\ell-a} \le \frac{K-A}{\ell} t.$$

For the last inequality write $d = A - a$; after cross-multiplication the difference between right and left numerators is $d\{K - A - \ell\} \ge 0$. For every $u < 1$,

$$\{p_\ell^{\mathrm{merge}} \le u\} = \left\{ p_{(A+\ell)} \le \frac{u\ell}{K-A} \right\},$$

so the preceding display bounds its probability by u; the case $u = 1$ is trivial. Taking $u = \alpha$ proves the prediction-set claim.

To show sharpness, take exactly A Byzantine zeros, so that $h = K - A$. Fix any $t \in [0, \ell/h]$. With probability ht/ℓ, choose a uniformly random ℓ-subset of the honest nodes, set their p-values to t, and set all other honest values to one; otherwise set every honest value to one. For each honest node and every $u < 1$, the probability of being at most u is zero when $u < t$ and t when $u \ge t$, hence the marginal is superuniform. Nevertheless, $p_{(A+\ell)} \le t$ with probability ht/ℓ. Therefore the multiplicative constant cannot be reduced under arbitrary dependence. Ties are counted with multiplicity and candidate inclusion uses the strict comparison $p_\ell^{\mathrm{merge}} > \alpha$, so exclusion is the event just bounded. □

Remark L.4 (Candidate regularity and an unconditional alternative). The true-label score may have a bounded density while incorrect-label scores have atoms. Hence that density alone does not control the uniform modulus ω_{size}: an arbitrarily small threshold increment crossing such an atom can add every incorrect label. The bound uniform over calibration thresholds therefore concerns N or the candidate CDFs.

A bounded true-label density gives an unconditional excess-size bound even when incorrect-label scores have atoms. Under the expected-size conditions of Corollary L.3, suppose each clean calibration score R_j has a density bounded by L conditional on the frozen system, and that W_{del} is fixed under that conditioning. Then every proposed rule satisfies

$$0 \le \mathbb{E}\big[|\mathcal{C}_\alpha(X)| - |\mathcal{C}^\circ_\alpha(X)|\big] \le \min\{M, MnLW_{\mathrm{del}}\}. \tag{29}$$

For $k \le n$, condition on X and put $s_y = s_{\mathcal{H}}(X, y)$. The outer sandwich can add candidate y only if $R_{(k)} \in [s_y - W_{\mathrm{del}}, s_y)$. Since an order statistic equals one of its inputs, this implies that at least one calibration score R_j lies in that interval. Independence from X, the marginal density bound, and a union bound give probability at most nLW_{del}. Summing over the M candidates and capping by M proves the display; independence among the calibration scores is not needed. The excess is zero when $k = n + 1$. This bound permits candidate atoms but pays an explicit factor n. The two bounds use different regularity information; when both are available, their minimum may be used.

M Implementation and exact arithmetic

This appendix gives the reference protocol for the three proposed rules and the exact arithmetic that implements their inclusion tests. All three use the same aligned, identity-resolved reports and registered reconstruction bounds, with their respective calibration and query comparisons from Theorem 4.3.

Freeze and aligned calibration. Before calibration the protocol fixes every model, retriever, corpus index, prompt, candidate ordering rule, score transform, quantizer, membership list, retry policy, corruption budget A, and optional honest-dropout budget D. All tuning uses data disjoint from calibration. Infrastructure errors are handled as in Appendix C and consume neither A nor D;

the deterministic fallback is reserved for an input-dependent retrieval or model failure selected by the frozen honest score map. For each example the hub sends every node the same input and ordered candidate list, and honest nodes score, quantize, and sign a vector. A missing or malformed authenticated report is replaced by a fixed valid sentinel, never silently removed; the experiments use the lowest codeword 0 in calibration and the maximum codeword $S_{\max}$ at query time. The hub keeps node identities and forms V from each vector's gold (correct-answer) coordinate.

Threshold and deployment. The index k is computed by integer arithmetic; $k = n + 1$ returns $\mathcal{Y}$, and otherwise the three rules apply the comparisons of Section 3, every one non-strict. More than $A + D$ absent or malformed slots trigger an abort or recalibration. The guarantees assume that at most D missing identities are honest, which absence alone cannot reveal, so the primary protocol sets $D = 0$. The matched-pipeline diagnostic further requires identical calibration and query maps, and Byzantine randomness that does not depend on phase, index, history, or gold-label exposure.

Robust p-value option. The hub can compute the p-values (18) from the aligned score tensor, which requires a complete matched calibration/query transcript for every honest node. If nodes instead transmit p-values, an exact rank costs $\lceil \log_2(n + 1) \rceil$ bits (nine when $n = 499$). Upward rounding preserves validity; downward or nearest rounding need not.

M.1 Exact decisions on the registered quantizer grid

The registered implementation decides inclusion in exact integer arithmetic, so a boundary tie falls on the side the theorems require. Floating-point evaluation of algebraically equal means can put the same boundary on opposite sides of a non-strict comparison. Write $L_{\mathrm{cal}} = 2^{b_{\mathrm{cal}}} - 1$ and $L_{\mathrm{qry}} = 2^{b_{\mathrm{qry}}} - 1$, so each decoded report is $S_{\max}$ times an integer index over L_r. For a feasible H, let C_H be the kth order statistic of its calibration index sums and $\Sigma_H(y)$ its query index sum. The fixed-set test for this H is exactly

$$2L_{\mathrm{cal}}\Sigma_H(y) \le 2L_{\mathrm{qry}}C_H + |H|(L_{\mathrm{cal}} + L_{\mathrm{qry}}),$$

obtained by multiplying the mean comparison and its two half-cell paddings by $2|H|L_{\mathrm{cal}}L_{\mathrm{qry}}/S_{\max}$. It preserves equality and allows unequal phase precisions. Joint-threshold, deletion, and guarded symmetric decisions use the corresponding exact rational thresholds, and α is treated as a rational number. The common point ranker of Equation (30) also compares integer code sums, so exact ties reach the fixed candidate order (Appendix O.3). Every pathwise assertion implied by the theory is checked before summarization; a failure would be an implementation or assumption error, not Monte Carlo uncertainty.

N Random Byzantine membership

This appendix proves Corollary 4.2, in which the Byzantine set $\mathcal{A}$ is drawn from a probability law instead of being fixed, and gives the tail probability ε_A for independent failures. The deployed rule does not change: the hub still uses the declared budget A and never estimates $\mathcal{A}$ or its law.

Failure law. Let $\mathcal{A} \subseteq [K]$ be a random variable on the same probability space, with $\mathcal{H} = [K] \setminus \mathcal{A}$ and $a = |\mathcal{A}|$, satisfying the following.

1. $\mathcal{A}$ is drawn once before calibration and held fixed for the entire calibration-to-query comparison.
2. For every $\mathcal{A}_0 \subseteq [K]$ with $\mathbb{P}\{\mathcal{A} = \mathcal{A}_0\} > 0$, the clean examples $Z_1, \ldots, Z_{n+1}$ are exchangeable conditional on $\{\mathcal{A} = \mathcal{A}_0\}$ and the frozen honest score functions.
3. The registered directional bounds (2) hold for every node that can be honest (they are properties of the registered quantizers), and the declared budget $A < K$ is a deterministic constant, fixed before any data are seen.

Drawing $\mathcal{A}$ independently of the clean examples is the simplest way to meet the second condition, but independence is not required. What the condition excludes is membership selected after inspecting the clean calibration scores, since conditioning on such a choice need not leave the examples

exchangeable. Write

$$\varepsilon_A = \mathbb{P}\{a > A\}$$

for the probability that the realized corruption exceeds the declared budget.

Conditions 1–3 restate the hypotheses of Corollary 4.2.

Proof of Corollary 4.2. For $k = n + 1$ all three rules return $\mathcal{Y}$ and coverage is one, so let $k \le n$. Fix $\mathcal{A}_0$ with $\mathbb{P}\{\mathcal{A} = \mathcal{A}_0\} > 0$ and $|\mathcal{A}_0| \le A$. Conditional on $\{\mathcal{A} = \mathcal{A}_0\}$ the honest set is the deterministic set $\mathcal{H}_0 = [K] \setminus \mathcal{A}_0$, the clean examples are exchangeable by the second condition, and $|\mathcal{H}_0| = K - |\mathcal{A}_0| \ge K - A$, so $\mathcal{H}_0 \in \mathfrak{H}_A$. Every hypothesis of Theorem 4.1 therefore holds in this conditional model, and part 1 of Theorem 4.3 gives the same bound for the two smaller rules, since each contains the oracle set, giving

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}_\alpha(X_{n+1}) \mid \mathcal{A} = \mathcal{A}_0\} \ge \frac{k}{n+1}.$$

The bound holds for every such $\mathcal{A}_0$ and the conditional probability is nonnegative on the remaining event, so averaging over $\mathcal{A}$ and discarding $\{a > A\}$ gives

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}_\alpha(X_{n+1})\} = \mathbb{E}\big[\mathbb{P}\{Y_{n+1} \in \mathcal{C}_\alpha(X_{n+1}) \mid \mathcal{A}\}\big] \ge \frac{k}{n+1}\,\mathbb{P}\{a \le A\}.$$

This is the first inequality of Equation (12). The second uses $k/(n+1) \ge 1-\alpha$ and $(1-\alpha)(1-\varepsilon_A) = 1 - \alpha - \varepsilon_A + \alpha\varepsilon_A \ge 1 - \alpha - \varepsilon_A$. □

Independent per-node failure. The corollary needs only the single number ε_A, which a per-node failure model supplies in closed form. Suppose node i is Byzantine independently with probability p_i. Then a has the Poisson–binomial distribution, the law of a sum of independent but not identically distributed indicators, and ε_A is its upper tail. In the two-class case of interest, where K_1 ordinary nodes each fail with probability p_1 and $K_2 = K - K_1$ suspect nodes each fail with the larger probability p_2 (the subscripts of p_1 and p_2 index classes, not nodes),

$$\varepsilon_A = \sum_{m=A+1}^{K} \sum_{\substack{m_1+m_2=m \\ 0\le m_\ell \le K_\ell}} \prod_{\ell=1}^{2} \binom{K_\ell}{m_\ell} p_\ell^{m_\ell} (1-p_\ell)^{K_\ell - m_\ell}.$$

With a single class this is $\mathbb{P}\{\mathrm{Bin}(K, p) > A\}$. The model describes which nodes are corrupt, not how a corrupt node reports; a node counted in $\mathcal{A}$ still reports arbitrarily.

Upper bounds on the per-node failure probabilities give a valid upper bound on ε_A.

Lemma N.1 (Monotone tails and the corner rule)**.** *Let $p = (p_i)_{i\in[K]}$ and $q = (q_i)_{i\in[K]}$ be per-node failure probabilities for independent membership with $p_i \le q_i$ for every i. Then $\varepsilon_A(p) \le \varepsilon_A(q)$ for every A. Consequently, for a componentwise box $\mathcal{U} = \{p : p_i \le \bar{p}_i \text{ for all } i\}$,*

$$\sup_{p\in\mathcal{U}} \varepsilon_A(p) = \varepsilon_A(\bar{p}),$$

so the robust declared budget

$$A^\star(\delta, \mathcal{U}) = \min\Big\{A' \in \{0, \dots, K-1\} : \sup_{p\in\mathcal{U}} \mathbb{P}_p\{a > A'\} \le \delta\Big\}$$

follows from one exact tail evaluation at the corner $\bar{p}$ of $\mathcal{U}$, with no search over $\mathcal{U}$.

Proof. Let $U_1, \dots, U_K$ be i.i.d. $\mathrm{Unif}[0,1]$ and set $a_p = \sum_i \mathbf{1}\{U_i \le p_i\}$ and $a_q = \sum_i \mathbf{1}\{U_i \le q_i\}$, which have the required marginals. For each i, $p_i \le q_i$ gives $\mathbf{1}\{U_i \le p_i\} \le \mathbf{1}\{U_i \le q_i\}$, so $a_p \le a_q$ pointwise and $\mathbb{P}\{a_p > A\} \le \mathbb{P}\{a_q > A\}$ for every A. The supremum over $\mathcal{U}$ is then attained at $\bar{p}$, which lies in $\mathcal{U}$. □

O Empirical design and reproducible methods

The full registered manifests, configurations, and implementation details are in the supplementary code's README. This section specifies what the completed synthetic and four-task RAG studies ran; findings and the independent output audit appear in Appendix P. Synthetic cells calibrate on $n = 499$

examples unless stated otherwise, and Table 5 gives each RAG task's calibration size. A clean tensor stores scores before offline attacks and quantization; an analysis cell applies one configuration to those scores. Several cells can reuse a tensor, so cell counts are not counts of independent datasets.

O.1 Registered synthetic models

Smooth copula model. The smooth model separates dependence between nodes from variation in their score levels. Here i indexes a node, j an example, and y a candidate. For each replicate, draw fixed node effects δ_i by randomly permuting evenly spaced points in $[-1, 1]$, so their mean is zero, and draw Y_j uniformly from $\{1, \dots, M\}$. With Φ the standard-normal CDF, let G_{jy} and E_{ijy} be independent arrays of i.i.d. standard Gaussians and, for dependence parameter $\gamma \in [0, 1]$, form

$$U_{ijy} = \Phi\{\sqrt{\gamma}\, G_{jy} + \sqrt{1-\gamma}\, E_{ijy}\}.$$

The shared G_{jy} couples the nodes; E_{ijy} supplies node-specific variation. Let $F^{-1}_{p,q}$ be the $\mathrm{Beta}(p, q)$ quantile function and $\eta \geq 0$ the magnitude of node effects. For $y = Y_j$, set

$$s^{\mathrm{syn}}_{ijy} = \mathrm{logit}^{-1}\left[\mathrm{logit}\{F^{-1}_{2,5}(U_{ijy})\} + \eta\delta_i\right],$$

with U_{ijy} clipped to $[10^{-9}, 1 - 10^{-9}]$, and use $F^{-1}_{5,2}$ when $y \neq Y_j$. Without node effects ($\eta = 0$) a correct answer's score has mean $2/7$ and an incorrect one's $5/7$, so smaller scores favor inclusion. Conditional on the node effects, examples are independent, while γ controls within-example node dependence and η node heterogeneity. The registered values are $\gamma \in \{0, 0.5, 0.9, 0.99\}$ and $\eta \in \{0, 0.5, 1, 2\}$. Randomized stable attacks draw their randomness from an independent per-example key, never from an example's index or phase. The label generates the joint law but is never an argument of the deployed score map or a phase-blind attack. Clipping and finite-precision storage can create atoms, so coverage uses the tie-safe rank inequality.

Atomic model. Discrete score laws test the conservative treatment of tied ranks. The two-point true-label law assigns probabilities $(0.6, 0.4)$ to $(0.2, 0.5)$; an incorrect label assigns $(0.2, 0.4, 0.4)$ to $(0.2, 0.5, 0.8)$. The four-point variant uses supports $(0.1, 0.3, 0.5, 0.7)$ and $(0.3, 0.5, 0.7, 0.9)$ with equal masses. Node i uses a shared draw with probability $\sqrt{\gamma}$, so two nodes share it with probability γ. The diagnostic fixes $(K, A, a, m, n, M, \alpha, \gamma) = (16, 2, 2, 2, 499, 4, 0.10, 0.99)$, with 200 replicates, 2,000 test examples, bit depths $\{1, 2, 4, 8\}$, and the `max` and `low_high` attacks.

Identification model. This experiment compares separate local distributions with aligned node averages in the two worlds of Proposition B.1, with $K = 2$, $A = m = 0$, and $\alpha = 0.25$. Each of 200 replicates uses nested sample sizes $n \in \{99, 999, 9{,}999\}$ and 5,000 test examples. The local arm thresholds the average of the two empirical node CDFs, $G_n(t)$, at the smallest pooled-grid t with $G_n(t) \geq k/(n+1)$; the aligned arm uses the kth order statistic of the paired node averages. We record sup-norm CDF error, threshold error against the exact $(1-\alpha)$ quantile, and true-label coverage; set size is undefined because no false-label law is specified.

Shift stress model. The shift control changes the future true-label law without Byzantine corruption, so it lies outside the exchangeability guarantee. At the operational factors with $a = 0$ and no attack, it reuses 200 calibration replicates of the operational clean tensor and changes only the correct-label test law from $\mathrm{Beta}(2, 5)$ to $\mathrm{Beta}(3, 4)$, with 5,000 test scores per node. Only true-label coverage is reported, and these runs are not pooled with theorem-confirming experiments.

Narrow-support low–high model. This model checks the zero-coverage failure of symmetric trimming with insufficient padding. Fix $K = 16$, $A = a = m = 2$, $n = 499$, $\alpha = 0.10$, 200 replicates, and 5,000 test examples. The 14 honest offsets are $(-0.15, -0.15, 0, \dots, 0, 0.15, 0.15)$ with ten zeros, $S \sim \mathrm{Unif}[0.45, 0.46]$, honest reports are exact, and the two Byzantine nodes report low–high. With $L_\xi = -0.025$, $U_\xi = 0.025$, and $\Delta_\xi = 0.05$, the exact population coverage at total paddings $g \in \{0, 0.025, 0.05, 0.10\}$ is $(0, 0, 450/500, 1)$, against which the Monte Carlo estimates are compared. The sharp p-merger construction at $K - A = 14$ and $\ell = 3$ is checked by exact enumeration of its $\binom{14}{3} = 364$ subset states plus the null state. With $t = \alpha\ell/(K - A) = 3/140$ and both Byzantine p-values zero, the all-one honest state has probability 0.9 and each three-subset state, which sets those three honest p-values to t, has probability $0.1/364$. Each honest p-value is then superuniform, whereas $\mathbb{P}\{p_{(A+\ell)} \leq t\} = 0.1$ exactly.

Table 4: Registered synthetic factor levels, not the Cartesian product of all experiments. The deployable trim is chosen from the registered budget A, never from the unobserved realized count a.

Factor	Values
Nodes K	$8, 16, 32, 64$
Declared budget A	stochastic grid at 3; operational recombinations at 2; offline sensitivity at $2, 4, 6$; analytic G/W (padding and width, Appendix H) phase diagram at $0, 1, 2, 3, 4, 6$
Actual Byzantine count a	$0, 1, 2, 3, 4, 6$ (report the induced a/K)
Symmetric trim m	$\max(0, A-2), A-1, A, A+\lceil 0.05K \rceil, \lfloor (K-1)/2 \rfloor$
Calibration size n	$49, 99, 249, 499, 999, 2499$
Candidate count M	$2, 4, 8, 16$
Bits per scalar	$1, 2, 3, 4, 6, 8, 16, 32$
Miscoverage α	$0.05, 0.10, 0.20$
Node dependence γ	$0, 0.5, 0.9, 0.99$
Heterogeneity η	$0, 0.5, 1, 2$

Primary configuration and offline reuse. The operational primary, the default for the synthetic results of Section 5, is

$$(K, a, A, m, n, M, b_{\mathrm{cal}}, b_{\mathrm{qry}}, \alpha, \gamma, \eta) = (16, 2, 2, 2, 499, 4, 8, 8, 0.1, 0.5, 0.5).$$

The retained $A = 3$ block is anchored at the registered baseline $(16, 2, 3, 3, 499, 4, 4, 4, 0.1, 0.5, 0.5)$ (cell 000), from which the one-factor sweeps vary one factor at a time over Table 4. Its variant with $a = A = m = 3$ and four bits is the *four-bit stress cell*; stable maximum (cell 006) and high–low (cell 102) there form the four-bit stress inference families. *Offline reuse* means applying A, m, and bit depth to a cached full-precision clean tensor instead of generating new scores; the operational cells and cell 102 reuse cell 000's tensor. Four bit-mismatch summaries, $(b_{\mathrm{cal}}, b_{\mathrm{qry}}) \in \{(2, 8), (4, 8), (8, 4), (8, 2)\}$ at the operational factors, test directional padding but are ineligible for the matched-pipeline and per-node p-value assertions. The reference attack is stable maximum, replaced only in the attack sweep. At the operational anchor it is the primary efficiency attack, high–low the worst-case-inflation endpoint, and low–high the exact-cancellation endpoint. The primary theorem-valid trim is $m = A$; $m < A$ is an assumption violation and $m > A$ an over-trimming diagnostic. The two interaction blocks cross $a \in \{0, 2, 3, 4\}$, $m \in \{2, 3, 4\}$, $b \in \{2, 4, 8\}$ and $K \in \{8, 16, 32\}$, $\gamma \in \{0, 0.9, 0.99\}$, $\eta \in \{0, 0.5, 2\}$.

Analysis scope and secondary comparators. The retained manifest has 94 analysis cells with $A = 3$. Adding the operational primary, the four bit-mismatch summaries, and 11 targeted offline summaries gives 110 stochastic summaries on 88 distinct clean-score cells. The targeted summaries are the two operational switching attacks, the $A = 2, a = 3$ low–high violation, four-bit stress high–low, and seven budget cells at $A = m \in \{2, 4, 6\}$. Each has 500 calibration replicates and 5,000 test examples. Only 23 tensors carry every candidate coordinate, so only they support set size, full-set rate, and accuracy; the others support coverage. The secondary synthetic comparators (Appendix O.4) run on the 45 full-candidate smooth cells backed by those 23 tensors, all with $K = 16$, and enter no primary simultaneous-inference family.

O.2 Attack catalog

Each attack fixes what the corrupted nodes report on each channel, with scores in $[0, S_{\max}]$, $S_{\max} = 1$, and minimum and maximum codewords $\mathrm{lo} = 0$ and $\mathrm{hi} = S_{\max}$ (Appendix E). The seven primary synthetic attacks transform a cached potential clean score $s_i^{\mathrm{pot}}(x, y)$ through the node's registered quantizer $Q_{i,r}$ on channel r. `min` and `max` (stable maximum) report the minimum or maximum codeword on both channels; `reflect` reports $Q_{i,r}(S_{\max} - s_i^{\mathrm{pot}})$; `shift_plus` adds $0.25 S_{\max}$ and clips; `uniform` reports a uniformly drawn codeword per key, candidate, and node without a phase flag. These five *stable* attacks use the same transformation in both phases. The two *switching* attacks, `low_high` (low–high) and `high_low` (high–low), report the minimum in one phase and the maximum in the other, low–high reporting the minimum in calibration. The real-data manifest adds `cal_only_max` and `query_only_max` (cal. max and query max), which report the maximum codeword on the named channel and the normally quantized score on the other, plus secondary negative-shift and bit-flip faults.

For each base clean cell and replicate, one salted uniform permutation of the K identities is drawn and its first a identities are corrupted, so corrupted sets are nested. The permutation is shared by every method, attack, A, m, and bit depth, and the attack interface never receives the gold label. In matched-pipeline experiments a stable attack is a fixed, phase-blind transformation that cannot inspect the example index, phase, calibration history, or label availability, which tests Lemma G.1. Switching attacks test the deletion coverage theorem and the symmetric low–high lower bound. Cells with $a > A$ test misspecification; theorem assertions are disabled in such deliberately ineligible cells.

O.3 Executed finite-answer RAG protocol

The executed study comprises OpenBookQA, MedQA-US, MedMCQA, and MMLU-Med (Mihaylov et al., 2018; Jin et al., 2021; Pal et al., 2022; Hendrycks et al., 2021). The three medical pools use the representation of the MIRAGE (Medical Information Retrieval-Augmented Generation Evaluation) benchmark (Xiong et al., 2024) with its complete Textbooks corpus. The results do not represent the complete MIRAGE suite.

Corpora, virtual nodes, and retrieval. Each corpus is split two ways into $K = 16$ virtual nodes, so these are controlled public-data federations rather than independent institutional deployments. OpenBookQA uses 1,326 distinct official science facts, and the medical tasks use all 125,847 passages from 18 Textbooks sources. The hash partition (H) assigns documents or passages by a frozen salted hash. The topic/source partition (T) uses frozen topic assignments from `all-MiniLM-L6-v2` embeddings (Reimers & Gurevych, 2019) for OpenBookQA and keeps whole books together for the medical tasks, assigning each book greedily to the least-loaded node. Each node retrieves four passages from its own shard by BM25 keyword ranking (Robertson & Zaragoza, 2009) ($k_1 = 1.2$, $b_{\mathrm{BM25}} = 0.75$) using the question alone, so candidate answers and labels cannot alter the evidence.

Frozen answer scoring. All answer scores use GPT-2 small (124 million parameters) (Radford et al., 2019) with the prompt

```
Question: [question]
Context: [space-joined retained passages]
Answer:
```

Each complete answer text y is scored in a separate teacher-forced pass as the continuation of one space followed by its tokens $o_{y1}, \ldots, o_{yT_y}$. The candidate weight uses the average log probability per token, avoiding a direct penalty for longer answers:

$$\ell_i(y) = \frac{1}{T_y} \sum_{t=1}^{T_y} \log p_{\mathrm{GPT2}}(o_{yt} \mid \mathrm{prompt}_i, o_{y,<t}), \qquad \pi_i(y) = \frac{\exp\{\ell_i(y)\}}{\sum_{z \in \mathcal{Y}} \exp\{\ell_i(z)\}}.$$

These weights are scoring devices, not calibrated posterior probabilities. Passages are shortened to fit the 1,024-token limit, while question and answer texts are never truncated; a question/candidate-only overflow yields the registered all-ones score vector, kept in every denominator. Candidate order is fixed by a salted public-content hash. With $z_i(x)$ the context node i retrieved, the primary score is

$$s_i(x, y) = 1 - \pi_i\{y \mid x, z_i(x)\} \in [0, 1],$$

so smaller scores favor inclusion. The secondary NLL score is the clipped negative log likelihood

$$s_i^{\mathrm{NLL}}(x, y) = \min\{-\log(\max[\pi_i(y \mid x, z_i(x)), 10^{-12}]), 12\}/12,$$

applied before quantization or attacks and separately calibrated, with the primary splits and corruption randomness. The no-RAG and centralized-union anchors (Appendix O.4) use the same scoring with empty context and with four passages retrieved from the whole corpus.

Table 5: Executed pool sizes and the split used within each repeat; pilot records are excluded from every held-out pool. Every task has four answer candidates, and $k = \lceil 0.9(n+1) \rceil$ for nominal coverage 90%.

Task	Pilot	Held-out N	Calibration n	Evaluation	k
OpenBookQA	500	1,000	333	667	301
MedQA-US	100	1,173	391	782	353
MedMCQA	100	4,083	499	3,584	450
MMLU-Med	100	989	329	660	297

Pools and repeated splits. Scoring decisions were made on pilot records disjoint from the held-out pools (Table 5). Each of 100 repeats permutes the fixed held-out pool without consulting labels or scores, using the first n records for calibration and the remaining $N - n$ for evaluation. Corrupt identities and attack randomness are paired across methods and partitions. This supports exchangeability for a uniformly selected held-out record, not independence between that record and its calibration subset, so efficiency is measured directly and the bounds in Equations (28) and (29) are not invoked.

Threats and cells. Primary cells use $(K, A, m, b, \alpha) = (16, 2, 2, 8, 0.10)$ and seven threats: clean, stable maximum with $a = 1, 2$, and, with $a = 2$, calibration-only maximum, query-only maximum, high–low, and low–high. Across four tasks and two partitions this gives 56 primary cells. Secondary cells add $a = 3 > A$ bound violations, a four-bit stress block $(A, m, b) = (3, 3, 4)$, the fault attacks, $\alpha \in \{0.05, 0.20\}$, secondary aggregators, and NLL scores, for 408 cells and 264,800 repeat-level rows in total.

Common point ranker and raw sets. Point prediction is separate from set construction. Every non-oracle federated set method shares one predictor, which ranks candidates by TM_A, the mean of the K query reports after the A largest and A smallest are discarded:

$$\widehat{y}_{\mathrm{common}}(x) = \arg\min_{y \in \mathcal{Y}} \mathrm{TM}_A\{w_1(x, y), \ldots, w_K(x, y)\}. \tag{30}$$

Ties use the fixed candidate order. Its accuracy is reported once per attacked tensor, not as a gain attributable to set construction. All reported set sizes and coverage use raw sets, including empty ones. The optional display augmentation

$$\mathcal{C}^+(x) = \begin{cases} \mathcal{C}(x), & \mathcal{C}(x) \neq \varnothing, \\ \{\widehat{y}_{\mathrm{common}}(x)\}, & \mathcal{C}(x) = \varnothing \end{cases} \tag{31}$$

cannot lower coverage, but it changes set utility, does not certify the forced answer, and never replaces the raw results.

O.4 Baseline definitions

Only the proposed rules, guarded symmetric trimming, and the p-value merger carry coverage guarantees against arbitrary reports in both phases, the last two under the conditions in their entries. The unguarded summaries and the Rob-FCP adaptation have no such guarantee. All non-oracle baselines receive the same quantized node tensor and candidate ordering, and no deployable baseline receives honest identities or realized honest ranges.

1. The **honest-mean oracle** (not deployable) and the three proposed rules are as defined in Sections 2 and 3.
2. **Guarded symmetric trimming** averages the reports left after discarding the $m = A$ largest and m smallest (TM_m, Appendix C) and adds a guard of $AS_{\max}/(K - A) + \rho_r$ per channel (Appendix F). It shows, in realized sets, the vacuity that Appendix H diagnoses at the four-bit stress cell. **Unguarded trimming** is the same trimmed mean with no guard, a matched-pipeline diagnostic.
3. The **p-value merger** (partial-conjunction/Rüger merger) keeps a candidate when the $(2A + 1)$th smallest of the K per-node conformal p-values (18) exceeds the cutoff (19), so the A smallest cannot exclude a candidate alone (Proposition G.2). It requires matched quantizers and superuniform honest ranks for the same future target.
4. The **all-node mean** averages all K reports; the **median** averages the two central order statistics when K is even; the **m-Winsorized mean** replaces the bottom and top $m = A$ values by $v_{(m+1)}$ and $v_{(K-m)}$ before averaging. All three use aligned calibration with no guard.
5. **Vector-level Krum** selects, for each calibration example and query, the node's M-score report minimizing the summed squared Euclidean distance to its $K - A - 2$ nearest other reports, with salted-identity tie breaks. It is run only when $K > 2A + 2$ and does not inherit the convergence theory of Blanchard et al. (2017).

6. The **median of means** averages within $2A + 1$ fixed, publicly seeded node groups and takes the median group mean; it is run only when $2A + 1 \leq K$.
7. The **local-marginal comparator** discards alignment (Section 3.1). It forms each node's empirical CDF $\widehat{F}_i$ of its own correct-answer calibration reports, sets q_{LM} to the smallest grid value with $\mathrm{TM}_A(\widehat{F}_1(t), \ldots, \widehat{F}_K(t)) \geq k/(n+1)$, and includes a query candidate when TM_A of its K reports is at most q_{LM}.
8. The **Rob-FCP adaptation** of Kang et al. (2024b) summarizes each node's correct-answer calibration scores as a histogram, keeps the $K - A$ nodes whose histograms lie closest to their nearest neighbors, and sets the threshold from their pooled scores. Queries use the unprotected all-node mean. Defaults were fixed before pilot outcomes and no filtering threshold was tuned; the original coverage theorem is not claimed here.
9. The **no-RAG** and **centralized-union anchors** score with empty context or with retrieval from the union corpus. They report top-one accuracy only and are not federated coverage baselines.

O.5 Metrics and uncertainty

Uncertainty is computed across complete calibration repeats (replicates), because queries within a repeat share the same calibrated rule, node effects, corrupt identities, and attack realization.

Repeat-level summaries. Each replicate r is first reduced to its within-replicate statistic T_r, such as coverage, mean set size, full-set rate, or accuracy. Across R_{rep} replicates, the estimate is the mean $\bar{T}$, and its Monte Carlo standard error is

$$\mathrm{se}(\bar{T}) = \frac{\mathrm{sd}(T_1, \ldots, T_{R_{\mathrm{rep}}})}{\sqrt{R_{\mathrm{rep}}}},$$

with sd the sample standard deviation. A method contrast is first reduced to its within-repeat paired difference and treated identically.

Finite-pool estimand. The real-data estimand averages a metric over random splits, corrupt identities, and attack draws within the observed pool. Fix a held-out pool $\mathcal{P}$ of size N, let C be uniform over its n-subsets, let Π be an independent uniform permutation of the K identities whose first a entries are corrupted, and let Ξ be any independent attack randomness. For a metric loss L,

$$\theta = \mathbb{E}_{C,\Pi,\Xi}\left[\frac{1}{N-n}\sum_{x\in\mathcal{P}\setminus C} L(C, x; \Pi, \Xi)\right].$$

We use 100 paired repeats per task, so the Monte Carlo standard error is the sample standard deviation divided by 10. These intervals quantify finite-pool, corrupt-identity, and attack randomization only. We do not use a question bootstrap, do not claim a superpopulation interval, and do not require a 95% lower bound to exceed $1 - \alpha$ as a test of the theorem.

Simultaneous intervals. A contrast is a metric difference, and a family is the group of contrasts covered by one simultaneous statement. Within each family, limited to coverage, mean size, and full-set rate, intervals use a fixed-seed, 10,000-draw Rademacher multiplier procedure. For replicate-level contrasts d_{rc} (signed coverage relative to $1 - \alpha$, or a paired method difference) with repeat means $\bar{d}_c$, each draw b uses independent equiprobable signs $\xi_{rb} \in \{-1, 1\}$ to form

$$T_b = \max_{c:\,\widehat{\mathrm{se}}(\bar{d}_c)>0}\left|\frac{R_{\mathrm{rep}}^{-1}\sum_r \xi_{rb}(d_{rc} - \bar{d}_c)}{\widehat{\mathrm{se}}(\bar{d}_c)}\right|.$$

The reported approximate 95% simultaneous intervals are $\bar{d}_c \pm q_{.95}\widehat{\mathrm{se}}(\bar{d}_c)$, with $q_{.95}$ the 9,500th smallest of the 10,000 draws; they are asymptotic, not exact. An endpoint the design proves deterministic receives a point interval at its known constant. An endpoint with merely zero observed standard deviation is reported with its estimate, and its adjusted interval is marked unavailable. A non-finite contrast value aborts the family.

Families and diagnostics. The executed analysis has 48 families with 1,978 contrasts: 20 coverage families (860 contrasts) and 28 efficiency families (1,118 paired contrasts), including one 70-contrast family per real task for each of coverage, mean size, and full-set rate. Separate 95% levels do not imply simultaneous coverage across families. Fixed-set, joint-threshold, and deletion are contrasted with the honest-mean oracle; guarded symmetric trimming and the p-merger with deletion; secondary comparators enter no family. Coverage families use every true-label tensor, while size and full-set families use only the full-candidate tensors. Every deterministic theorem inequality is checked before aggregation in exact arithmetic on the registered grid, so it needs no confidence interval. Undercoverage is $[1-\alpha-\widehat{\text{coverage}}]_+$ with $[u]_+ = \max\{u, 0\}$, and retrieval recall is not reported.

P SUPPORTING EMPIRICAL RESULTS AND OUTPUT AUDIT

The supporting results explain where the observed set-size gains come from and how their uncertainty was checked. We describe the numerical audit, then the complete synthetic reference conditions and sensitivity studies, then the complete real-data comparisons and the score-quality context they sit in. Further supporting displays (per-task threat plots, the NLL-score and real-data α sensitivities, and the discrete-score and shifted-query controls) are in the supplement. Settings, cell IDs and attack names follow Appendix O. A cell is *eligible* for a method when it meets the assumptions of that method's guarantee; deliberately invalid cells, such as $a > A$, are kept as negative controls. Set sizes are *raw*: an empty set counts as zero and is never replaced by a forced answer (Equation (31)).

P.1 VERIFICATION AND CORRECTED INFERENCE

Independent reaggregation reproduces the reported means, set-status identities, and paired method differences, checks every output against its recorded inputs, and reproduces all 48 original family critical values. It covers all 77,500 synthetic result files (510,000 method rows) and 72 real-data analysis bundles (264,800 repeat rows). The original files remain immutable.

The constant-column correction. A metric or paired difference that takes the same value in every observed repeat does not justify a zero-width interval, so the registered rule reports its estimate and zero standard error but leaves its interval unavailable unless the design proves it deterministic. Floating-point roundoff in the original aggregation had given some such columns spurious positive standard errors; the correction detects exact equality and otherwise centers with accurate `fsum` summation. When the same 10,000 sign draws per family are replayed, 141 synthetic and 94 real endpoints become unavailable (149 and 114 in total), and some family critical values change. The largest change to a retained bound is 1.12×10^{-5} (synthetic) and 2.45×10^{-4} (real). No synthetic interval changes sign; three real intervals (fixed-set, joint-threshold and deletion coverage-minus-0.9 for MMLU-Med/hash/low–high) move from including zero to a lower endpoint of about 6.41×10^{-5}, a borderline change that is disclosed rather than used as a claim. Scores, splits, attacks and point estimates are unchanged.

Theorem diagnostics. No proposed-envelope containment failure occurs in a theorem-eligible primary cell. Positive failure counters in the retained outputs occur in deliberate $a > A$ tests. Some counters are stored on oracle rows, which are always flagged eligible; the audit instead judges each envelope check by whether $a \le A$ in its cell. Grid membership, hierarchy, and guarded-symmetric containment use exact integer/rational decisions. Diagnostics for outer score bounds and the decoded-mean cancellation reference (the honest-decoded padded set $\mathcal{C}_\alpha^{\mathrm{dec},\mathcal{H}}$ of Corollary E.1, which all three rules equal under exact-budget low–high) use the registered 10^{-12} floating tolerance. Across 500 synthetic replicates, 57 checked combinations of audit and method yield 28,500 repeat-level checks and 570 million coordinate comparisons of budget nesting and high–low extremality, with zero violations. The audit matched the aggregate check counts to their source records; it did not regenerate every stored record of which candidates each set contains, nor the underlying model scores.

P.2 COMPLETE SYNTHETIC REFERENCE CONDITIONS AND COMPARISON INTERVALS

Fixed-set is never larger than either relaxation and covers at least 90% in every reference condition within the declared budget; attack direction sets how much inflation remains. At the low–high endpoint with $a = A$, all three proposed rules coincide with the honest nodes' own comparison

plus reconstruction padding; their common mean size is 0.965, with 8.92% empty sets, so a mean size below one does not mean that every query receives one correct answer. Exceeding the budget, $a = 3 > A = 2$, lowers proposed coverage to 77.13%. Table 6 reports all six reference conditions, including the four-bit stress cell (four-bit reports with $a = A = m = 3$) and the deliberate corruption-budget violation, so the favorable operational result (eight bits, $a = A = 2$) can be read alongside its limits.

Table 6: Fixed-set reduces deletion's set size under stable maximum; the p-merger is smaller under high–low. Entries are coverage (%) above raw mean size, including empty sets, over 500 repeats (nominal 90%). Labels give cell ID and (a, A, b): actual corruption, declared budget, and common report depth. The final $a > A$ row lies outside the guarantees, which require $a \leq A$. The oracle uses honest identities and full precision; rounded 100.00 values need not imply perfect coverage.

Condition; ID; (a, A, b)	Fixed-set	Joint	Deletion	Guarded sym.	p-merger	Oracle
Clean	93.77	95.36	96.25	99.97	98.74	90.02
103; $(0, 2, 8)$	1.036	1.088	1.128	2.530	1.404	0.953
Stable max	94.11	97.84	98.31	99.98	99.03	90.04
094; $(2, 2, 8)$	1.049	1.243	1.298	2.648	1.487	0.955
Low–high	90.54	90.54	90.54	99.82	97.05	90.04
099; $(2, 2, 8)$	0.965	0.965	0.965	1.946	1.222	0.955
High–low	99.73	99.84	99.89	100.00	99.03	90.04
100; $(2, 2, 8)$	1.836	1.995	2.126	3.214	1.487	0.955
Four-bit high–low	100.00	100.00	100.00	100.00	99.72	90.01
102; $(3, 3, 4)$	3.793	3.868	3.934	4.000	1.884	0.956
Low–high violation	77.13	77.13	77.13	99.09	94.57	90.01
101; $(3, 2, 8)$	0.784	0.784	0.784	1.465	1.110	0.956

Fixed membership removes a substantial part of the extra set size introduced by the relaxations: under operational stable maximum, fixed-set removes about three quarters of deletion's excess over the oracle, and its mean size is 19.19% below deletion's. The high–low and four-bit rows also show that fixed membership does not make fixed-set smaller than the p-merger under every attack.

The registered uncertainty calculations confirm the reversal between fixed-set and the p-merger. Under operational high–low, their mean sizes are 1.836 and 1.487, respectively. Table 7 reports the size differences together with conservative intervals obtained from the existing simultaneous comparisons.

Table 7: The fixed-set–p-merger size advantage reverses with the attack. Entries are mean-size differences (fixed-set minus p-merger); negative values favor fixed-set. Conservative intervals are linear combinations of corrected approximate 95% simultaneous bounds. The first three rows use the operational-attacks family; the final row uses the four-bit-stress family. Simultaneity applies within each family, not across both.

Condition; ID	Original family	Mean difference	Lower	Upper
Stable max; 094	operational attacks	-0.4381	-0.4467	-0.4295
Low–high; 099	operational attacks	-0.2572	-0.2612	-0.2531
High–low; 100	operational attacks	+0.3491	+0.3267	+0.3716
Four-bit high–low; 102	four-bit stress	+1.9089	+1.8940	+1.9238

The operational-attacks family is the family of mean-size contrasts in the operational attack cells, including cells 094, 099 and 100.

P.3 Identification and padding diagnostics

Table 8 tests alignment in the two worlds of Proposition B.1 (upper block) and the padding requirement of Proposition F.3 (lower block); designs are in Appendix O.

Table 8: Alignment tracks the honest-mean distribution in both worlds; local marginals overcover in World 1. Insufficient padding gives zero coverage in the separate narrow-support construction. Entries average 200 repeats. Identification uses $n = 9{,}999$ and nominal 75%; padding uses nominal 90%. CDF (cumulative distribution function) error is the maximum absolute difference from truth; threshold error is signed. Padding's population coverage is known analytically. Means are descriptive.

World	Calibration summary	CDF error	Threshold error	Coverage (%)
0	Aligned	0.00855	0.00000	75.064
0	Local marginal	0.00627	+0.00013	75.080
1	Aligned	0.00858	-0.00002	75.063
1	Local marginal	0.24985	+0.09357	99.682

Narrow-support padding	Observed coverage (%)	Exact population (%)
0.0	0.000	0.0
0.025	0.000	0.0
0.05	90.188	90.0
0.1	100.000	100.0

The identification rows show why preserving paired examples matters: in World 1 the local-marginal rule's discrepancy is excessive conservatism, while World 0 shows that local summaries can work in one dependence structure without identifying the relevant law in both. The padding rows illustrate the proved padding requirement and the cost of excessive guarding.

P.4 Complete communication and declared-budget sweeps

The synthetic resolution sweep isolates report precision, which the real-data four-bit stress block (supplement) confounds with a larger budget. Table 9 gives the complete equal-depth sweep, with the same bit depth in both phases, and the four conditions with unequal calibration and query depths.

Table 9: Finer reports reduce fixed-set size, with little change beyond eight bits in this sweep. Entries are 500-repeat descriptive means; b_c/b_q denotes calibration/query bits. Stable maximum uses $a = 2$, with $A = m = 3$ in the equal-depth block and $A = m = 2$ otherwise. Equal-depth cells are independently generated. Daggers mark unmatched-map p-merger results outside its guarantee; directional padding remains eligible.

ID	A	b_c/b_q	FS cov. (%)	FS size	Joint size	Del. size	Sym. size	p size
021	3	1/1	100.00	4.000	4.000	4.000	4.000	4.000
022	3	2/2	100.00	3.768	3.788	3.823	4.000	3.242
023	3	3/3	99.84	2.010	2.473	2.664	4.000	2.149
000	3	4/4	98.71	1.377	1.785	1.932	4.000	1.796
024	3	6/6	96.25	1.136	1.452	1.561	3.978	1.628
025	3	8/8	95.26	1.090	1.385	1.485	3.957	1.588
026	3	16/16	94.93	1.079	1.369	1.467	3.950	1.578
027	3	32/32	94.95	1.078	1.369	1.467	3.950	1.582
094	2	8/8	94.11	1.049	1.243	1.298	2.648	1.487
095	2	2/8	99.90	2.137	2.425	2.502	3.939	1.467†
096	2	4/8	96.66	1.156	1.382	1.452	2.982	1.484†
097	2	8/4	96.64	1.156	1.382	1.453	2.981	1.490†
098	2	8/2	99.89	2.209	2.452	2.574	3.727	1.457†

The equal-depth means are descriptive; they do not identify an optimal bit depth. At one bit, every displayed method returns all four candidates: coverage remains conservative while answer selection disappears. In the unequal-depth rows the p-merger loses its guarantee, which needs matched scoring and quantization in both phases (Appendix G).

Fixed membership also limits the cost of declaring a conservative corruption budget. Table 10 holds clean scores, reports, and actual corrupt identities fixed while increasing A, so these comparisons isolate the declared-budget effect. It covers clean, stable-maximum, and high–low conditions at every displayed budget.

Table 10: Fixed membership limits inflation as the declared budget grows, under clean and stable-maximum reports. Entries are 500-repeat descriptive raw mean sizes, including empty sets; FS denotes fixed-set, with coverage in percent (nominal 90%). Reports use eight bits. Within each attack block, only budget/trim ($m = A$) changes; clean scores and attack realizations are paired.

Attack; ID	a	A	Fixed-set	Joint	Deletion	Sym.	p-merger	FS cov.
Clean; 103	0	2	1.036	1.088	1.128	2.530	1.404	93.77
Clean; 104	0	4	1.106	1.234	1.340	4.000	1.569	95.75
Clean; 107	0	6	1.189	1.431	1.646	4.000	1.845	97.13
Stable max; 094	2	2	1.049	1.243	1.298	2.648	1.487	94.11
Stable max; 105	2	4	1.134	1.553	1.710	4.000	1.737	96.19
Stable max; 108	2	6	1.238	1.996	2.278	4.000	2.320	97.61
High–low; 100	2	2	1.836	1.995	2.126	3.214	1.487	99.73
High–low; 106	2	4	2.351	2.701	2.964	4.000	1.737	99.94
High–low; 109	2	6	3.050	3.468	3.708	4.000	2.320	99.99

A wider budget permits more possible honest groups, but fixed membership avoids much of the additional set inflation that deletion incurs. The clean block shows that a conservative budget also costs efficiency when no node is actually corrupt, while the high–low block shows the larger sets caused by switching attacks.

P.5 Broad synthetic sensitivity and assumption controls

Coverage holds well beyond the reference settings of Appendix P.2: every eligible cell of the broad sensitivity study has mean coverage at or above nominal. Table 11 summarizes coverage across changes in node count, actual corruption, trimming, calibration size, candidate count, dependence, heterogeneity, nominal level, and attack type, together with the two registered interaction blocks, which vary (a, m, b) and (K, γ, η) jointly (Appendix O).

Table 11: Every eligible cell has mean coverage at least nominal. Ranges span 500-repeat cell means, expressed as excess over each cell's nominal coverage in percentage points (pp). Counts follow fixed-set (FS) / symmetric / p-merger order. Each method's range is descriptive, not a confidence interval, and covers only its eligible cells. Corruption and trim violations remain in the complete numeric summaries; Table 6 includes the deliberate budget violation.

Registered block	Eligible cells	FS excess (pp)	Sym. excess (pp)	p excess (pp)
Node count K	3/3/3	7.20–9.75	9.40–10.00	9.46–10.00
Actual corruption a	3/3/3	8.22–8.95	10.00–10.00	9.50–9.73
Symmetric trim m	4/2/4	8.69–8.72	10.00–10.00	9.65–9.66
Calibration size n	5/5/5	8.68–8.86	10.00–10.00	9.65–9.87
Candidate count M	3/3/3	8.70–8.72	10.00–10.00	9.65–9.66
Nominal α	2/2/2	4.54–16.36	5.00–20.00	4.92–18.75
Dependence γ	3/3/3	6.13–10.00	10.00–10.00	8.01–10.00
Heterogeneity η	3/3/3	8.67–9.26	10.00–10.00	9.64–9.78
Stable/switching attacks	6/6/6	7.27–9.99	10.00–10.00	9.01–9.65
a, m, b interaction	20/12/20	4.83–10.00	10.00–10.00	9.01–10.00
K, γ, η interaction	24/24/24	5.40–10.00	9.77–10.00	7.93–10.00

Two further controls (supplement) separate tie handling from exchangeability: on discrete-score cells every method, including the oracle, covers conservatively, while under the shifted-query law, which breaks clean-example exchangeability, all three proposed methods and the p-merger (87.16%) fall below nominal.

The secondary synthetic aggregators test whether a conventional robust summary is sufficient under changes in reporting behavior. These methods use the same scores as the primary comparisons but do not have the proposed guard for the clean honest-mean comparison.

Table 12: Low–high switching causes undercoverage for several aggregators; Krum and local-marginal calibration remain above nominal coverage. Entries are coverage (%) above raw mean size, including empty sets, over 500 paired repeats; settings match Table 6. MoM denotes median of means. These descriptive means lie outside the primary simultaneous families. Switching need not preserve exchangeability; no primary Byzantine-envelope guarantee is asserted for these rows.

Condition; source ID	All-node mean	Median	Winsorized	Unguarded trim	Local marginal	MoM	Krum
Stable max	90.05	90.10	90.00	90.01	95.80	90.03	90.17
094	0.956	0.968	0.956	0.957	1.110	0.975	1.012
Low–high	57.67	81.61	75.30	77.76	92.68	67.57	90.16
099	0.579	0.841	0.764	0.791	1.011	0.685	1.012
High–low	99.05	95.32	97.00	96.43	98.78	98.06	90.18
100	1.440	1.112	1.177	1.143	1.385	1.383	1.013
Four-bit high–low	99.88	97.74	98.63	98.17	99.06	99.20	93.50
102	2.094	1.288	1.372	1.299	1.487	1.798	1.113

Table 12 shows that several unguarded methods give smaller sets and near-nominal coverage under stable maximum, then lose coverage under low–high switching: failures of particular aggregation rules under a specified attack, not a universal failure of robust aggregation. Under high–low and in the four-bit stress cell, the same methods remain competitive.

P.6 Complete real-data coverage and paired uncertainty

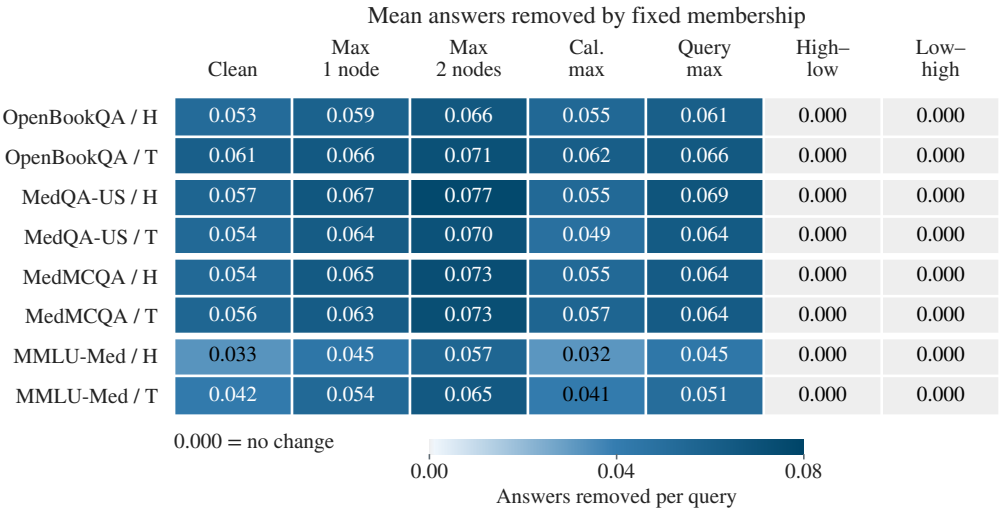
Table 13 summarizes the eight task and partition pairs under stable maximum at $a = A = 2$: fixed-set saves 0.057–0.077 answers per query (paired standard error at most 0.003), removing 34–45% of deletion's excess over the oracle. The oracle keeps 3.509–3.567 of four answers, and the hub's top-ranked answer is 24.10–28.12% accurate against 25% chance (Appendix P.7).

Table 13: Fixed membership removes 33.9–44.7% of the extra answers deletion retains above the honest-mean oracle, in all eight comparisons, with fixed-set coverage above 90%. GPT-2 scores under stable maximum, $a = A = 2$, eight bits, 100 repeats; coverage $\pm$ one Monte Carlo standard error. Figure 5 gives all 56 cells.

Task	Part.	Fixed cov.	Oracle	Fixed	Joint	Del.	Saved	Cut	Merger	Rob-FCP†
OpenBookQA	H	92.76 ± 0.21	3.525	3.654	3.673	3.721	**0.066**	**33.9%**	4.000	91.54/3.591
OpenBookQA	T	92.84 ± 0.20	3.540	3.674	3.691	3.746	**0.071**	**34.7%**	4.000	91.70/3.612
MedQA-US	H	92.28 ± 0.19	3.557	3.659	3.690	3.736	**0.077**	**43.3%**	3.929	90.96/3.604
MedQA-US	T	92.38 ± 0.15	3.567	3.667	3.696	3.737	**0.070**	**41.4%**	3.931	91.33/3.619
MedMCQA	H	92.42 ± 0.15	3.540	3.647	3.674	3.721	**0.073**	**40.5%**	3.964	91.04/3.589
MedMCQA	T	92.47 ± 0.14	3.544	3.654	3.678	3.727	**0.073**	**39.9%**	3.970	91.04/3.594
MMLU-Med	H	91.56 ± 0.21	3.519	3.596	3.630	3.653	**0.057**	**42.4%**	3.895	90.06/3.529
MMLU-Med	T	91.60 ± 0.19	3.509	3.590	3.625	3.656	**0.065**	**44.7%**	3.895	90.40/3.537

H/T: hash/topic-source partition. Sizes are mean answers retained; Saved is Del. − Fixed and Cut is that saving as a share of deletion's excess over the oracle, both before rounding. Merger: partial-conjunction p-value merger; guarded symmetric trimming returns all four answers and is omitted. †Coverage/size of the Rob-FCP adaptation, which defends calibration only, so it lacks a guarantee here; no tested attack targets it.

Figure 5 gives the paired saving from fixed membership in every primary cell, the complete version of the comparison Section 5 summarizes. Fixed-set is smaller than deletion in 40 of the 56 cells, in all 100 repeats of each, and equal in the remaining 16, never larger, as the proved hierarchy requires. The 16 ties are exactly the two switching columns: low–high, where the rules coincide (Appendix P.2), and high–low, which drives all of them to the full answer set.



H: hash; T: topic/source. 100 paired repeats per cell.

Figure 5: Fixed membership removes answers in every cell except the two switching columns, where the rules coincide or saturate. Deletion minus fixed-set mean raw set size over 100 paired repeats for each of the 56 primary real-data cells. Rows are task and partition; columns are the seven threats, with *Max, j node(s)* denoting stable maximum at j corrupt nodes, and *Cal. max* and *Query max* denoting the maximum sent on one channel only. Coverage for every cell is in Table 14.

The full primary table (Table 14) gives coverage and size for all tasks, partitions, and threats at $K = 16$, $A = 2$, eight-bit reports, and nominal 90% coverage. Every task shows the same attack-dependent pattern. Fixed-set is smaller than deletion under stable maximum. High–low drives them to all four answers while the p-merger stays smaller on the medical tasks, and guarded symmetric trimming returns all four candidates throughout. The consistency across partitions concerns these fixed question pools and score functions; it is not evidence of useful clinical answer selection.

The Rob-FCP adaptation remains empirically competitive across the complete primary design. It is smaller than fixed-set in 44 of 56 cells and has lower coverage in 43. Secondary comparator ranges appear in Table 18; threat labels in Tables 14 and 18 follow Figure 5.

Table 14: Fixed-set mean coverage exceeds 90% in every primary cell, with smaller sets than deletion except where they coincide. Entries are coverage (%) / raw mean set size over 100 repeats on fixed question pools, including empty sets. H/T denote hash/topic-source partitions. These are repeat means; registered stable-maximum uncertainty appears in Table 15.

Task	Part.	Threat	Oracle	Fixed	Joint	Deletion	Sym.	Merger
OpenBookQA	H	Clean	90.10/3.526	92.50/3.638	92.70/3.647	93.56/3.691	100.00/4.000	100.00/4.000
OpenBookQA	H	Max, 1	90.06/3.525	92.62/3.647	92.93/3.661	93.88/3.706	100.00/4.000	100.00/4.000
OpenBookQA	H	Max, 2	90.07/3.525	92.76/3.654	93.17/3.673	94.11/3.721	100.00/4.000	100.00/4.000
OpenBookQA	H	Cal. max	90.07/3.525	93.35/3.680	93.52/3.690	94.42/3.735	100.00/4.000	100.00/4.000
OpenBookQA	H	Query max	90.07/3.525	91.93/3.614	92.30/3.630	93.22/3.675	100.00/4.000	100.00/4.000
OpenBookQA	H	High–low	90.07/3.525	100.00/4.000	100.00/4.000	100.00/4.000	100.00/4.000	100.00/4.000
OpenBookQA	H	Low–high	90.07/3.525	91.49/3.593	91.49/3.593	91.49/3.593	100.00/4.000	100.00/4.000
OpenBookQA	T	Clean	90.01/3.541	92.38/3.656	92.58/3.667	93.64/3.717	100.00/4.000	100.00/4.000
OpenBookQA	T	Max, 1	89.98/3.540	92.57/3.664	92.88/3.678	93.90/3.730	100.00/4.000	100.00/4.000
OpenBookQA	T	Max, 2	89.96/3.540	92.84/3.674	93.20/3.691	94.22/3.746	100.00/4.000	100.00/4.000
OpenBookQA	T	Cal. max	89.96/3.540	93.35/3.700	93.56/3.711	94.55/3.762	100.00/4.000	100.00/4.000
OpenBookQA	T	Query max	89.96/3.540	91.93/3.632	92.18/3.646	93.27/3.698	100.00/4.000	100.00/4.000
OpenBookQA	T	High–low	89.96/3.540	100.00/4.000	100.00/4.000	100.00/4.000	100.00/4.000	100.00/4.000
OpenBookQA	T	Low–high	89.96/3.540	91.45/3.609	91.45/3.609	91.45/3.609	100.00/4.000	100.00/4.000
MedQA-US	H	Clean	90.02/3.557	91.96/3.645	92.19/3.656	93.33/3.702	100.00/4.000	98.60/3.924
MedQA-US	H	Max, 1	90.01/3.556	92.10/3.652	92.62/3.674	93.72/3.719	100.00/4.000	98.62/3.926
MedQA-US	H	Max, 2	89.99/3.557	92.28/3.659	93.05/3.690	94.08/3.736	100.00/4.000	98.64/3.929
MedQA-US	H	Cal. max	89.99/3.557	93.27/3.699	93.52/3.710	94.46/3.753	100.00/4.000	98.64/3.929
MedQA-US	H	Query max	89.99/3.557	91.21/3.614	91.66/3.634	92.84/3.683	100.00/4.000	98.40/3.914
MedQA-US	H	High–low	89.99/3.557	100.00/4.000	100.00/4.000	100.00/4.000	100.00/4.000	98.64/3.929
MedQA-US	H	Low–high	89.99/3.557	91.01/3.606	91.01/3.606	91.01/3.606	100.00/4.000	98.38/3.913
MedQA-US	T	Clean	89.96/3.565	92.16/3.656	92.47/3.669	93.46/3.711	100.00/4.000	98.49/3.926
MedQA-US	T	Max, 1	89.96/3.567	92.26/3.661	92.77/3.682	93.81/3.724	100.00/4.000	98.55/3.928
MedQA-US	T	Max, 2	90.01/3.567	92.38/3.667	93.12/3.696	94.06/3.737	100.00/4.000	98.62/3.931

Task	Part.	Threat	Oracle	Fixed	Joint	Deletion	Sym.	Merger
MedQA-US	T	Cal. max	90.01/3.567	93.36/3.705	93.59/3.716	94.50/3.754	100.00/4.000	98.62/3.931
MedQA-US	T	Query max	90.01/3.567	91.45/3.625	91.87/3.645	92.94/3.689	100.00/4.000	98.17/3.913
MedQA-US	T	High–low	90.01/3.567	100.00/4.000	100.00/4.000	100.00/4.000	100.00/4.000	98.62/3.931
MedQA-US	T	Low–high	90.01/3.567	91.25/3.617	91.25/3.617	91.25/3.617	100.00/4.000	98.10/3.910
MedMCQA	H	Clean	89.86/3.538	92.16/3.635	92.39/3.645	93.31/3.689	100.00/4.000	99.13/3.960
MedMCQA	H	Max, 1	89.87/3.538	92.27/3.640	92.70/3.659	93.65/3.705	100.00/4.000	99.17/3.962
MedMCQA	H	Max, 2	89.89/3.540	92.42/3.647	92.99/3.674	93.94/3.721	100.00/4.000	99.21/3.964
MedMCQA	H	Cal. max	89.89/3.540	93.22/3.684	93.42/3.693	94.30/3.738	100.00/4.000	99.21/3.964
MedMCQA	H	Query max	89.89/3.540	91.43/3.605	91.87/3.623	92.88/3.669	100.00/4.000	99.06/3.956
MedMCQA	H	High–low	89.89/3.540	100.00/4.000	100.00/4.000	100.00/4.000	100.00/4.000	99.21/3.964
MedMCQA	H	Low–high	89.89/3.540	91.22/3.596	91.22/3.596	91.22/3.596	100.00/4.000	98.86/3.946
MedMCQA	T	Clean	89.89/3.543	92.14/3.640	92.41/3.651	93.40/3.696	100.00/4.000	99.28/3.966
MedMCQA	T	Max, 1	89.88/3.543	92.30/3.646	92.70/3.665	93.69/3.710	100.00/4.000	99.30/3.968
MedMCQA	T	Max, 2	89.89/3.544	92.47/3.654	92.99/3.678	94.04/3.727	100.00/4.000	99.35/3.970
MedMCQA	T	Cal. max	89.89/3.544	93.22/3.688	93.44/3.698	94.43/3.745	100.00/4.000	99.35/3.970
MedMCQA	T	Query max	89.89/3.544	91.48/3.612	91.88/3.629	92.95/3.676	100.00/4.000	99.20/3.963
MedMCQA	T	High–low	89.89/3.544	100.00/4.000	100.00/4.000	100.00/4.000	100.00/4.000	99.35/3.970
MedMCQA	T	Low–high	89.89/3.544	91.24/3.602	91.24/3.602	91.24/3.602	100.00/4.000	98.95/3.950
MMLU-Med	H	Clean	89.81/3.518	91.40/3.589	91.66/3.600	92.17/3.622	100.00/4.000	97.99/3.891
MMLU-Med	H	Max, 1	89.80/3.519	91.50/3.593	91.99/3.615	92.59/3.639	100.00/4.000	98.03/3.893
MMLU-Med	H	Max, 2	89.82/3.519	91.56/3.596	92.36/3.630	92.98/3.653	100.00/4.000	98.06/3.895
MMLU-Med	H	Cal. max	89.82/3.519	92.62/3.640	92.89/3.649	93.45/3.672	100.00/4.000	98.06/3.895
MMLU-Med	H	Query max	89.82/3.519	90.69/3.558	91.19/3.579	91.71/3.603	100.00/4.000	97.67/3.874
MMLU-Med	H	High–low	89.82/3.519	100.00/4.000	100.00/4.000	100.00/4.000	100.00/4.000	98.06/3.895
MMLU-Med	H	Low–high	89.82/3.519	90.53/3.551	90.53/3.551	90.53/3.551	100.00/4.000	97.66/3.873
MMLU-Med	T	Clean	89.74/3.510	91.42/3.583	91.72/3.595	92.38/3.625	100.00/4.000	97.89/3.888
MMLU-Med	T	Max, 1	89.75/3.510	91.54/3.587	92.08/3.611	92.75/3.641	100.00/4.000	97.95/3.891
MMLU-Med	T	Max, 2	89.73/3.509	91.60/3.590	92.37/3.625	93.11/3.656	100.00/4.000	98.01/3.895
MMLU-Med	T	Cal. max	89.73/3.509	92.63/3.635	92.95/3.648	93.66/3.677	100.00/4.000	98.01/3.895
MMLU-Med	T	Query max	89.73/3.509	90.61/3.549	91.12/3.571	91.80/3.600	100.00/4.000	97.57/3.870
MMLU-Med	T	High–low	89.73/3.509	100.00/4.000	100.00/4.000	100.00/4.000	100.00/4.000	98.01/3.895
MMLU-Med	T	Low–high	89.73/3.509	90.43/3.542	90.43/3.542	90.43/3.542	100.00/4.000	97.56/3.869

The registered real-data contrasts quantify remaining robustness overhead and the cost of alternative constructions. In Table 15, fixed-set minus oracle measures that overhead; symmetric or merger minus deletion compares those alternatives with the scalable deletion rule. The corrected intervals are simultaneous within each registered task/metric family, without adjustment across all tasks or all three metrics. "Full" rows report the full-set rate, the share of queries whose set contains all M candidates.

Table 15: Under stable maximum ($a = A = 2$), all displayed contrasts indicate excess conservatism. Entries are paired mean [lower, upper] with approximate 95% simultaneous Monte Carlo intervals within each 70-contrast task/metric family, for fixed question pools. Size is in candidates; full-set rate is in percentage points (pp). H/T denote hash/topic-source partitions. Positive values mean the first method exceeds its reference.

Task	Part.	Metric	Fixed − oracle	Sym. − del.	Merger − del.
OpenBookQA	H	Size	0.129 [0.123, 0.136]	0.279 [0.257, 0.301]	0.279 [0.257, 0.301]
OpenBookQA	H	Full (pp)	8.180 [7.762, 8.598]	19.135 [17.778, 20.492]	19.135 [17.778, 20.492]
OpenBookQA	T	Size	0.134 [0.127, 0.141]	0.254 [0.230, 0.278]	0.254 [0.230, 0.278]
OpenBookQA	T	Full (pp)	8.168 [7.709, 8.627]	17.577 [16.031, 19.123]	17.577 [16.031, 19.123]
MedQA-US	H	Size	0.101 [0.096, 0.106]	0.264 [0.250, 0.278]	0.193 [0.180, 0.205]
MedQA-US	H	Full (pp)	7.231 [6.863, 7.600]	23.331 [22.220, 24.442]	16.446 [15.447, 17.445]
MedQA-US	T	Size	0.100 [0.095, 0.104]	0.263 [0.250, 0.276]	0.194 [0.181, 0.206]
MedQA-US	T	Full (pp)	7.522 [7.160, 7.884]	23.347 [22.312, 24.381]	16.673 [15.723, 17.622]
MedMCQA	H	Size	0.108 [0.104, 0.111]	0.279 [0.263, 0.296]	0.243 [0.225, 0.261]
MedMCQA	H	Full (pp)	7.346 [7.090, 7.602]	23.052 [21.840, 24.264]	19.855 [18.401, 21.309]
MedMCQA	T	Size	0.110 [0.106, 0.114]	0.273 [0.258, 0.289]	0.243 [0.226, 0.260]
MedMCQA	T	Full (pp)	7.449 [7.174, 7.724]	22.524 [21.349, 23.699]	19.828 [18.451, 21.205]
MMLU-Med	H	Size	0.077 [0.073, 0.082]	0.347 [0.327, 0.368]	0.242 [0.224, 0.261]
MMLU-Med	H	Full (pp)	5.177 [4.847, 5.508]	27.764 [26.243, 29.284]	18.835 [17.412, 20.258]
MMLU-Med	T	Size	0.081 [0.077, 0.085]	0.344 [0.325, 0.363]	0.239 [0.223, 0.255]
MMLU-Med	T	Full (pp)	5.468 [5.164, 5.773]	27.668 [26.279, 29.057]	18.774 [17.518, 20.030]

These contrasts are not the direct fixed-set-minus-deletion saving, which remains a descriptive paired comparison, not a registered interval contrast; its paired estimate and Monte Carlo standard error are retained as machine-readable artifacts. To express the saving relative to deletion's excess size, let $\bar{S}_u$ be method u's repeat-mean raw set size. The reported ratio, the Cut column of Table 13, is $(\bar{S}_{\mathrm{del}} - \bar{S}_{\mathrm{fs}})/(\bar{S}_{\mathrm{del}} - \bar{S}_{\mathrm{oracle}})$, computed separately for each stable-maximum task/partition cell. This ratio of repeat means is neither a mean of repeat-level ratios nor an additional registered confidence contrast.

P.7 SCORE-QUALITY ANCHORS AND SECONDARY CALIBRATION

The real-data gains reduce robustness overhead in a weak-score regime. Table 16 compares the common ranker (Equation (30)) with no-retrieval and whole-corpus anchors. With broad oracle sets and near-chance point accuracy, a smaller robust set here is a reduction in overhead, not reliable forced-answer selection.

Table 16: Whole-corpus retrieval has the highest observed accuracy, but all scorers remain near 25% chance. Entries are percentage means ± one Monte Carlo standard error over 100 splits of fixed question pools. The common ranker is shared by federated methods; No RAG omits retrieval, and Union RAG retrieves four whole-corpus passages. No RAG and Union RAG scores are reused across partitions.

Task	Partition	Common ranker	No RAG	Union RAG
OpenBookQA	Hash	24.62 ± 0.09	25.00 ± 0.09	27.38 ± 0.10
OpenBookQA	Topic/source	24.59 ± 0.09	25.00 ± 0.09	27.38 ± 0.10
MedQA-US	Hash	24.10 ± 0.09	24.57 ± 0.08	26.03 ± 0.09
MedQA-US	Topic/source	24.71 ± 0.08	24.57 ± 0.08	26.03 ± 0.09
MedMCQA	Hash	28.12 ± 0.03	26.89 ± 0.02	29.72 ± 0.03
MedMCQA	Topic/source	27.52 ± 0.02	26.89 ± 0.02	29.72 ± 0.03
MMLU-Med	Hash	27.21 ± 0.10	27.40 ± 0.10	28.83 ± 0.11
MMLU-Med	Topic/source	27.17 ± 0.11	27.40 ± 0.10	28.83 ± 0.11

The registered secondary analyses (supplement) show that the full sets are partly a matter of score design. Under the separately calibrated NLL score of Appendix O.3, no rule returns all four candidates for every query in every repeat of any cell; across its 56 cells fixed-set mean size is 3.53–3.92 and mean coverage 90.20–98.58%. It is a secondary analysis, not an outcome-selected replacement for the primary score, and leaves the weak-model limitation intact. The real-data α sensitivity shows that lower nominal coverage permits smaller sets.

P.8 REAL-DATA BOUNDARY CONTROLS AND SECONDARY COMPARATORS

Exceeding the declared budget can break coverage. The real-data violation repeats the low–high test with $a = 3 > A = 2$, outside the eligible primary cells; real fixed-set coverage falls to 68.74–75.92%.

Table 17: Underestimating corruption lowers all three proposed rules' coverage to 68.74–75.92%, below nominal 90%. The deliberate violation uses $a = 3 > A = 2$, low–high reporting, and eight-bit reports. Entries are descriptive mean coverage percentages on fixed question pools; the guarantees, which require $a \leq A$, do not apply.

Task	Partition	Fixed	Joint	Deletion	Symmetric	Merger
OpenBookQA	Hash	68.74	68.74	68.74	100.00	100.00
OpenBookQA	Topic/source	69.38	69.38	69.38	100.00	100.00
MedQA-US	Hash	72.65	72.65	72.65	100.00	98.06
MedQA-US	Topic/source	72.51	72.51	72.51	100.00	97.76
MedMCQA	Hash	72.90	72.90	72.90	100.00	98.52
MedMCQA	Topic/source	72.08	72.08	72.08	100.00	98.46
MMLU-Med	Hash	75.92	75.92	75.92	100.00	97.35
MMLU-Med	Topic/source	74.94	74.94	74.94	100.00	97.25

The symmetric and merger results remain at or above nominal in these cells, but that does not restore a guarantee whose declared budget has been exceeded.

The secondary real-data comparators show both competitive outcomes and attack-sensitive failures. Table 18 shows ranges over all four tasks and both partitions at four reference threats; the complete numerical artifact retains all seven threats and individual standard errors.

Table 18: Low–high reporting drives several comparators below 90% coverage; Krum and the Rob-FCP adaptation remain above it in these cells. Entries span repeat means across four tasks and both partitions (eight fixed-pool cells), not confidence intervals. Coverage and full-set rate are percentages; size counts retained answers. All seven threats and individual Monte Carlo standard errors remain in the numeric artifact.

Method	Threat	Coverage range	Size range	Full-set range
All-node mean	Clean	89.74–90.15	3.509–3.565	62.11–70.00
All-node mean	Max, 2	89.70–90.10	3.508–3.568	62.02–69.94
All-node mean	High–low	100.00–100.00	4.000–4.000	100.00–100.00
All-node mean	Low–high	54.57–60.94	2.126–2.329	5.83–8.34
Median	Clean	90.08–90.72	3.523–3.602	63.45–72.38
Median	Max, 2	90.04–90.79	3.520–3.603	63.22–72.50
Median	High–low	91.13–91.88	3.571–3.644	66.61–75.21
Median	Low–high	88.64–89.30	3.460–3.548	59.38–68.80
Winsorized	Clean	89.77–90.04	3.507–3.582	62.23–70.60
Winsorized	Max, 2	89.80–90.10	3.508–3.582	62.41–70.49
Winsorized	High–low	91.88–92.96	3.609–3.685	68.94–76.43
Winsorized	Low–high	86.39–87.28	3.384–3.444	54.49–62.74
Unguarded trim	Clean	89.85–90.07	3.510–3.581	62.50–70.73
Unguarded trim	Max, 2	89.84–90.08	3.512–3.582	62.74–70.56
Unguarded trim	High–low	91.55–92.50	3.597–3.671	68.16–75.69
Unguarded trim	Low–high	86.99–87.90	3.407–3.465	55.97–63.99
Median of means	Clean	89.82–90.20	3.511–3.585	62.44–70.69
Median of means	Max, 2	89.85–90.19	3.513–3.586	62.66–70.77
Median of means	High–low	92.32–93.01	3.626–3.694	70.14–76.81
Median of means	Low–high	85.95–87.15	3.360–3.437	53.02–62.92
Krum	Clean	90.16–90.93	3.524–3.614	63.53–72.63
Krum	Max, 2	90.18–90.90	3.522–3.615	63.02–72.47
Krum	High–low	90.19–90.90	3.521–3.615	63.00–72.40
Krum	Low–high	90.18–90.95	3.522–3.613	63.02–72.60
Local marginal	Clean	90.16–91.27	3.533–3.620	64.10–73.81
Local marginal	Max, 2	89.85–90.89	3.515–3.599	63.00–72.76
Local marginal	High–low	91.55–92.84	3.596–3.685	68.10–77.66
Local marginal	Low–high	89.42–90.40	3.496–3.580	61.74–71.54
Rob-FCP adaptation	Clean	91.05–92.29	3.574–3.661	66.67–75.85
Rob-FCP adaptation	Max, 2	90.06–91.70	3.529–3.619	63.65–74.29
Rob-FCP adaptation	High–low	100.00–100.00	4.000–4.000	100.00–100.00
Rob-FCP adaptation	Low–high	90.06–91.70	3.529–3.619	63.65–74.29

Several robust summaries remain near nominal under stable maximum but undercover under low–high in Table 18. Krum and the Rob-FCP adaptation remain above nominal throughout the displayed low–high cells, while local-marginal coverage spans both sides of nominal. Thus an aggregator can be competitive on these fixed tasks without inheriting the proposed guarantee against arbitrary reports in both phases.

The scoring audit confirms complete model outputs for all 246,330 held-out score vectors, with zero fallbacks caused by a question or answer exceeding the token limit.

Q Assumption and claim checklist

The coverage and minimality guarantees (Theorems 4.1 and 4.3) rest on the five conditions below and on nothing stronger hidden in the proofs; items 2 and 3 also record what the baselines and the efficiency bounds add.

1. Clean calibration and future examples are exchangeable (their joint law is unchanged by reordering) conditional on the scoring system frozen before calibration.
2. The honest node set is fixed, authenticated, and contains at least $K - A$ nodes, with $A < K$ (baseline requirements: Appendix F and Proposition G.2).
3. Honest reports satisfy registered directional reconstruction-error bounds. Neither a bounded decoded-report alphabet nor an honest-range certificate (a registered bound on the range of the honest scores) is required for the proposed rules' coverage when every honest node responds. A bounded decoded-report alphabet is required for the pathwise upper efficiency sandwich; the expected-size bounds additionally require fixed certificate widths and the future X to be independent of the clean calibration split conditional on the frozen system, as stated in Corollary L.3. The alternative bound (29) also assumes a bounded true-label density.

4. Honest nodes evaluate every shared calibration candidate and every future/query candidate with the same frozen score functions s_i, whose honest average $s_{\mathcal{H}}$ (Equation (1)) is the clean target.
5. The finite candidate space, candidate ordering, and coordinate convention are frozen across phases. Honest nodes evaluate every candidate; during calibration the hub selects the correct-answer (gold) coordinate only after receiving the complete score vectors.

Beyond Theorem 4.1's stated exclusions, the theorems require no i.i.d. nodes, unbiased noise, symmetric quantization or Dvoretzky–Kiefer–Wolfowitz (DKW) approximation; the p-merger's extra assumption is in Proposition G.2.

R RESTRICTED SUBSET FAMILIES AND A PROMPT-INDUCED PANEL FAILURE

This section makes two points. First, the fixed-set rule of Section 3 can search any family of subsets chosen from data instead of $\mathfrak{H}_A$. Coverage then costs one tail probability, but clean-score integrity holds only while the true honest set is still searched (Proposition R.1). Second, when bad reports are *induced* rather than constructed, the proposed rules hold coverage in every regime, while the unguarded all-node mean does not when only calibration is misled.

A misled judge emits a finite report, so Theorem 4.1 already covers it; the measurements locate ordinary misbehavior inside the proved envelope rather than test the worst case. The clean reference (1) is the unquantized faithful-instruction mean over the judges that drew the faithful instruction.

The panel. Sixteen nodes each hold a disjoint shard of the $1{,}326$ OpenBookQA evidence facts. Each retrieves four passages per question by BM25 (Robertson & Zaragoza, 2009) on its own shard, querying with the question and its four options. Each then scores all four options of $1{,}000$ held-out questions with `Llama-3.2-3B-Instruct` (Grattafiori et al., 2024). The score of an option is one minus the model's probability of its letter after "Answer:", renormalized over the four letters, so smaller scores favor inclusion as in Section 2. Every question is scored under a faithful instruction, to answer from the retrieved evidence, and an injected one, to answer with the first listed option whatever the evidence says.

Each of 500 replicates permutes the questions into an audit pool of 100, a calibration split of $n = 333$ and 567 evaluation questions. It draws the misled judges independently, fourteen with probability 0.02 and two with probability 0.5, and reports eight-bit scores at $K = 16$ and budgets $A = 2, \dots, 5$. Only the two trust-based comparators read the audit pool. A judge's *audit error* is the fraction of audit questions on which its smallest reported score is not the correct answer's alone (ties count as errors). Over all judge–question pairs, the audit error is 0.2990 under the faithful instruction and 0.5874 under the injected one. The injection thus makes a judge content-blind in one direction rather than noisy. The injected audit error stays well below one partly because the correct answer is listed first on 26.4% of questions.

R.1 A SUBSET FAMILY CHOSEN FROM DATA

For a family $\mathcal{G}$ of nonempty subsets of $[K]$, define the restricted margin and rule by

$$\Delta_{\mathcal{G},k}(V, w) = \min_{H \in \mathcal{G}} \left\{ \frac{1}{|H|} \sum_{i \in H} w_i - T_k(H; V) \right\},$$

$$\mathcal{C}^{\mathcal{G}}_{\alpha}(x) = \{ y : \Delta_{\mathcal{G},k}\{V, w(x, y)\} \le g_\rho \},$$

so that $\mathcal{C}^{\mathfrak{H}_A}_{\alpha}$ is the fixed-set rule (8). Extending Equation (1), write $s_H(x, y) = |H|^{-1} \sum_{i \in H} s_i(x, y)$ for the clean mean over a nonempty $H \subseteq [K]$, $R^H_j = s_H(X_j, Y_j)$ for its clean calibration scores, and $R^H_{(k)}$ for the kth smallest of $R^H_1, \dots, R^H_n$.

Proposition R.1 (Restricted subset families)**.** *Assume $k \le n$ and that the registered directional bounds* (2) *hold for the honest nodes. Assume also that $\mathcal{A}$ is drawn once before calibration and held fixed for the whole calibration-to-query comparison; this is the first condition of the failure law in Appendix N and has a deterministic honest set as its degenerate case. Let $\mathcal{G}$ be a random family of*

nonempty subsets of $[K]$, chosen as a measurable function of audit data $\mathcal{D}$, and assume the clean examples $Z_1, \dots, Z_{n+1}$ are exchangeable conditional on $(\mathcal{D}, \mathcal{A})$. Write

$$\varepsilon_{\mathcal{G}} = \mathbb{P}\{\text{ no } H \in \mathcal{G} \text{ satisfies } H \subseteq \mathcal{H}\}.$$

Then, uniformly over all finite Byzantine calibration and query reports,

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}_\alpha^{\mathcal{G}}(X_{n+1})\} \;\geq\; \frac{k}{n+1}\,(1-\varepsilon_{\mathcal{G}})\,. \tag{32}$$

Moreover, clean-score integrity for $s_{\mathcal{H}}$, that is $\mathcal{C}_\alpha^{\circ}(x) \subseteq \mathcal{C}_\alpha^{\mathcal{G}}(x)$ for every query and admissible transcript, holds whenever $\mathcal{H} \in \mathcal{G}$, and it can fail when $\mathcal{H} \notin \mathcal{G}$. For $k = n+1$, $\mathcal{C}_\alpha^{\mathcal{G}}$ returns $\mathcal{Y}$ and coverage is one.

Proof. Let $E_{\mathcal{G}}$ be the event that some $H \in \mathcal{G}$ satisfies $H \subseteq \mathcal{H}$, so $\mathbb{P}(E_{\mathcal{G}}) = 1 - \varepsilon_{\mathcal{G}}$.

Step 1: one all-honest subset suffices, pathwise. Fix any $H_0 \in \mathcal{G}$ with $H_0 \subseteq \mathcal{H}$; it is nonempty. Averaging the directional bounds (2) over $i \in H_0$ gives, for each calibration example j,

$$R_j^{H_0} - \rho_{\mathrm{cal}}^{-} \;\leq\; c_j(H_0; V) \;\leq\; R_j^{H_0} + \rho_{\mathrm{cal}}^{+},$$

and by monotonicity of order statistics

$$R_{(k)}^{H_0} - \rho_{\mathrm{cal}}^{-} \;\leq\; T_k(H_0; V) \;\leq\; R_{(k)}^{H_0} + \rho_{\mathrm{cal}}^{+}.$$

At the query, on the true label, honesty of H_0 gives

$$\frac{1}{|H_0|}\sum_{i \in H_0} w_i(X_{n+1}, Y_{n+1}) \;\leq\; R_{n+1}^{H_0} + \rho_{\mathrm{qry}}^{+}.$$

The minimum defining $\Delta_{\mathcal{G},k}$ runs over a family containing H_0, so

$$\Delta_{\mathcal{G},k} \;\leq\; R_{n+1}^{H_0} + \rho_{\mathrm{qry}}^{+} - R_{(k)}^{H_0} + \rho_{\mathrm{cal}}^{-}.$$

Hence on the clean rank event $\{R_{n+1}^{H_0} \leq R_{(k)}^{H_0}\}$ for the H_0-mean, $\Delta_{\mathcal{G},k} \leq \rho_{\mathrm{qry}}^{+} + \rho_{\mathrm{cal}}^{-} = g_\rho$ and $Y_{n+1} \in \mathcal{C}_\alpha^{\mathcal{G}}(X_{n+1})$.

Step 2: the choice of H_0 does not break exchangeability. Condition on $(\mathcal{D}, \mathcal{A})$. On $E_{\mathcal{G}}$, let H_0 be the lexicographically least $H \in \mathcal{G}$ with $H \subseteq \mathcal{H}$. Since $\mathcal{G}$ is a function of $\mathcal{D}$ and $\mathcal{H} = [K] \setminus \mathcal{A}$, H_0 is deterministic given $(\mathcal{D}, \mathcal{A})$ and does not depend on the calibration scores. The clean examples are exchangeable given $(\mathcal{D}, \mathcal{A})$, and $R_j^{H_0}$ applies one fixed function to each, so Lemma L.1 gives

$$\mathbb{P}\{R_{n+1}^{H_0} \leq R_{(k)}^{H_0} \mid \mathcal{D}, \mathcal{A}\} \;\geq\; \frac{k}{n+1}.$$

Step 3: average. The event $E_{\mathcal{G}}$ is $(\mathcal{D}, \mathcal{A})$-measurable, so combining Steps 1 and 2 and discarding the complementary event,

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}_\alpha^{\mathcal{G}}(X_{n+1})\} \;\geq\; \mathbb{E}\left[\mathbf{1}_{E_{\mathcal{G}}}\,\frac{k}{n+1}\right] \;=\; \frac{k}{n+1}\,(1-\varepsilon_{\mathcal{G}})\,,$$

which is Equation (32).

Step 4: integrity when $\mathcal{H} \in \mathcal{G}$. Apply Step 1 with $H_0 = \mathcal{H}$. The clean rank event for the $\mathcal{H}$-mean is $\{Y_{n+1} \in \mathcal{C}_\alpha^{\circ}(X_{n+1})\}$ by Equation (3), and the same pathwise chain for an arbitrary candidate y gives $\mathcal{C}_\alpha^{\circ}(x) \subseteq \mathcal{C}_\alpha^{\mathcal{G}}(x)$. This is the containment argument of Theorem 4.3 with $\mathfrak{H}_A$ replaced by $\mathcal{G}$.

Step 5: integrity can fail when $\mathcal{H} \notin \mathcal{G}$. Take $K = 2$, $a = 0$, so $\mathcal{H} = \{1, 2\}$, and $\mathcal{G} = \{\{1\}\}$. Let the registered reconstruction errors be zero, so $g_\rho = 0$, and take $n = 1$ and $\alpha = 1/2$, so $k = 1$. Let both nodes report 0 at calibration, so $T_1(\{1\}; V) = 0$ and $R_{(1)} = 0$. For a candidate y, let node 1 report $+1$ and node 2 report -1. The oracle's query score is $0 \leq R_{(1)}$, so it admits y; but $\Delta_{\mathcal{G},1} = 1 > 0 = g_\rho$, so $\mathcal{C}_\alpha^{\mathcal{G}}$ excludes it.

Finally, for $k = n+1$ the full-set convention of Section 2 returns $\mathcal{Y}$. □

An audit run on its own held-out block of questions, as here, meets the exchangeability hypothesis; what the hypothesis excludes is an audit that inspects the clean calibration scores. Neither $\varepsilon_{\mathcal{G}}$ nor whether $\mathcal{H} \in \mathcal{G}$ is observable at deployment.

R.2 THE TWO COMPARATORS AS COROLLARIES

Both comparators act on the judges' audit errors. *V1, trust pruning*, removes the two judges with the highest audit error and runs the fixed-set rule on the other fourteen at budget $\max(0, A-2)$. *V2, an audited budget*, flags the judges whose audit error exceeds 0.4 and runs the fixed-set rule on all sixteen at the audited budget A'. This is the smallest budget whose tail under the corner rule of Lemma N.1, with flagged judges failing with probability $\bar{p}_2 = 0.5$ and the rest with $\bar{p}_1 = 0.02$, is at most $\delta = 0.01$.

Corollary R.2 (An audited budget)**.** *Let $A' < K$ be an audited budget, chosen from the audit data as in Proposition R.1, and take $\mathcal{G} = \mathfrak{H}_{A'}$. Then $\mathcal{H} \in \mathcal{G}$ if and only if $a \le A'$, so $\varepsilon_{\mathcal{G}} = \mathbb{P}\{a > A'\}$, and integrity holds on $\{a \le A'\}$. If $A' \ge A_0$ on every realization for a constant $A_0 < K$, then $\varepsilon_{\mathcal{G}} \le \varepsilon_{A_0}$, the failure law's own tail. If the audit returns a deterministic A', the proposition reduces to Corollary 4.2 at A'.*

Proof. Since $A' < K$, every member of $\mathfrak{H}_{A'}$ is nonempty, and a set $H \subseteq \mathcal{H}$ with $|H| \ge K - A'$ exists if and only if $a \le A'$, which is also the condition $\mathcal{H} \in \mathfrak{H}_{A'}$. If $A' \ge A_0$ always, then $\{a > A'\} \subseteq \{a > A_0\}$; for deterministic A', Equation (32) is Equation (12) at A'. □

Corollary R.3 (Removing the least-trusted nodes)**.** *Let $S \subseteq [K]$ with $|S| \le A$ be a removed set chosen as in Proposition R.1, and take $\mathcal{G} = \{H \in \mathfrak{H}_A : H \cap S = \emptyset\}$, the family the fixed-set rule enumerates when it is run on the $K - |S|$ survivors at budget $A - |S|$. Then $\mathcal{H} \in \mathcal{G}$ if and only if $a \le A$ and every removed node was Byzantine, $S \subseteq \mathcal{A}$; on that event Proposition R.1's integrity guarantee applies. Further,*

$$\varepsilon_{\mathcal{G}} = \mathbb{P}\{|\mathcal{H} \setminus S| < K - A\}.$$

Proof. A set H lies in $\mathcal{G}$ and satisfies $H \subseteq \mathcal{H}$ exactly when $H \subseteq \mathcal{H} \setminus S$ and $|H| \ge K - A$. Since $K - A \ge 1$, such an H exists if and only if $|\mathcal{H} \setminus S| \ge K - A$. Membership $\mathcal{H} \in \mathcal{G}$ asks $a \le A$ and $\mathcal{H} \cap S = \emptyset$, that is $S \subseteq \mathcal{A}$. □

An audited budget keeps every node in the reference, but its advertised tail can understate the true one, since the audit supplies a realized count of flagged nodes rather than a bound on the law. Removal changes the estimand to the survivors' mean, so V1's smaller sets answer a different question.

R.3 WHAT THE MEASUREMENTS SHOW

The proposed rules hold coverage in every regime; the unguarded mean does not. A judge that drew the injected instruction is misled at calibration only, at the query only, or in both phases (Table 19); the three regimes share each replicate's misled set, oracle and audit. The three proposed rules clear $k/(n+1)$ in every regime and budget, over all replicates and over the budget-respecting ones ($a \le A$). Misleading only calibration drops the all-node mean to 0.8862; with per-cell Monte Carlo standard errors of about 0.001, that is about 15 standard errors below $k/(n+1)$.

Why the calibration channel is the weak point. The injection compresses the upper tail of the correct-answer score, which lowers the calibration threshold. Across all questions it raises the node-mean correct-answer score (median 0.294 to 0.694). But the threshold reads an upper order statistic, and on the 100 questions in the faithful panel's upper decile the injected judges score *lower*: 0.656 against 0.925. The node-mean score's ninetieth percentile is lowest when about a quarter of the panel is misled, and the realized corruption averages 1.27 judges, so the sparse regime is the most exposed.

Fixed membership removes 45–75% of deletion's excess. At $A = 2$ fixed-set saves 0.21–0.30 answers per query over deletion across the three regimes, 45–75% of deletion's excess over the honest-mean oracle (Table 20). Each extra unit of budget costs deletion more than fixed-set, so from $A = 2$ to $A = 5$ the saving roughly doubles, to 0.43–0.58 answers per query (Table 20; Figure 4b shows the both-phases regime).

Trust-based pruning answers a different question. At $A = 2$, V1 loses clean-score integrity on 308 of the 464 budget-respecting replicates, while fixed-set and V2 lose it on none (Table 21). The losses sit where pruning discards honest judges (Corollary R.3); the table notes give the breakdown and the limits of V2's tail.

Table 19: Every proposed rule clears its guaranteed floor in every regime and at every declared budget, while misleading only the calibration channel drops the unguarded all-node mean to 0.8862, below the nominal 0.90. Panel headings name the phases in which a misled judge misreports. Because membership is drawn, ε_A is the exact probability that more than A judges are misled and the guaranteed floor is $\{k/(n+1)\}(1-\varepsilon_A)$ (Corollary 4.2). The final panel keeps, for the proposed rules, only replicates with $a \le A$. Every entry in the final panel clears $k/(n+1) = 0.9012$, the tightest being fixed-set at $A = 2$ with only the calibration channel misled, at 0.9115 over 464 replicates. The honest-mean oracle (not deployable) and the all-node mean (no guarantee) are references. Means over 500 replicates; $n = 333$.

Rule	$A = 2$	$A = 3$	$A = 4$	$A = 5$
Misled at: both phases				
ε_A	0.0777	0.0090	0.0007	3.7×10^{-5}
Guaranteed floor	0.8312	0.8931	0.9006	0.9012
Honest-mean oracle	0.9010	0.9010	0.9010	0.9010
All-node mean	0.9009	0.9009	0.9009	0.9009
Fixed-set	0.9247	0.9316	0.9379	0.9444
Joint-threshold	0.9332	0.9420	0.9495	0.9564
Deletion	0.9516	0.9642	0.9735	0.9791
p-merger	0.9814	0.9808	0.9802	0.9779
Misled at: calibration only				
ε_A	0.0777	0.0090	0.0007	3.7×10^{-5}
Guaranteed floor	0.8312	0.8931	0.9006	0.9012
Honest-mean oracle	0.9010	0.9010	0.9010	0.9010
All-node mean	0.8862	0.8862	0.8862	0.8862
Fixed-set	0.9105	0.9172	0.9237	0.9301
Joint-threshold	0.9151	0.9240	0.9325	0.9405
Deletion	0.9362	0.9523	0.9636	0.9689
p-merger	0.9820	0.9812	0.9805	0.9782
Misled at: query only				
ε_A	0.0777	0.0090	0.0007	3.7×10^{-5}
Guaranteed floor	0.8312	0.8931	0.9006	0.9012
Honest-mean oracle	0.9010	0.9010	0.9010	0.9010
All-node mean	0.9172	0.9172	0.9172	0.9172
Fixed-set	0.9349	0.9411	0.9468	0.9523
Joint-threshold	0.9384	0.9456	0.9523	0.9585
Deletion	0.9539	0.9654	0.9730	0.9784
p-merger	0.9797	0.9795	0.9790	0.9770
Proposed rules, restricted to replicates with $a \le A$				
Fixed-set (both phases)	0.9243	0.9317	0.9379	0.9444
Joint-threshold (both phases)	0.9329	0.9421	0.9495	0.9564
Deletion (both phases)	0.9516	0.9643	0.9735	0.9791
Fixed-set (calibration only)	0.9115	0.9173	0.9237	0.9301
Joint-threshold (calibration only)	0.9162	0.9241	0.9325	0.9405
Deletion (calibration only)	0.9373	0.9524	0.9636	0.9689
Fixed-set (query only)	0.9339	0.9411	0.9468	0.9523
Joint-threshold (query only)	0.9374	0.9457	0.9523	0.9585
Deletion (query only)	0.9534	0.9655	0.9730	0.9784

Table 20: Fixed-set is the smallest of the three proposed rules in every panel and removes 45–75% of deletion's excess over the oracle at $A = 2$; each extra unit of declared budget costs it 0.048 to 0.067 candidates per query, against 0.108 to 0.175 for deletion. Entries are mean numbers of candidates per query out of $M = 4$. No rule returned an empty set on any replicate, so every miss is a nonempty set that excludes the gold answer. Panels are split by misled phase, as in Table 19. The oracle's gap to a proposed rule is the price of not knowing which judges are honest; the all-node mean carries no guarantee. Means over 500 replicates; $n = 333$.

Rule	$A = 2$	$A = 3$	$A = 4$	$A = 5$
Mean set size, misled at: both phases				
Honest-mean oracle	1.718	1.718	1.718	1.718
All-node mean	1.653	1.653	1.653	1.653
Fixed-set	1.869	1.935	2.000	2.067
Joint-threshold	1.959	2.056	2.145	2.230
Deletion	2.167	2.342	2.505	2.642
p-merger	2.820	2.756	2.726	2.651
Mean set size, misled at: calibration only				
Honest-mean oracle	1.718	1.718	1.718	1.718
All-node mean	1.606	1.606	1.606	1.606
Fixed-set	1.787	1.839	1.890	1.938
Joint-threshold	1.831	1.901	1.964	2.024
Deletion	1.995	2.133	2.257	2.365
p-merger	2.833	2.771	2.741	2.669
Mean set size, misled at: query only				
Honest-mean oracle	1.718	1.718	1.718	1.718
All-node mean	1.806	1.806	1.806	1.806
Fixed-set	1.986	2.053	2.119	2.186
Joint-threshold	2.022	2.103	2.181	2.263
Deletion	2.203	2.364	2.495	2.625
p-merger	2.785	2.728	2.700	2.624

Table 21: Fixed-set never loses clean-score integrity. Pruning the two least-trusted judges (V1) returns sets 0.034 to 0.054 candidates per query smaller across declared budgets when only the calibration channel is misled but excludes an answer the honest-mean oracle admits on up to 308 of 464 budget-respecting replicates; the audited budget (V2) keeps every judge in the reference and loses integrity on no budget-respecting replicate here. "Lost integrity" counts budget-respecting replicates on which a rule excluded at least one query–candidate pair the honest-mean oracle admits, so fixed-set's zeros confirm part 1 of Theorem 4.3. The notes after the table define both comparators and qualify their rows. Means over 500 replicates; $n = 333$.

Misled at	Rule	Coverage	Cov. ($a \le A$)	Size	Nodes in ref.	Lost integrity ($a \le A$)
By regime, declared $A = 2$						
Both phases	Fixed-set	0.9247	0.9243	1.869	16	0 / 464
	V1, trust pruning	0.9041	0.9039	1.736	14	308 / 464
	V2, audited budget	0.9316	0.9303	1.940	16	0 / 464
Calibration only	Fixed-set	0.9105	0.9115	1.787	16	0 / 464
	V1, trust pruning	0.9030	0.9039	1.733	14	308 / 464
	V2, audited budget	0.9172	0.9172	1.840	16	0 / 464
Query only	Fixed-set	0.9349	0.9339	1.986	16	0 / 464
	V1, trust pruning	0.9052	0.9039	1.745	14	308 / 464
	V2, audited budget	0.9408	0.9396	2.055	16	0 / 464
By declared budget, misled at calibration only						
$A = 2$	Fixed-set	0.9105	0.9115	1.787	16	0 / 464
	V1, trust pruning	0.9030	0.9039	1.733	14	308 / 464
	V2, audited budget	0.9172	0.9172	1.840	16	0 / 464
$A = 3$	Fixed-set	0.9172	0.9173	1.839	16	0 / 498
	V1, trust pruning	0.9116	0.9117	1.797	14	242 / 498
	V2, audited budget	0.9172	0.9173	1.840	16	0 / 498
$A = 4$	Fixed-set	0.9237	0.9237	1.890	16	0 / 500
	V1, trust pruning	0.9186	0.9186	1.852	14	163 / 500
	V2, audited budget	0.9172	0.9172	1.840	16	0 / 500
$A = 5$	Fixed-set	0.9301	0.9301	1.938	16	0 / 500
	V1, trust pruning	0.9256	0.9256	1.904	14	80 / 500
	V2, audited budget	0.9172	0.9172	1.840	16	0 / 500

What the audit classified, per replicate (common to all three regimes)

A'	Replicates	Share	Flagged	Misled missed	Faithful flagged
2	82	0.164	0.000	0.000	0.000
3	364	0.728	1.387	0.000	0.022
4	38	0.076	3.000	0.000	0.237
5	14	0.028	4.286	0.000	2.214
7	2	0.004	7.000	0.000	4.500

Notes to Table 21. Both comparators read an audit pool of 100 questions disjoint from calibration and evaluation. V1 prunes the two judges with the highest audit error and runs the fixed-set rule on the rest at budget $\max(0, A - 2)$; V2 flags judges whose audit error exceeds 0.4 and runs the fixed-set rule on all sixteen at the audited budget A', the smallest budget whose tail under the corner law (flagged judges at $\bar{p}_2 = 0.5$, the rest at $\bar{p}_1 = 0.02$) is at most $\delta = 0.01$. V2 ignores A, so its row repeats down the by-budget panel. Broken out by realized misled count at $A = 2$, V1 loses integrity on 86 of the 86 replicates at a realized count of 0, on 222 of the 230 at 1, and on 0 of the 148 at 2: pruning two judges preserves the honest mean only when exactly two were misled and the audit ranked them first, and the losses sit where it discards honest judges instead. The upper panel holds $A = 2$ and varies the regime: on budget-respecting replicates the calibration-only and query-only rows differ by 2.2 points for both fixed-set and V2. V1's rows coincide on all 464 of them, an invariance of the construction rather than robustness: the audit is not split by regime, and there the two pruned judges include every misled one. Its marginal coverage moves by 0.2 points only through the 36 over-budget replicates. The audited budget exceeds the declared $A = 2$ on 418 of the 500 replicates (modal $A' = 3$, on 364), so V2 is a restriction of the declared family only at larger declared budgets; the bottom panel is common to all three regimes. V2's advertised tail assumes the flagged judges bound how many carry a failure probability above $\bar{p}_1$ (Lemma N.1); the flagged count is instead a realized one, and it falls below the law's 2 suspects on 305 of the 500 replicates. Where no judge is flagged (82 replicates) it returns $A' = 2$ and advertises a tail of 0.0037, against 0.0777 under the law. The audit missed no misled judge but flagged 57 faithful ones in total, so $A' \ge a$ held on all 500 replicates while A' is not a function of the realized count; the tail Proposition R.1 spends, $\mathbb{P}\{a > A'\}$, is bounded through the smallest audited budget, $A' \ge 2$, by the law's 0.0777. Fixed-set is decided in exact rational arithmetic and both comparators in floating point; the two ways of deciding fixed-set disagree on 2,496 of 13,608,000 decisions.