DISAGREEMENT HAS A DEPTH: FAIL-SAFE CONSENSUS OVER LABEL HIERARCHIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Coordinated failures can flip a majority vote, yet sources that disagree on a label often agree on its ancestors: disagreement has a depth. We ask for the most specific statement that stays consistent with a strict majority of the uncorrupted sources, right or wrong, when up to f of M sources are corrupted. For label reports, the finest such rule, even among set-valued rules and rules that see the whole stream of examples, returns the deepest subtree holding more than $(M + f)/2$ reports. We derive its exact loss of depth, with a barrier at $f = M/3$, and its exact risk under per-example and persistent adversaries, certifiable from i.i.d. labelled held-out data. For probability reports, backing off from the likeliest label to its first ancestor with mean probability at least θ is safe exactly when $\theta > \frac{1}{2} + \frac{f}{2M}$. Calibration to 95% coverage crosses this threshold on all four datasets, human labels included. Safe backing off is never finer than the finest safe subtree-valued rule for probability reports, so backing off matched to that rule's clean error crosses the threshold on all three machine-source datasets, with 1.2 to 2.7 times the rule's per-example worst-case error for two or three of ten corrupted sources. Safety costs abstentions and buys little against the non-adaptive failures we tested.

1 INTRODUCTION

Many predictions are aggregated from several sources that each output a label from a taxonomy: members of a model ensemble (Lakshminarayanan et al., 2017), modalities of a multi-omic assay (Hao et al., 2021), human annotators (Peterson et al., 2019), or cooperating language-model agents (Du et al., 2024). Sources fail: a model can be trained on poisoned data (Levine & Feizi, 2021), an antibody panel can break, an annotator can be careless or adversarial (Jagabathula et al., 2017). Majority voting and probability averaging tolerate rare independent failures, but a few coordinated ones can flip the answer unnoticed, and abstaining whenever the vote is indecisive discards the whole prediction.

Work on label hierarchies has mostly changed how a single model is trained, decoded or calibrated (Bertinetto et al., 2020; Karthik et al., 2021; Valmadre, 2022; Goren et al., 2024). We ask what a hierarchy changes when the predictions of several sources, some of which may fail, are combined. Figure 1a shows the phenomenon on one blood cell: six of ten sources name the same subtype, seven the same cell type and nine the same lineage, so a few corrupted sources can destroy fine agreement while coarse agreement survives. *Disagreement has a depth.* Given the observed reports, what is the most specific statement that remains valid under every admissible corruption of up to f sources, and what does it cost in specificity and in error?

We call a rule *f-safe* if its output never contradicts, and never claims more than, what a strict majority of the uncorrupted sources supports. Safety is relative to that majority, not to the truth: when most working sources share an error, a safe rule may share it, and we measure how often. The answer is one rule, Q_f: the deepest subtree holding more than $(M + f)/2$ of the M reports. Majority-rule consensus of trees (Margush & McMorris, 1981) aggregates hierarchical votes, and Byzantine quorum systems (Malkhi & Reiter, 1998) size quorums so that correct servers overlap; neither asks, for an observed profile, which statement stays valid under every admissible corruption. Answering it yields their quota, applied to every subtree at once so that the answer loses depth instead of vanishing, and what neither gives: for label reports, a finest safe output even among set-valued and stateful rules, its

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.

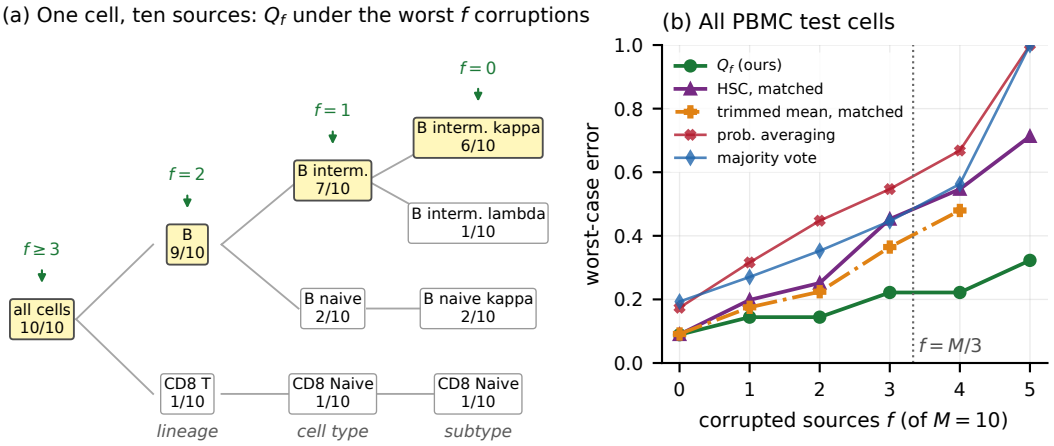

Figure 1: Corruption removes specificity before it forces contradiction. (a) A test cell of the PBMC blood atlas and the labels reported by ten sources, with counts per node (true path shaded). Under the worst corruption of f sources, the finest safe rule Q_f names the subtype for $f = 0$, the cell type for $f = 1$, the lineage for $f = 2$, and abstains from $f = 3$ on. (b) Worst-case error over the 20,000 test cells when f of the ten sources are corrupted; HSC and a trimmed-mean variant are re-calibrated at each f to the clean error of Q_f (for them and probability averaging, attack lower bounds).

exact loss of depth, its exact error against the truth with a held-out certificate, and the threshold at which backing off on averaged probabilities stops being safe.

The characterization makes a sharp prediction about a rule in current use. Hierarchical selective classification (HSC) (Goren et al., 2024), run on the averaged probabilities, climbs from the most likely label to its first ancestor with mass at least θ; it is f-safe on all probability reports if and only if $\theta > \theta_c = \frac{1}{2} + \frac{f}{2M}$, a boundary that clean calibration need not respect. Our contributions:

- **Optimal safe information** (Theorem 1, Proposition 7). For label reports, Q_f is f-safe and every f-safe rule contains $Q_f(s)$ on every profile s, even a rule that outputs an arbitrary set of labels, knows which source sent which report, or sees every example of a persistently corrupted stream.
- **The exact price of safety** (Theorems 2 and 3, Corollary 2). Relative to the honest majority, corruption only removes depth, by an amount the honest reports fix, and unanimous honest answers survive if and only if $M > 3f$. The worst-case error is exact for per-example and persistent adversaries and certifiable from i.i.d. labelled held-out data, and a label-free part of it can warn that deployment data differ. For rules that output one subtree, the analysis extends to probability reports, stricter anchors and weighted sources.
- **The boundary, tested** (Proposition 1, Figure 2, Tables 1 to 3). Calibration to 95% coverage crosses θ_c on CIFAR-100, FGVC-Aircraft, a CITE-seq atlas and CIFAR-10H human annotations. Safe climbing is never finer than the finest safe subtree-valued probability rule, so climbing matched to that rule's clean error crosses θ_c on all three machine-source datasets, with 1.2 to 2.7 times the rule's per-example worst-case error for two or three of ten corrupted sources, while climbing on vote shares, which nearly coincides with Q_f, stays safe. Safety costs abstentions and gains little under the non-adaptive failures we tested; a learned reliability model can be steered by an adaptive adversary; and the certificate fails where its i.i.d. assumption does (Section 5).

2 Related Work

Hierarchical prediction, abstention and conformal prediction. Hierarchies make mistakes less severe (Bertinetto et al., 2020; Karthik et al., 2021), support evaluation at several operating points (Valmadre, 2022), and allow backing off to an ancestor when a model is unsure (Deng et al., 2012). HSC climbs the tree until the predicted mass of a node passes a threshold (Goren et al., 2024), and conformal methods return internal nodes or node sets with coverage guarantees (Mortier et al., 2026; Zhang et al., 2025; den Hengst et al., 2025); robust conformal prediction protects coverage against bounded input perturbations (Gendler et al., 2022) or malicious federated clients (Kang

et al., 2024). These methods assume intact predictors at test time; we derive the threshold that climbing needs when the predictors themselves may be corrupted, and show that split-conformal calibration of climbing to common coverage targets crosses it. Ensemble disagreement signals error (Lakshminarayanan et al., 2017; Jiang et al., 2022) and coarse predictions can repair fine ones (Jain et al., 2023). Ultrametric and p-adic (Albeverio et al., 1999; Murtagh, 2004; Martins, 2025) and hyperbolic (Nickel & Kiela, 2017; Ganea et al., 2018) representations embed hierarchies; we need none.

Certified robustness by voting. Partition-and-vote defences certify that a prediction is stable under poisoned training points through the vote margin (Levine & Feizi, 2021; Wang et al., 2022; Rezaei et al., 2023; Jia et al., 2021; Cao et al., 2021), and smoothing certifies stability under input perturbations (Cohen et al., 2019; Levine & Feizi, 2020; Wang et al., 2024); adaptive hierarchical certification coarsens a single smoothed segmentation model against input perturbations (Anani et al., 2024), whereas we coarsen an aggregate against corrupted sources. Stability certificates preserve an observed decision, which anchors it to the clean output; safety preserves compatibility with the honest majority, which is finer and keeps unanimous answers for $M > 3f$ (Theorem 2) rather than, up to ties, $M > 4f$ (Proposition 5).

Robust aggregation, crowdsourcing and consensus. Byzantine-robust aggregation protects gradient or model averaging (Blanchard et al., 2017; Yin et al., 2018; Karimireddy et al., 2022; Parsa et al., 2026). Robust federated inference (Dhasade et al., 2026) aggregates the probability vectors of several models, some Byzantine, by robust averaging or a trained aggregator and bounds the loss of flat accuracy; we coarsen instead of committing, and evaluate its robust averaging followed by climbing, not its trained aggregator. Selective ensembles abstain when the ensemble mode fails a test, for consistency rather than robustness (Black et al., 2022). Crowdsourcing methods estimate annotator reliability (Dawid & Skene, 1979), also over a taxonomy (Otani et al., 2015; Van Horn et al., 2018), or identify adversarial annotators (Jagabathula et al., 2017; Steinhardt et al., 2016), and learning from untrusted data returns a short list of candidates (Charikar et al., 2017). Majority-rule consensus of trees is classical (Margush & McMorris, 1981; Barthélemy & Janowitz, 1991): Q_f is the ℓ-majority consensus of the reported paths at $\ell = \frac{1}{2} + \frac{f}{2M}$. What is new is that, for label reports, this quota gives the finest f-safe output, even among set-valued and stateful rules, with its exact risk and a held-out certificate (Table 5).

Byzantine quorums and consensus. Agreement without message authentication needs more than three times as many processes as faults (Lamport et al., 1982), and validity conditions specify which decisions faulty processes may not force (Neiger, 1994; Civit et al., 2023). Byzantine quorum systems (Malkhi & Reiter, 1998; Cachin et al., 2011) size quorums so that any two overlap in enough correct servers: dissemination quorums of more than $(M + f)/2$ servers exist if and only if $M > 3f$, and masking quorums need $M > 4f$. Our quotas and barriers coincide with these. Vector consensus (Vaidya & Garg, 2013; Mendes & Herlihy, 2013) and approximate agreement on trees (Nowak & Rybicki, 2019) use rounds of communication; our sources do not communicate.

3 SETTING

Label hierarchies as nested balls. A label hierarchy is a rooted tree whose leaves are the labels $\mathcal{Y}$. We identify each node with the set of labels below it, a *ball*; $B_k(y)$ is the depth-k ball containing y, so $B_0(y) = \mathcal{Y}$ and $B_L(y) = \{y\}$ at the leaf depth L. Our results rest on one structural fact:

Lemma 1 (Nested or disjoint)**.** *Two balls are either nested or disjoint.*

A rule outputs a ball: $B_k(y)$ asserts the first k decisions of y and abstains below them (notation: Table 4).

Sources, corruption and safety. Standing assumptions. **(A1)** The tree has all leaves at depth L (shallower leaves are padded with single-child nodes) and at least two children at the root. **(A2)** Each of M sources reports a label; $s = (s_1, \ldots, s_M)$ is a *profile* and $n_s(B)$ counts its reports inside ball B (Section 4: probability vectors). **(A3)** An unknown set Z of at most $f < M$ sources is *corrupted*: their reports are replaced by arbitrary labels chosen by an adversary who knows everything, including

the true label; a rule sees the observed profile s' and which source sent each report, but not Z. Let $h = s'|_{[M]\setminus Z}$ be the *honest reports*. Two disjoint balls cannot both hold more than half of h, so by Lemma 1 the balls that do form a chain from the root; its deepest element $C(h)$, the *honest-majority ball*, is the most specific statement a strict majority of the working sources supports.

Definition 1 (f-safety)**.** A rule A is *f-safe* if $A(s') \supseteq C(s'|_{[M]\setminus Z})$ for every profile s' and every set Z of at most f sources.

By Lemma 1, $A(s') \supseteq C(h)$ means the output neither contradicts nor claims more than the honest majority, for every placement of the corrupted sources. The honest majority can be wrong, so safety is not correctness; Theorem 3 prices the difference.

Threat regimes and conventions. (A) *Per-example oracle*: Z and the corrupted reports are chosen for each example knowing its label, the pointwise model of certified defences. (B) *Persistent*: the same Z on every example, as with a poisoned model or a compromised annotator. (C) *Label-free*: a persistent set fails without knowing the truth (collusion, poisoning, noise, broken sensors). Each regime is a special case of the one before, so worst-case risks cannot increase from (A) to (C). Quotas are strict: more than $(M + f)/2$ reports means at least $\tau_f = \lfloor (M + f)/2 \rfloor + 1$; for even M, $\tau_{2j} = \tau_{2j+1}$, so with $M = 10$ the rules for $f = 0, 1$ coincide, as do those for $f = 2, 3$. At most one ball per depth holds more than $M/2$ reports, so ties never affect the quorum rules.

4 The finest safe consensus and its price

We characterize the safe rules, price safety in depth and in error, and extend the picture to other anchors and to probability reports. Proofs are in Appendix B; every closed form is verified computationally, exhaustively on small instances and by constructive and random attacks on larger ones (Appendix C).

Theorem 1 (The finest safe rule)**.** *Under (A1) to (A3), for $0 \le f < M$ let $Q_f(s)$ be the deepest ball B with $n_s(B) > (M + f)/2$. Then (i) Q_f is f-safe; (ii) for every profile s some set Z of at most f sources has $C(s|_{[M]\setminus Z}) = Q_f(s)$; (iii) hence every f-safe rule A, whether it outputs a ball or any set of labels and whether or not it uses source identities, satisfies $A(s) \supseteq Q_f(s)$ on every profile s: Q_f is the finest f-safe rule under set inclusion.*

Proof sketch. (i) A ball with more than $(M + f)/2$ observed reports keeps more than $(M - |Z|)/2$ honest ones. (ii) If c is the count of the most supported child of $Q_f(s)$, declaring $z = \max(0, 2c - M) \le f$ of its reports corrupted leaves no child with an honest majority, while $Q_f(s)$ keeps one. □

The same holds for rules that see the whole stream of examples: which sets are admissible does not depend on the reports, so the set of (ii) is an admissible persistent corruption, and statefulness cannot buy specificity (Proposition 7); a rule that estimates source reliability gives up the guarantee, and whether it lowers the error depends on how the honest sources behave (Section 5.3). Specificity is not accuracy: coarser safe rules can err less.

Theorem 2 (The price in specificity)**.** *Under (A1) to (A3), let Z be the corrupted set, h the honest reports and s' any observed profile.*

(i) Corruption only coarsens, relative to the honest majority *(Theorem 1(i) restated): $Q_f(s') \supseteq C(h)$.*

(ii) Exact resolution: $\operatorname{depth} Q_f(s') \ge \rho_f(h)$*, the depth of the deepest ball with $n_h(B) > (M+f)/2$ ($\rho_f(h) = 0$ if there is none), and the adversary attains it by reporting labels in the top-level ball with the fewest honest reports.*

(iii) One-third barrier: *if $M > 3f$ and all honest reports lie in a ball B, then $C(h) \subseteq Q_f(s') \subseteq B$. If $M \le 3f$, some corruption of every profile drives Q_f, and hence every f-safe rule, to the root, even when all honest sources report the same label.*

Part (iii) is the availability condition of dissemination quorums (Malkhi & Reiter, 1998); relative to the clean output, corruption can also make Q_f finer and wrong (Theorem 3). Minimizing (ii) over Z gives:

Corollary 1 (Worst case over corruptions of a clean profile)**.** *For a clean profile s with label y, the least depth of Q_f over all corruptions of at most f sources is the largest k such that some depth-k*

ball holds at least $\tau_f + f$ clean reports (0 if none), and, for $k \geq 1$, $Q_f(s') \subseteq B_k(y)$ under every such corruption if and only if $n_s(B_k(y)) \geq \tau_f + f$.

Theorem 3 (The price in error)**.** *Under (A1) to (A3), let s be the clean profile of an example with label y and $E_f(s, y) = \mathbf{1}[\max_{u \not\ni y} n_s(u) > (M - f)/2]$, the maximum over balls u that exclude y.*

(i) Regime A: *some corruption of at most f sources makes Q_f exclude y if and only if $E_f(s, y) = 1$, which also equals $\max_{|Z| \leq f} \mathbf{1}[y \notin C(s|_{[M]\setminus Z})]$; the output can leave the path of y at depth j exactly when a sibling of $B_j(y)$ holds more than $(M - f)/2$ clean reports. So $R_f^\star = \mathbb{E}[E_f(s, y)]$ is the exact, attained per-example oracle risk.*

(ii) Regime B: *if the same Z with $|Z| = f$ is corrupted on every example, the worst-case risk is exactly $R_Z = \Pr[\max_{u \not\ni y} n_{s|_{[M]\setminus Z}}(u) > (M - f)/2] \leq R_f^\star$.*

(iii) A label-free part: *call s* contested *($K_f(s) = 1$) if two disjoint balls each hold more than $(M - f)/2$ clean reports, and otherwise let $U_f(s)$ be the deepest ball that does. Then $E_f(s, y) = K_f(s) + (1 - K_f(s))\, \mathbf{1}[y \notin U_f(s)]$.*

Corollary 2 (Held-out certificate and a label-free warning)**.** *Let N labelled held-out examples with uncorrupted reports be i.i.d. draws from the deployment distribution, let e of them have $E_f = 1$, and for $|Z| = f$ let e_Z of them have $E_{f,Z} = 1$, where $E_{f,Z} = \mathbf{1}[\max_{u \not\ni y} n_{s|_{[M]\setminus Z}}(u) > (M - f)/2]$ is the indicator of Theorem 3(ii). With probability at least $1 - \delta$, $R_f^\star \leq \mathrm{Beta}^{-1}(1 - \delta;\, e + 1,\, N - e)$, the one-sided Clopper-Pearson bound (Clopper & Pearson, 1934) ($\mathrm{Beta}^{-1}(q; a, b)$ is the q-quantile of $\mathrm{Beta}(a, b)$, and the bound is 1 if $e = N$), and with probability at least $1 - \delta$ the bounds $R_Z \leq \mathrm{Beta}^{-1}(1 - \delta/\binom{M}{f};\, e_Z + 1,\, N - e_Z)$ hold for all Z at once. The contested share $\mathbb{E}[K_f] \leq R_f^\star$ needs no labels; a change in it between held-out and uncorrupted deployment reports warns that their distributions differ, though not why.*

Which majority to protect. Anchoring to the truth allows only the root, since all sources may err. The honest majority gives the most specific finest rule; stricter anchors need larger quotas and tolerate fewer faults. Containing the α-supermajority ball of the honest reports needs more than $\alpha M + (1 - \alpha)f$ reports and keeps unanimous answers if and only if $(1 - \alpha)M > (2 - \alpha)f$. Containing the clean majority ball for every clean profile within f replacements needs more than $M/2 + f$ reports, the rule Q_{2f}; for $2f < M$ its worst-case error equals the clean error of the majority ball, and it keeps unanimous answers only if $M > 4f$, as do masking quorums and, up to ties, stability certificates for voting (Levine & Feizi, 2021; Jia et al., 2021). With weights and a weight budget, the ball-valued characterization holds with a per-ball quota, but its set-valued part can fail (App. B).

Probability reports. Now each source reports a probability vector and a corrupted one any vector in the simplex. Let $\mu_m(B)$ be the mass source m puts on B, $n_s(B) = \sum_m \mu_m(B)$, $\mathrm{top}_j(B)$ the sum of the j largest $\mu_m(B)$, $S_{M-f}(u)$ the sum of the $M - f$ largest $\mu_m(u)$, and $C(h)$ the deepest ball holding more than half of the honest mass.

Proposition 1 (Probability reports)**.** *(i) The balls with $n_s(B) - \mathrm{top}_j(B) > (M - j)/2$ for $j = 0, \ldots, f$ form a chain; its deepest element $Q_f^{\mathrm{p}}(s)$ is f-safe, a ball-valued rule is f-safe if and only if it contains $Q_f^{\mathrm{p}}(s)$ on every profile, and $Q_f^{\mathrm{p}} = Q_f$ for point masses. (ii) Some corruption of at most f sources makes Q_f^{p} exclude y if and only if some ball $u \not\ni y$ has $S_{M-f}(u) > (M - f)/2$. (iii) Climbing, which starts from the most likely label of the mean probability and outputs its deepest ancestor with mean mass at least θ, is f-safe if and only if $\theta > \theta_c = \frac{1}{2} + \frac{f}{2M}$. (iv) Hence every f-safe ball-valued rule, in particular climbing with $\theta > \theta_c$, contains Q_f^{p} on every observed profile, so on every example it is at most as deep, and its worst-case error at most as high, as that of Q_f^{p}.*

Part (iii) concerns probability vectors (for one-hot reports the threshold is $\lfloor (M + f)/2 \rfloor / M$, App. E); calibration, which picks θ for a target clean error, is not designed to respect it, and what crossing it costs is measured in Section 5.2. Unlike Theorem 1, Proposition 1 concerns ball-valued rules: with probability reports a set of labels can be safe and strictly finer than Q_f^{p} (App. B). By (i), any aggregator followed by climbing is unsafe on some profile or never finer than Q_f^{p}, and a climbing output that does not contain $Q_f^{\mathrm{p}}(s)$ is unsafe: some placement of the corrupted sources makes it contradict or

exceed the honest majority. We call it an *overclaim*; above θ_c there are none. Proposition 3 (App. B) gives certified upper bounds on the worst-case error of climbing on the mean, trimmed mean or median; for the mean at $\theta = \frac{1}{2} + \frac{f'}{2M}$ the bound equals, up to ties, the exact risk of $Q^{\mathrm{p}}_{f'}$ designed for $f' \le f$ faults and attacked by f. Computing Q_f takes $O(ML)$ time, 4 to 11 μs per example (App. D).

5 Experiments

Four questions test Section 4. **Q1** Does hierarchical disagreement leave usable safe information (Theorem 2)? **Q2** Does θ_c separate safe from unsafe climbing, and what does crossing it cost at a fixed clean error (Proposition 1)? **Q3** What changes under persistent and label-free failures, and against rules that learn source reliability (Proposition 7)? **Q4** When does the held-out certificate transfer (Corollary 2)?

Data and sources. *CIFAR-100* (Krizhevsky, 2009): 20 superclasses $\to$ 100 classes, 10,000 test images. *FGVC-Aircraft* (Maji et al., 2013): manufacturer $\to$ family $\to$ variant, 3,333 test images. *PBMC CITE-seq* (Hao et al., 2021): lineage $\to$ type $\to$ subtype, 20,000 test cells from two donors unseen in training. Each has $M = 10$ temperature-calibrated sources, so the budgets $f = 2, 3$ of Tables 1 to 3 satisfy $M > 3f$: linear probes on the frozen features of ten ImageNet backbones, or, for PBMC, small networks that each see one of ten disjoint gene or antibody panels. *CIFAR-10H* (Peterson et al., 2019): CIFAR-10 test images with about 50 human labels each under a three-level hierarchy; for each image we sample M of its human reports (five draws), so its sources are per-image reports, not persistent annotators.

Rules, attacks and metrics. Quorum rules: Q_f, Q^{p}_f and the clean anchor Q_{2f}. Baselines: HSC (Goren et al., 2024); the same climbing with split-conformal thresholds (Papadopoulos et al., 2002; Lei et al., 2018), on vote shares, or on the coordinate-wise trimmed mean or median or the geometric median, the robust averaging rules of Byzantine-robust aggregation (Yin et al., 2018; Dhasade et al., 2026); one-coin Dawid-Skene (Dawid & Skene, 1979) followed by climbing; probability averaging; majority and top-down votes; and the flat rule with the same guarantee (a label if it holds more than $(M + f)/2$ reports, else abstain). Thresholds are set on clean validation data; a rule *matched* to Q has the clean validation error of Q. An *error* is an output that excludes the true label, a *top-level error* one that names the wrong lineage, superclass or manufacturer; *depth* is the mean output depth over L (other specificity measures agree, App. E). Worst cases are *exact* for quorum rules; for climbing we report an *attack lower bound* and, where available, the *certified upper bound* of Proposition 3 (proof in App. B, attacks in App. D).

5.1 Q1: Safe information survives corruption

Sources are more reliable near the root (Table 24): the majority ball Q_0 stops above the leaves for 21%, 61% and 24% of the machine-source examples, where probability averaging errs on 52%, 64% and 45% of them (9.7%, 12.4% and 8.6% elsewhere). Under the worst corruption of two of ten sources, Q_2 still answers at least at the top level for 67% of the CIFAR-100 images and 79% of the cells, but for only 24% of the aircraft, whose weak sources are wrong at the top level 42% of the time. The thesis test swaps only the tree: with a randomly permuted hierarchy these shares fall to 53%, 16% and 52%, close to the 52%, 14% and 49% that get any answer without a hierarchy, so the gain over the flat guarantee comes from the hierarchy's semantics. The price is availability: some corruption drives Q_2 to the root on 32.8%, 75.9% and 21.0% of the examples, a persistent pair on up to 30.3%, 74.1% and 19.9%, and matched HSC on at least 31.5%, 83.6% and 15.5% (Tables 1 and 14).

Real sources err together (intra-class correlation 0.30 to 0.48 for machines, 0.09 to 0.19 for humans), so the share certified correct at level k, $\Pr[X_k \ge \tau_f + f]$ with X_k the number of clean reports in $B_k(y)$ (Corollary 1), is matched within 3.2 points by a beta-binomial fitted with this one parameter, where independent sources would be off by up to 41; ten probes of one backbone raise the worst case of Q_2 from 14.6% to 25.0% (HSC: 34.5%). Top-down stability, the certificate of partition-and-vote defences, certifies more decisions (66% against 50% of CIFAR-100 leaves at $f = 2$) but never abstains (worst case 34%, 69% and 35% against 15%, 19% and 14% for Q_2; Tables 10 and 25).

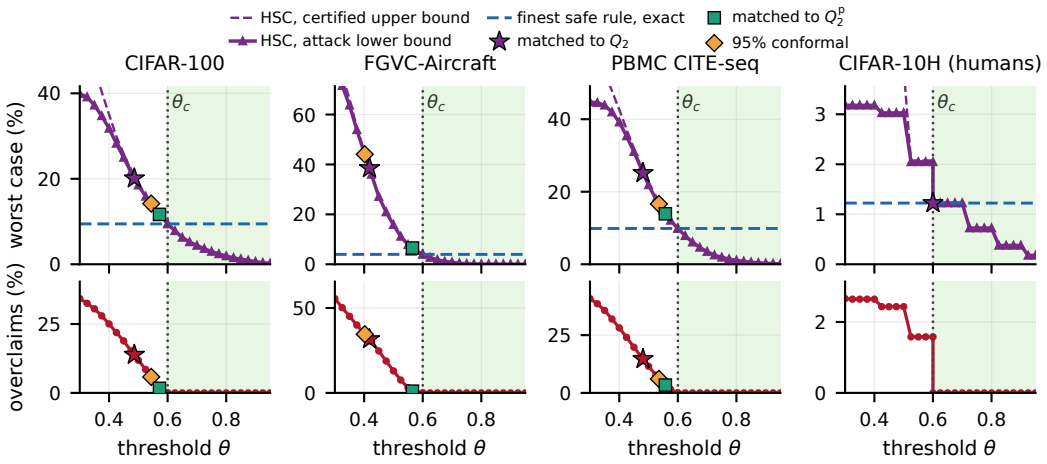

Figure 2: Crossing the safety boundary, $f = 2$ of $M = 10$ sources. *Top:* worst-case error of climbing on the mean (HSC) against its threshold, with the exact worst case of the finest safe ball-valued rule for the same reports (Q_2^{p}; Q_2 for one-hot human reports). *Bottom:* share of test examples on which climbing overclaims. Proposition 1 predicts no overclaim right of $\theta_c = 0.6$ (shaded). Stars: thresholds matched to Q_2 (CIFAR-10H: mean of five draws); squares: matched to Q_2^{p}; diamonds: split-conformal thresholds at 95% coverage (CIFAR-10H: $\theta = 0$, off the plot).

Table 1: Rules under $f = 2$ of 10 corrupted sources; each block compares rules on the same reports, and a matched row has the clean validation error of the rule it is matched to. Errors in %; depth: mean output depth over L; "worst": worst-case error; "none": share that some corruption drives to the root. Quorum rules are exact; for climbing, "worst" and "none" are attack lower bounds (certified bounds: equal for HSC matched to Q_f^{p}, at most 0.5 points higher for HSC matched to Q_f, up to 2.5 times for the trimmed mean; Table 11). More: Tables 6 and 9.

	CIFAR-100				FGVC-Aircraft				PBMC CITE-seq			
Rule	clean	depth	worst	none	clean	depth	worst	none	clean	depth	worst	none
probability reports; HSC matched to Q_f^{p}												
Q_f^{p} (ours)	3.3	0.68	9.4	41.4	0.7	0.17	4.0	94.2	2.2	0.77	9.9	20.7
HSC, matched	3.4	0.69	11.7	39.9	0.6	0.17	6.5	93.5	2.5	0.78	13.9	19.6
label reports; climbing on vote shares matched to Q_f												
Q_f (ours)	5.8	0.76	14.6	32.8	4.5	0.39	18.6	75.9	5.1	0.81	14.4	21.0
vote shares, matched	5.8	0.76	14.6	32.9	4.4	0.39	18.6	77.3	5.1	0.81	14.4	21.7
Q_{2f}, clean anchor	3.5	0.69	9.5	45.3	2.0	0.29	9.5	88.5	2.6	0.74	8.9	38.0
flat guarantee	4.7	0.70	12.2	48.1	2.3	0.30	9.5	85.6	3.9	0.67	10.9	51.2
averaged probabilities, matched to Q_f												
HSC	5.6	0.76	20.1	31.5	4.6	0.33	38.5	83.6	5.4	0.83	25.2	15.5
trimmed mean + climb	5.8	0.78	18.5	20.3	4.4	0.36	24.8	61.7	6.2	0.85	22.4	5.4

5.2 Q2: The safety boundary, crossed

Figure 2 shows Proposition 1 at work: above θ_c climbing never overclaims, so safe climbing has a clean error of at most 2.7%, 0.3% and 1.6% at $f = 2$ with machine sources and 0.4% with human reports. Common targets need more: 95% split-conformal coverage sets $\theta = 0.54$, 0.40 and 0.54, below θ_c even for $f = 1$ (worst case 14.2%, 44.1% and 16.6% at $f = 2$), and $\theta = 0$ on CIFAR-10H.

What crossing θ_c costs. Fixing the report model isolates the threshold. Safe climbing is never finer than the finest safe ball-valued probability rule Q_f^{p} (Proposition 1(iv)), so HSC matched to it crosses θ_c on every machine-source dataset ($\theta = 0.56$ to 0.61 for $f = 2, 3$), and at nearly the same test clean error and depth its worst case is 1.23, 1.64 and 1.41 times that of Q_2^{p} (1.34, 2.72 and 1.84 times at $f = 3$); its certified bound equals the attack, so these ratios are exact; paired bootstrap intervals

Table 2: Failures that need no labels (regime C), $f = 3$ of 10, mean over all $\binom{10}{3}$ failed sets: error (%) / depth; poisoning is scored on the poisoned group. HSC and the trimmed mean are matched to Q_3.

Data	Failure	Q_f	Q_f^{p}	HSC	trimmed	prob. avg.
CIFAR-100	collusion, runner-up	3.7 / 0.57	1.2 / 0.39	13.2 / 0.70	11.8 / 0.76	35.6 / 1.00
	collusion, bandwagon	11.2 / 0.87	5.7 / 0.75	15.5 / 0.94	16.8 / 0.97	19.2 / 1.00
	coordinated poisoning	6.2 / 0.54	1.2 / 0.30	3.8 / 0.55	8.8 / 0.69	32.0 / 1.00
FGVC-Aircraft	collusion, runner-up	5.3 / 0.25	1.1 / 0.09	26.9 / 0.42	18.6 / 0.46	73.2 / 1.00
	collusion, bandwagon	13.2 / 0.56	1.9 / 0.24	28.8 / 0.75	27.2 / 0.75	45.3 / 1.00
	coordinated poisoning	4.1 / 0.07	0.0 / 0.00	4.4 / 0.05	5.4 / 0.11	48.9 / 1.00
PBMC CITE-seq	collusion, runner-up	3.8 / 0.66	1.8 / 0.58	18.0 / 0.83	15.5 / 0.87	37.3 / 1.00
	collusion, bandwagon	10.9 / 0.90	5.1 / 0.82	16.5 / 0.97	17.3 / 0.99	18.2 / 1.00
	coordinated poisoning	0.0 / 0.94	0.0 / 0.91	0.1 / 0.98	0.7 / 1.00	0.8 / 1.00

exclude zero, and the ratios stay between 1.19 and 4.30 over five retrained pools per image dataset (App. E). With vote reports the same protocol stays safe: vote-share climbing matched to Q_f lands just above the one-hot threshold 0.6 and nearly coincides with Q_f.

Across report models. Matched to the vote rule Q_f, HSC on averaged probabilities has 1.38, 2.07 and 1.74 times its worst case (1.68, 2.28 and 2.04 at $f = 3$; Figure 1b shows every f on PBMC), within 0.5 points of its certified bound. This gap mixes the threshold with the report model. Robust averaging narrows it by attack, to 1.2 to 1.6 times for the trimmed mean and median and 0.97 to 1.33 times for the geometric median, but their certified bounds are up to 2.5 times their attack values or unavailable; over whole parameter ranges, vote quorums have the lowest worst case where calibration lands and the probability quorum the lowest certified one at intermediate depths (Figure 3, App. E). Paired bootstrap intervals exclude zero for every difference with HSC and the trimmed mean.

5.3 Q3: Persistent and label-free failures

Regime B. When the same pair is corrupted on every example, the exact worst case of Q_2 over all pairs is 9.9%, 11.1% and 9.6% (regime A: 14.6%, 18.6% and 14.4%), and HSC keeps 1.5, 2.6 and 1.9 times as many errors (Table 3). By Theorem 3(ii) this equals the error of the honest-majority ball of the other eight sources, so knowing the corrupted pair would buy depth, not accuracy. *Learning reliability* (Table 13). Dawid-Skene, re-estimated on the attacked stream, discounts a pair that reports a wrong leaf on every example and beats Q_2 on CIFAR-100 and PBMC (worst over pairs 7.9% and 5.3%). A pair that reports honestly except where attacking flips the output drives it to 11.8% on CIFAR-100, and with the weak aircraft sources either attack does (16.3% and 14.8% against 11.1%); on PBMC it stays below Q_2 at a lower depth (0.70 against 0.81). Learning reliability thus gives up the guarantee (Proposition 7) to beat naive pairs. *Regime C* (Table 2). In *collusion* the failed sources see the honest outputs and back one alternative, the runner-up or, in the *bandwagon* variant, the honest favourite: with three colluders backing the runner-up, Q_3 errs on 3.7%, 5.3% and 3.8%, and HSC, thresholded to the same attacked depth, errs 1.45, 1.95 and 1.26 times as often, while bandwagon collusion hurts Q_3 more (ratios 1.01, 1.19 and 0.99; Table 10). Under poisoning, label noise, sensor noise, blur and panel failures, every calibrated rule stays accurate and none dominates: HSC is more accurate on most CIFAR-100 failures (poisoning: 3.8% against 6.2% at the same depth), Q_f at least as accurate on PBMC and under sensor noise and blur on FGVC-Aircraft (App. E).

5.4 Q4: When the certificate transfers

Estimated on the validation split ($N = 5{,}000$, 3,333, 10,000 and 5,000) with a one-sided 95% Clopper-Pearson bound, the certificate for all errors covers the exact test worst case for every machine-source dataset and $f \leq 4$ and in every CIFAR-10H draw (Tables 3 and 18), and the Bonferroni-corrected bound covers every fixed set. It fails only where its sampling assumption does: on PBMC, whose validation and test cells come from different donors, the top-level certificate misses the test value at every f (1.3% against 3.6% at $f = 2$). The label-free part of Theorem 3(iii) warns: the top-level contested share rises from 0.3% to 0.6% at $f = 2$ and from 1.2% to 3.5% at $f = 3$ ($p \leq 2 \cdot 10^{-4}$), and after a Holm correction over twelve tests only these and PBMC all errors at $f = 3$ fire (Table 19). On

Table 3: Certificate $R_f^\star$ of Q_f on validation data, its 95% bound and the exact test worst case (regime A; bold: missed); regime B: Bonferroni-corrected bound, exact test worst case, HSC attack lower bound. Errors in %. CIFAR-10H: mean of five draws (top-level bound held in four); regime B n/a.

		oracle: all errors			oracle: top level		persistent corruption		
Data	f	valid.	bound	test	bound	test	bound	Q_f test	HSC test
CIFAR-100	2	15.3	16.1	14.6	9.1	8.4	12.4	9.9	14.8
	3	21.7	22.6	21.8	13.2	12.9	16.0	13.6	26.1
FGVC-Aircraft	2	18.0	19.1	18.6	13.6	12.5	13.2	11.1	29.2
	3	32.5	33.8	33.5	23.9	22.2	20.4	18.3	63.8
PBMC CITE-seq	2	16.1	16.8	14.4	1.3	**3.6**	13.8	9.6	17.9
	3	21.6	22.3	22.2	2.0	**5.8**	16.7	14.1	35.8
CIFAR-10H	2	1.4	1.7	1.2	0.2	0.1	n/a	n/a	n/a
	3	2.3	2.7	2.1	0.3	0.2	n/a	n/a	n/a

random splits of pooled data it false-alarms in 3.5% to 8% of tests at level 5%, and detects samples with 20%, 30% and 50% of cells from other donors in 33%, 57% and 99% of draws (Table 20).

6 Discussion and limitations

Operating envelope. Safety pays against failures that target the answer (per-example or adaptive persistent attacks, collusion on the runner-up) when a wrong specific answer costs more than a coarse one: calibrated climbing then has 1.2 to 2.7 times the worst case of the finest safe ball-valued rule at the same clean error, and a learned reliability model can be steered by an adaptive adversary. It does not pay against bandwagon collusion or the non-adaptive failures we tested; correlated sources make the safe answer itself wrong more often (14.6% to 25.0%); and two corruptions, persistent or not, make it abstain on 20% to 76% of machine-source examples, so coarsening is itself an attack surface. For ten models over a taxonomy facing such failures, we would report Q_f (votes) or Q_f^{p} (probabilities) at the budget of the threat model, certify its worst case on held-out data, and never calibrate climbing below θ_c.

Choosing and designing. Set f from the threat model, then choose the anchor and report type by their held-out certificates and depth (Q_{2f} caps the worst case at the clean error of the majority ball); reporting the chain $Q_0(s) \subseteq Q_1(s) \subseteq \cdots$ gives the safe answer for every budget. The two-thirds rule of iNaturalist for community taxa (iNaturalist, 2026) is a deployed supermajority rule, safe against up to a third of unreliable identifiers. A failing group of sources counts as that many failures, so surviving the loss of any one of g equal groups needs $g \geq 4$, more than our two CITE-seq modalities.

Limitations. The results need a tree (or a tree-shaped projection fixed in advance); for probability reports they cover ball-valued rules; a threshold for node-set conformal predictors remains open. Rules modelling honest sources can be more accurate when the model holds, and we test one (Dawid-Skene) only under persistent attacks. Certificates need a trusted labelled run and i.i.d. data; our certification split also chose the probes' weight decay and temperatures (App. D). Uncertainty is conditional on the trained sources (retrained pools in App. E); the warning was tested on one natural shift; and we use $M = 10$ sources, at most three levels and simulated coordinated failures; deeper taxonomies, larger pools and language-model agents remain untested.

7 Conclusion

Disagreement among sources has a depth that decides what can be asserted safely when some fail. For label reports, no f-safe rule, not even a set-valued or stateful one, is more specific than the deepest subtree holding more than $(M + f)/2$ reports; relative to the honest majority, corruption then only removes specificity, and the error that remains is exact and certifiable from i.i.d. held-out data. Backing off on probabilities is safe exactly above a sharp threshold, which 95% coverage calibration crossed on all four datasets; at the clean error of the finest safe ball-valued probability rule, crossing it cost 1.2 to 2.7 times that rule's per-example worst-case error at $f = 2, 3$ on machine sources.

AI USE STATEMENT

Generative AI was used throughout this work. It helped draft, polish and review the text, search for and format references, and propose the title; coding agents processed the datasets and implemented the methods, attacks, experiments and verification tests; AI took part in the discussion of the results, generated the plots, and helped to create slides for internal meetings. Generative AI was not used to generate or annotate data. Every closed form was verified computationally, by exhaustive enumeration of corruptions on small instances and by constructive and random attacks on larger ones (Appendix C), the code was tested before any result was reported, and every reference was checked against its publisher or archive page. The authors reviewed all AI-assisted content and take full responsibility for the paper, its claims and its code.

ETHICS STATEMENT

This work studies how to aggregate predictions so that a bounded number of faulty or adversarial sources cannot make the aggregate contradict the working majority. The single-cell data are a public, de-identified dataset released for research, and the image benchmarks and the human annotations are public research datasets without personal information. The rule we characterize can refuse to commit to fine labels, which in a clinical or safety-critical deployment shifts decisions to a coarser level or to a human; this is intended, but operators must not read a coarse output as a correct fine one. The guarantee is relative to the honest majority of sources: a safe output can still be wrong when most uncorrupted sources share an error, which we state and measure, and its held-out certificate holds only for i.i.d. data from the deployment distribution. Forced coarsening is itself an attack surface, since corrupted sources can deny a fine answer; the worst-case depth quantifies this. Attack procedures are described only at the level needed to evaluate worst cases and are standard in certified-robustness work.

REPRODUCIBILITY STATEMENT

All definitions and standing assumptions are stated in Section 3; complete proofs are in Appendix B, and Appendix C describes the exhaustive, constructive and randomized checks of every closed form. Datasets, splits, source construction, calibration, the attacks used for each rule and all hyper-parameters are given in Appendix D. All aggregation experiments run on cached source predictions on a laptop CPU and use fixed seeds. The code, a Python package with experiment scripts, implements every rule, the closed-form worst cases, the attacks and the certificates, and includes tests that repeat the checks of Appendix C against brute-force enumeration.

REFERENCES

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, 2021.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *IEEE/CVF International Conference on Computer Vision*, pp. 9616–9625, 2019.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763, 2024.

Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defenses against general poisoning attacks. In *International Conference on Learning Representations*, 2021.

Srikanth Jagabathula, Lakshminarayanan Subramanian, and Ashwin Venkataraman. Identifying unreliable and adversarial workers in crowdsourced labeling tasks. *Journal of Machine Learning Research*, 18(93):1–67, 2017.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12503–12512, 2020.

Shyamgopal Karthik, Ameya Prabhu, Puneet K. Dokania, and Vineet Gandhi. No cost likelihood manipulation at test time for making better mistakes in deep networks. In *International Conference on Learning Representations*, 2021.

Jack Valmadre. Hierarchical classification at multiple operating points. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18034–18045, 2022.

Shani Goren, Ido Galil, and Ran El-Yaniv. Hierarchical selective classification. In *Advances in Neural Information Processing Systems*, volume 37, pp. 111047–111073, 2024.

T. Margush and F. R. McMorris. Consensus n-trees. *Bulletin of Mathematical Biology*, 43(2): 239–244, 1981.

Dahlia Malkhi and Michael Reiter. Byzantine quorum systems. *Distributed Computing*, 11(4): 203–213, 1998.

Jia Deng, Jonathan Krause, Alexander C. Berg, and Li Fei-Fei. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3450–3457, 2012.

Thomas Mortier, Alireza Javanmardi, Yusuf Sale, Eyke Hüllermeier, and Willem Waegeman. Conformal prediction in hierarchical classification with constrained representation complexity. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1324–1332, 2026.

Botong Zhang, Shuo Li, and Osbert Bastani. Conformal structured prediction. In *International Conference on Learning Representations*, 2025.

Floris den Hengst, Inès Blin, Majid Mohammadi, Syed Ihtesham Hussain Shah, and Taraneh Younesian. Hierarchical conformal classification. *arXiv preprint arXiv:2508.13288*, 2025.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2022.

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, 2024.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J. Zico Kolter. Assessing generalization of SGD via disagreement. In *International Conference on Learning Representations*, 2022.

Kanishk Jain, Shyamgopal Karthik, and Vineet Gandhi. Test-time amendment with a coarse classifier for fine-grained classification. In *Advances in Neural Information Processing Systems*, volume 36, pp. 63372–63384, 2023.

Sergio Albeverio, Andrei Khrennikov, and Brunello Tirozzi. p-adic dynamical systems and neural networks. *Mathematical Models and Methods in Applied Sciences*, 9(9):1417–1437, 1999.

Fionn Murtagh. On ultrametricity, data coding, and computation. *Journal of Classification*, 21(2): 167–184, 2004.

André F. T. Martins. Learning with the p-adics. *arXiv preprint arXiv:2512.22692*, 2025.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Wenxiao Wang, Alexander Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (Deterministic) finite aggregation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22769–22783, 2022.

Keivan Rezaei, Kiarash Banihashem, Atoosa Chegini, and Soheil Feizi. Run-off election: Improved provable defense against data poisoning attacks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29030–29050, 2023.

Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Intrinsic certified robustness of bagging against data poisoning attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7961–7969, 2021.

Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Provably secure federated learning against malicious clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6885–6893, 2021.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320, 2019.

Alexander Levine and Soheil Feizi. (De)randomized smoothing for certifiable defense against patch attacks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6465–6475, 2020.

Yanting Wang, Hongye Fu, Wei Zou, and Jinyuan Jia. MMCert: Provable defense against adversarial attacks to multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24655–24664, 2024.

Alaa Anani, Tobias Lorenz, Bernt Schiele, and Mario Fritz. Adaptive hierarchical certification for segmentation using randomized smoothing. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1531–1556, 2024.

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, 2017.

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, 2018.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022.

Javad Parsa, Amir Hossein Daghestani, André M. H. Teixeira, and Mikael Johansson. Byzantine-robust federated learning with learnable aggregation weights. In *International Conference on Learning Representations*, 2026.

Akash Dhasade, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Maxime Jacovella, Anne-Marie Kermarrec, and Rafael Pinot. Robust federated inference. In *International Conference on Learning Representations*, 2026.

Emily Black, Klas Leino, and Matt Fredrikson. Selective ensembles for consistent predictions. In *International Conference on Learning Representations*, 2022.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Naoki Otani, Yukino Baba, and Hisashi Kashima. Quality control for crowdsourced hierarchical classification. In *2015 IEEE International Conference on Data Mining*, pp. 937–942, 2015. doi: 10.1109/ICDM.2015.83.

Grant Van Horn, Steve Branson, Scott Loarie, Serge Belongie, and Pietro Perona. Lean multiclass crowdsourcing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2714–2723, 2018.

Jacob Steinhardt, Gregory Valiant, and Moses Charikar. Avoiding imposters and delinquents: Adversarial crowdsourcing and peer prediction. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 47–60, 2017.

Jean-Pierre Barthélemy and Melvin F. Janowitz. A formal theory of consensus. *SIAM Journal on Discrete Mathematics*, 4(3):305–322, 1991.

Leslie Lamport, Robert Shostak, and Marshall Pease. The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982.

Gil Neiger. Distributed consensus revisited. *Information Processing Letters*, 49(4):195–201, 1994.

Pierre Civit, Seth Gilbert, Rachid Guerraoui, Jovan Komatovic, and Manuel Vidigueira. On the validity of consensus. In *Proceedings of the 2023 ACM Symposium on Principles of Distributed Computing*, pp. 332–343, 2023.

Christian Cachin, Rachid Guerraoui, and Luís Rodrigues. *Introduction to Reliable and Secure Distributed Programming*. Springer, second edition, 2011.

Nitin H. Vaidya and Vijay K. Garg. Byzantine vector consensus in complete graphs. In *Proceedings of the 2013 ACM Symposium on Principles of Distributed Computing*, pp. 65–73, 2013.

Hammurabi Mendes and Maurice Herlihy. Multidimensional approximate agreement in Byzantine asynchronous systems. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, pp. 391–400, 2013.

Thomas Nowak and Joel Rybicki. Byzantine approximate agreement on graphs. In *33rd International Symposium on Distributed Computing (DISC 2019)*, volume 146 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 29:1–29:17, 2019.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, volume 2430 of *Lecture Notes in Computer Science*, pp. 345–356. Springer, 2002. doi: 10.1007/3-540-36755-1_29.

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.

iNaturalist. What are the community taxon and the observation taxon? https://help.inaturalist.org/en/support/solutions/articles/151000173076, 2026. iNaturalist Help; accessed September 2026.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, 2017.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

A NOTATION AND POSITIONING

Table 4 lists the symbols used in the paper, and Table 5 states what the closest literatures give and what this work adds.

Table 4: Notation.

Symbol	Meaning
$\mathcal{Y}$, L	labels (leaves of the hierarchy) and their common depth
B, $B_k(y)$	a ball (the set of labels below a node); the depth-k ball containing label y
M, f	number of sources; largest number of corrupted sources the rule must tolerate
$[M]$, y	the set of sources $\{1, \dots, M\}$; the true label of an example
s, s'	a profile of reports, one per source (the clean profile in Corollaries 1, 2 and Theorem 3); an observed profile after corruption
Z, h, n	corrupted sources ($\lvert Z\rvert \le f$); honest reports $s'\rvert_{[M]\setminus Z}$; their number $M - \lvert Z\rvert$
$n_s(B)$	number of reports of profile s inside ball B (total mass for probability reports)
$C(h)$	honest-majority ball: deepest ball holding more than half of the honest reports
τ_f	least integer above $(M+f)/2$, $\lfloor (M+f)/2 \rfloor + 1$
$Q_f(s)$	finest f-safe rule: deepest ball with $n_s(B) > (M+f)/2$
A_t	output on example t of a rule that sees the whole stream of profiles (Proposition 7)
$\rho_f(h)$	depth of the deepest ball with more than $(M+f)/2$ honest reports (0 if none)
$E_f(s,y)$, $R_f^\star$	worst-case error indicator of Q_f on an example (regime A); its mean, the worst-case risk
$E_{f,Z}$, R_Z	worst-case error indicator and risk of Q_f when the same set Z is corrupted on every example (regime B)
$K_f(s)$, $U_f(s)$	contested indicator (two disjoint balls above $(M-f)/2$ clean reports); deepest such ball otherwise
N, e, e_Z, δ	held-out sample size; its numbers of examples with $E_f = 1$ and with $E_{f,Z} = 1$; one minus the confidence level
$\mathrm{Beta}^{-1}(q; a, b)$	q-quantile of the $\mathrm{Beta}(a, b)$ distribution (Clopper-Pearson bound)
X_k	number of clean reports inside $B_k(y)$
a_k, κ_k	mean source accuracy at level k; intra-class correlation of the correctness of the sources
$Q_{2f}(s)$, $D_f(s')$	clean-anchored rule, deepest ball with $n_s(B) > M/2 + f$; profiles within f reports of s'
α, $C_\alpha(h)$	supermajority level; deepest ball holding more than $\alpha\lvert h\rvert$ honest reports
$\mu_m(B)$, $\mathrm{top}_j(B)$	mass that source m puts on B; sum of the j largest $\mu_m(B)$
$S_{M-f}(u)$	sum of the $M-f$ largest $\mu_m(u)$
f', $\mathbf{e}_a$	design budget of $Q^{\mathrm{p}}_{f'}$ attacked by f sources (Lemma 3); point mass on label a
$Q^{\mathrm{p}}_f(s)$	finest f-safe ball-valued rule for probability reports (Proposition 1)
θ, θ_c, θ_c^+	climbing threshold; safety threshold $\frac{1}{2} + \frac{f}{2M}$ (climbing is f-safe iff $\theta > \theta_c$); $\theta_c + 10^{-6}$
w_m, W, F, F_B	source weight; total weight; weight budget; largest corruptible weight inside B (Prop. 8)
n_w, g_u, ε_u, t_ℓ	top-down vote: count of the chosen child, its margin over sibling u, tie indicator, flip cost

Table 5: What the closest literatures give, and what they do not give for the question studied here.

Line of work	Gives	Does not give
Majority-rule consensus of trees (Margush & McMorris, 1981; Barthélemy & Janowitz, 1991)	the ℓ-majority consensus of reported paths	safety under corruption; optimality among set-valued or stateful rules; error against the truth
Byzantine quorum systems (Malkhi & Reiter, 1998; Cachin et al., 2011)	quorum sizes and the $M > 3f$ condition for overlapping correct servers	the most specific statement for an observed profile; its loss of depth; its label-error risk
Stability certificates for voting (Levine & Feizi, 2021; Jia et al., 2021)	decisions that f replacements cannot change	coarsening; honest-majority safety, which keeps unanimous answers for $M > 3f$ instead of $M > 4f$
Hierarchical selective and conformal prediction (Goren et al., 2024; Mortier et al., 2026)	risk or coverage control by backing off on clean data	a threshold that keeps backing off safe when sources are corrupted
Robust aggregation of predictions (Dhasade et al., 2026; Yin et al., 2018)	bounds on flat accuracy under Byzantine sources	abstention by coarsening; exact worst-case risk; held-out certificate
Reliability models for crowdsourcing (Dawid & Skene, 1979)	lower error when sources behave as the model assumes	worst-case safety (Proposition 7)
This work	for label reports, the finest safe output on every profile, even among set-valued and stateful rules; its exact loss of depth and exact error, certified on held-out data; the threshold θ_c for probability reports	

B PROOFS

Throughout, (A1) to (A3) hold, s' is an observed profile of M reports, Z a set of at most f sources, $h = s'|_{[M]\setminus Z}$ and $n = |h| = M - |Z|$; $\operatorname{depth} B$ is the depth of a ball B, and $\tau_f = \lfloor (M+f)/2 \rfloor + 1$ is the least integer above $(M+f)/2$.

Proof of Lemma 1. Let $B_j(y)$ and $B_k(y')$ with $j \le k$ share a label z. Then $B_j(y) = B_j(z)$ and $B_k(y') = B_k(z)$, and $B_k(z) \subseteq B_j(z)$ because sharing k decisions implies sharing $j \le k$. (Equivalently, the balls are those of the ultrametric with $d(y,y) = 0$ and $d(y,y') = 2^{-k}$ for distinct labels sharing exactly k decisions, which coding the decisions as base-p digits turns into p-adic balls; no result below uses this coding.) □

Lemma 2 (Majority chain)**.** *For a multiset h of $n \ge 1$ reports, the balls B with $n_h(B) > n/2$ form a chain that contains the root. Its deepest element $C(h)$ satisfies: a ball B holds a strict majority of h if and only if $B \supseteq C(h)$.*

Proof. Two balls with strict majorities intersect (otherwise their counts add to more than n), so they are nested by Lemma 1; the root holds all n reports. A finite chain has a deepest element. If $B \supseteq C(h)$ then $n_h(B) \ge n_h(C(h)) > n/2$. Conversely, a ball with a strict majority is on the chain and therefore contains its deepest element. □

Proof of Theorem 1. The balls with $n_s(B) > (M+f)/2 \ge M/2$ form a chain containing the root (as in Lemma 2; the root holds $M > (M+f)/2$ reports because $f < M$), so $Q_f(s)$ is well defined.

(i) Let $B = Q_f(s')$. Removing the $|Z| \le f$ corrupted reports removes at most $|Z|$ reports from B, so $n_h(B) \ge n_{s'}(B) - |Z| > \frac{M+f}{2} - |Z| \ge \frac{M-|Z|}{2} = \frac{n}{2}$, using $f \ge |Z|$. By Lemma 2, $B \supseteq C(h)$.

(ii) Let $B = Q_f(s)$. If B is a leaf, $Z = \emptyset$ works, since $n_s(B) > M/2$. Otherwise let B' be a child of B with the largest count $c = n_s(B')$. Because B' is deeper than $Q_f(s)$, $c \le (M+f)/2$,

so $z = \max(0, 2c - M) \le f$ and $z \le c$. Let Z be z sources that report in B' and $h = s|_{[M]\setminus Z}$, with $|h| = M - z$. Then $n_h(B) = n_s(B) - z > \frac{M+f}{2} - z \ge \frac{M-z}{2}$, so B holds a strict majority of h. The child B' keeps $c - z \le (M - z)/2$ honest reports. Any other child B'' keeps $n_s(B'')$ reports, and $n_s(B'') \le c$ by the choice of B' while $n_s(B'') \le M - c$ because B' and B'' are disjoint. If $z = 0$ then $c \le M/2$ and $n_s(B'') \le c \le (M - z)/2$; if $z = 2c - M > 0$ then $n_s(B'') \le M - c = (M - z)/2$. So no child of B holds a strict majority of h, and by Lemma 2 $C(h) = B$.

(iii) A rule whose output contains $C(s|_{[M]\setminus Z})$ for every admissible Z contains $Q_f(s)$ by (ii), whether it outputs a ball or any set of labels and whatever it knows about the sources, because (ii) exhibits one admissible Z with $C(s|_{[M]\setminus Z}) = Q_f(s)$. Conversely, a ball-valued rule with $A(s) \supseteq Q_f(s)$ for every s is f-safe by (i). □

Proof of Theorem 2. (i) is Theorem 1(i).

(ii) Every ball B with $n_h(B) > (M + f)/2$ also has $n_{s'}(B) \ge n_h(B) > (M + f)/2$, so it is on the chain defining $Q_f(s')$ and $Q_f(s') \subseteq B$; hence $\operatorname{depth} Q_f(s') \ge \rho_f(h)$. For attainment, let B^* be the deepest ball with $n_h(B^*) > (M + f)/2$, or the root if there is none, and let T be a top-level ball with the fewest honest reports. The root has at least two children, so $n_h(T) \le n/2 = (M - |Z|)/2$. Let every corrupted source report a leaf in T. A ball inside T then holds at most $n_h(T) + |Z| \le (M + |Z|)/2 \le (M + f)/2$ observed reports; a ball outside T receives no corrupted report and keeps exactly its honest count; and the root is the only ball that contains T, because T is a top-level ball. So the non-root balls with more than $(M + f)/2$ observed reports are exactly those with more than $(M + f)/2$ honest reports, and $Q_f(s') = B^*$.

(iii) If all honest reports lie in B then $n_{s'}(B) \ge n = M - |Z| \ge M - f > (M + f)/2$ whenever $M > 3f$, so $Q_f(s') \subseteq B$, and $Q_f(s') \supseteq C(h)$ by (i). If $M \le 3f$ and $|Z| = f$, every ball has $n_h(B) \le M - f \le (M + f)/2$, so $\rho_f(h) = 0$ and, by (ii), every profile can be driven to the root. For the second claim, let $M - f$ sources report a label y_1 and f sources a label y_2 in another top-level ball. Every non-root ball holds at most $M - f \le (M + f)/2$ reports, so $Q_f(s')$ is the root and, by Theorem 1(iii), every f-safe rule outputs the root; yet when Z is the set of sources reporting y_2, all honest sources report y_1. □

Proof of Corollary 1. A corruption of the clean profile s is a choice of Z and of the reports of Z. By Theorem 2(ii), for a fixed Z the least depth over the reports is $\rho_f(s|_{[M]\setminus Z})$, the depth of the deepest ball with at least τ_f reports in $s|_{[M]\setminus Z}$. Removing Z lowers the count of every ball by at most $|Z| \le f$, so a ball with $n_s(B) \ge \tau_f + f$ keeps at least τ_f reports for every Z. Conversely, the balls with $n_s(B) \ge \tau_f > M/2$ form a chain; taking Z to be f sources that report inside its deepest element (it holds at least $\tau_f \ge f$ reports) lowers every ball of the chain by exactly f, and balls off the chain hold fewer than τ_f reports anyway. So the least depth is the largest k such that some depth-k ball has $n_s(B) \ge \tau_f + f$, or 0.

For the second claim, let $k \ge 1$. If $n_s(B_k(y)) \ge \tau_f + f$, then for every Z the ball $B_k(y)$ keeps at least τ_f honest, hence observed, reports, so $Q_f(s') \subseteq B_k(y)$ as in the proof of Theorem 2(ii). If $n_s(B_k(y)) < \tau_f + f$, let Z be $\min(f, n_s(B_k(y)))$ sources reporting inside $B_k(y)$ and park their reports as in Theorem 2(ii). Then $n_h(B_k(y)) < \tau_f$, so neither $B_k(y)$ nor any ball inside it is on the honest chain of ρ_f, and $Q_f(s') = B^* \not\subseteq B_k(y)$ because B^* is not inside the non-root ball $B_k(y)$. (For $k = 0$ the inclusion always holds, while $n_s(\mathcal{Y}) = M < \tau_f + f$ when $M \le 3f$.) □

Proof of Theorem 3. (i) Let $u \not\ni y$ be a ball with $n_s(u) > (M - f)/2$. If at least f sources report outside u, corrupt f of them and move their reports into u: $n_{s'}(u) = n_s(u) + f > (M + f)/2$, so $Q_f(s') \subseteq u$ excludes y. Otherwise corrupt every source outside u (fewer than f) and move them into u: $n_{s'}(u) = M > (M + f)/2$ because $f < M$. Conversely, if some corruption makes $y \notin Q_f(s')$, then $C(h) \subseteq Q_f(s')$ excludes y and $u = C(h)$ satisfies $n_s(u) \ge n_h(u) > (M - |Z|)/2 \ge (M - f)/2$. Moreover $\max_{|Z|\le f} \mathbf{1}[y \notin C(s|_{[M]\setminus Z})] = E_f(s, y)$: removing $\min(f, M - n_s(u))$ sources that report outside such a u leaves u with a strict majority of the remaining reports, so $C(h) \subseteq u$ excludes y, and conversely $u = C(h) \not\ni y$ has $n_s(u) \ge n_h(u) > (M - f)/2$. For the depth of the error, every ball excluding y lies inside a sibling of some $B_j(y)$, and the argument above applied to that sibling u shows that it can be made to contain the output exactly when $n_s(u) > (M - f)/2$, in which case the output agrees with y on the first $j - 1$ decisions only.

(ii) Fix Z with $|Z| = f$ and let $h = s|_{[M]\setminus Z}$. If some $u \not\ni y$ has $n_h(u) > (M-f)/2$, let every corrupted source report a leaf of u: $n_{s'}(u) = n_h(u) + f > (M+f)/2$ and $Q_f(s') \subseteq u$. Conversely, if $Q_f(s') = B \not\ni y$, then $n_h(B) \geq n_{s'}(B) - f > (M-f)/2$. So on every example the worst case over the reports of Z is $\mathbf{1}[\max_{u \not\ni y} n_h(u) > (M-f)/2]$, and its mean is R_Z. This indicator is at most $E_f(s,y)$ because $n_h(u) \leq n_s(u)$, so $\max_Z R_Z \leq R^\star_f$. Restricting to $|Z| = f$ loses nothing: if $Z \subseteq Z'$ with $|Z'| = f$, removing the extra $f - |Z|$ sources lowers $n_h(u)$ by at most $f - |Z|$ while the extra corrupted reports raise $n_{s'}(u)$ by exactly that much, so Z' is at least as damaging. The maximum over Z takes $\binom{M}{f}$ evaluations, 45 or 120 for our $M = 10$, but grows combinatorially with M.

(iii) If $K_f(s) = 1$, one of the two disjoint balls excludes y, so $E_f(s,y) = 1$. Otherwise any two balls holding more than $(M-f)/2$ clean reports intersect, so these balls form a chain containing the root, with deepest element $U_f(s)$. Some ball of the chain excludes y if and only if its deepest element does, so $E_f(s,y) = \mathbf{1}[y \notin U_f(s)]$. Two disjoint balls above the threshold at different depths imply two such balls at the shallower depth (the ancestor of the deeper one), so K_f can be computed depth by depth.

All three parts hold verbatim for errors at depth at most k (for example top-level errors, $k = 1$), with u and the balls in K_f and U_f restricted to depth at most k; these are the top-level certificate and contested share of Section 5.4. □

Proof of Corollary 2. If the N held-out examples are i.i.d. draws from the deployment distribution, the held-out indicators E_f are independent Bernoulli variables with mean $R^\star_f$, and the one-sided Clopper-Pearson bound $\bar{R} = \mathrm{Beta}^{-1}(1-\delta;\, e+1,\, N-e)$ satisfies $\Pr[R^\star_f \leq \bar{R}] \geq 1-\delta$ (Clopper & Pearson, 1934). For a fixed Z with $|Z| = f$, the indicators $E_{f,Z}$ are independent Bernoulli variables with mean R_Z by Theorem 3(ii), so the bound computed from e_Z at level $\delta/\binom{M}{f}$ fails with probability at most $\delta/\binom{M}{f}$; a union bound over the $\binom{M}{f}$ sets gives simultaneous coverage with probability at least $1-\delta$. Finally $K_f \leq E_f$ pointwise by Theorem 3(iii), so $\mathbb{E}[K_f] \leq R^\star_f$ on any distribution, and K_f does not involve y. □

Proposition 2 (Clean anchor)**.** *Let $2f < M$ and, for an observed profile s', let $D_f(s')$ be the set of profiles that differ from s' in at most f reports. The finest ball-valued rule A with $A(s') \supseteq C(s)$ for every s' and every $s \in D_f(s')$ is Q_{2f}, the deepest ball with $n_{s'}(B) > M/2 + f$ (the root if there is none). Some corruption of at most f sources of a clean profile s makes Q_{2f} exclude the true label y if and only if $C(s)$ excludes y. If the $M-f$ honest sources all report inside a ball B, $Q_{2f}(s') \subseteq B$ for every corruption if and only if $M > 4f$.*

Proof. If $n_{s'}(B) > M/2 + f$, every $s \in D_f(s')$ has $n_s(B) > M/2$, so $B \supseteq C(s)$ by Lemma 2; hence Q_{2f} has the property. Conversely, let B' be a child of $Q_{2f}(s')$, so $c = n_{s'}(B') \leq M/2 + f$, and let $s \in D_f(s')$ move $\min(f,c)$ of the reports in B' to a leaf outside $Q_{2f}(s')$ (or outside B' if $Q_{2f}(s')$ is the root). Then $n_s(B') \leq M/2$, so $C(s) \not\subseteq B'$. A rule with the property contains $C(s)$ for all these s and $C(s') \subseteq Q_{2f}(s')$, so its output is not inside any child of $Q_{2f}(s')$ and not disjoint from $Q_{2f}(s')$; being a ball, it contains $Q_{2f}(s')$. For the risk: if $C(s) \not\ni y$, some wrong ball u has $n_s(u) > M/2$, and moving $\min(f, M - n_s(u))$ reports into u gives $n_{s'}(u) > M/2 + f$ (if fewer than f reports are outside u, $n_{s'}(u) = M > M/2 + f$ because $2f < M$). Conversely, if $Q_{2f}(s') \not\ni y$, the output ball has more than $M/2 + f$ observed reports and hence more than $M/2$ clean ones, so it contains $C(s)$, which then excludes y. The last claim follows because the worst case puts all f corrupted reports outside B, leaving $M-f$ reports, and $M - f > M/2 + f$ if and only if $M > 4f$. □

Proof of Proposition 1. (i) A ball with $n_s(B) - \mathrm{top}_j(B) > (M-j)/2$ for $j = 0$ holds more than $M/2$ of the mass, so these balls are nested (Lemma 1); the root has $n_s = M$ and $\mathrm{top}_j = j$, and $M - j > (M-j)/2$. If B satisfies the condition and $|Z| = j \leq f$, the honest mass in B is at least $n_s(B) - \mathrm{top}_j(B) > (M-j)/2 = |h|/2$, so $B \supseteq C(h)$ as in Lemma 2, which holds verbatim for masses. Conversely, if a child B' of $Q^{\mathrm{p}}_f(s)$ fails the condition, some $j \leq f$ has $n_s(B') - \mathrm{top}_j(B') \leq (M-j)/2$; taking Z to be the j sources with the largest mass on B' leaves B' without an honest majority, so $C(h) \not\subseteq B'$. A safe ball-valued output therefore lies inside no child

of $Q_f^{\mathrm{p}}(s)$, and it meets $Q_f^{\mathrm{p}}(s)$ because both contain $C(s)$; so it contains $Q_f^{\mathrm{p}}(s)$. For point masses $\mathrm{top}_j(B) = \min(j, n_s(B))$, and $n_s(B) - \min(j, n_s(B)) > (M-j)/2$ for all $j \le f$ is equivalent to $n_s(B) > (M+f)/2$.

(ii) If $S_{M-f}(u) > (M-f)/2$, corrupt the f sources with the smallest mass on u into point masses in u. For $j \le f$ the j largest observed masses on u are ones, so $n_{s'}(u) - \mathrm{top}_j(u) = S_{M-f}(u) + f - j > (M-f)/2 + f - j \ge (M-j)/2$, and the output lies in u. Conversely, if the output excludes y, it lies inside a maximal wrong ball u, which then satisfies the condition of (i) in s': $n_{s'}(B) - \mathrm{top}_j(B)$ is the sum of the $M-j$ smallest masses on B, which can only grow from B to a ball $u \supseteq B$ because $\mu_m(u) \ge \mu_m(B)$ for every m. With $j = f$, the sum of the $M-f$ smallest observed masses on u exceeds $(M-f)/2$; it is at most the sum of the masses of any $M-f$ honest sources, hence at most $S_{M-f}(u)$. For point masses $S_{M-f}(u) = \min(n_s(u), M-f)$, and (ii) reduces to Theorem 3(i).

(iii) If $\theta > \frac{1}{2} + \frac{f}{2M}$, the output of climbing is the root or a ball with $n_{s'}(B) \ge \theta M > (M+f)/2$, whose honest mass exceeds $(M+f)/2 - |Z| \ge |h|/2$; so it contains $C(h)$. If $\theta \le \frac{1}{2} + \frac{f}{2M}$, take two leaves a and b in different top-level balls, let f sources report a point mass on a and the other $M-f$ sources put mass $\frac{1}{2}$ on a and $\frac{1}{2}$ on b, and let Z be the first f sources. The honest mass is split evenly between two disjoint balls, so $C(h)$ is the root, while the mean mass of a is $\frac{1}{2} + \frac{f}{2M} \ge \theta$ and no leaf has more. Climbing therefore starts from a (or, if $f = 0$ and a tie is broken towards b, from b, whose mass is then also $\frac{1}{2} \ge \theta$) and outputs that leaf, which does not contain $C(h)$.

(iv) By (i), an f-safe ball-valued rule A satisfies $A(s') \supseteq Q_f^{\mathrm{p}}(s')$ on every observed profile s', and climbing with $\theta > \theta_c$ is f-safe by (iii). Hence $\mathrm{depth}\, A(s') \le \mathrm{depth}\, Q_f^{\mathrm{p}}(s')$, and if $A(s')$ excludes y so does $Q_f^{\mathrm{p}}(s') \subseteq A(s')$: every corruption that makes A err makes Q_f^{p} err. □

The rule of (i) also has a certified depth: B survives every corruption of f sources if it still satisfies the condition after its f largest contributions are set to zero. When the design budget is at least the attack budget, as for Q_f^{p} attacked by f sources, the certified depth is the exact worst-case depth. Let B be the deepest surviving ball and B' its heaviest child. Corrupt the f largest contributors to B' into point masses in the top-level ball T with the least honest mass. T does not contain B: the honest mass of B exceeds $(M-f)/2$ because B survives, and that of T does not. So B still qualifies, B' sees exactly its zeroed profile and fails, every other child of B holds at most $M/2$ and fails, and no ball inside T qualifies, since at $j = f$ it keeps at most the honest mass of T. For a smaller design budget the parked point masses can make T qualify, but corruptions that split their mass still attain the certified depth, so it is exact for every design budget. If B is not the root, report the f largest contributors to B' as point masses outside B: B keeps more than $M/2$, so no ball disjoint from B qualifies, B' sees its zeroed profile and the other children hold at most $M/2$. If B is the root, let them put no mass on B' and spread it over the other top-level balls without raising any above $M/2$, which is possible with at least three top-level balls or when B' keeps honest mass $H \ge M/2$; with two top-level balls and $H < M/2$, give each of them mass $(M/2 - H)/f \in [0, 1]$ on B' and the rest on the other ball, so both hold exactly $M/2$ (the other ball has at most $M/2$ clean mass, so $H \ge M/2 - f$).

The set-valued part of Theorem 1(iii) does not extend to probability reports. On a tree with two top-level balls of two labels each ($00, 01$ and $10, 11$), let $M = 4$, $f = 1$, write $\mathbf{e}_a$ for the point mass on label a, and let the reports be $\frac{2}{3}\mathbf{e}_{00} + \frac{1}{3}\mathbf{e}_{11}$, $\mathbf{e}_{00}$, $\frac{1}{2}\mathbf{e}_{00} + \frac{1}{6}\mathbf{e}_{01} + \frac{1}{3}\mathbf{e}_{11}$ and $\mathbf{e}_{11}$. The top-level ball $\{00, 01\}$ holds mass $\frac{7}{3}$ but only $\frac{4}{3} \le \frac{3}{2}$ without its largest contribution, so Q_1^{p} is the root; yet the honest-majority balls over $|Z| \le 1$ are $\{00\}$, $\{00, 01\}$ and $\{11\}$, so the set-valued output $\{00, 01, 11\}$ is safe on this profile. Proposition 1 is therefore stated for ball-valued rules.

Lemma 3 (Design and attack budgets)**.** *Let $Q_{f'}^{\mathrm{p}}$ be designed for f' corruptions and attacked by f. Some corruption makes it exclude y if and only if some ball $u \not\ni y$ satisfies the condition of Proposition 1(i) with budget f' in the profile made of f point masses in u and the $M-f$ largest clean masses of u; for $f' \le f$ this reads $S_{M-f}(u) > (M - 2f + f')/2$.*

Proof. The argument of Proposition 1(ii) applies with the design budget f' in the condition and the attack budget f in the corruption: point masses in u on the f sources with the least mass on u are optimal, because raising a corrupted mass never lowers the sum of the $M-j$ smallest observed

masses. For $f' \le f$ the j largest observed masses are ones for $j \le f'$, so u qualifies if and only if $S_{M-f}(u) + f - j > (M-j)/2$ for all $j \le f'$, and the binding constraint is $j = f'$. □

Proposition 3 (Certified upper bounds for climbing)**.** *Let climbing act on the mean, the coordinate-wise trimmed mean or the coordinate-wise median of the reports. An f-corruption can make its output exclude y only if some sibling u of a ball $B_j(y)$ can reach aggregated mass θ. For the mean, the largest reachable mass of u is $(S_{M-f}(u) + f)/M$; for the trimmed mean and the median, every coordinate stays between the clean order statistics shifted by f positions. The share of examples on which some such u can reach θ is therefore an upper bound on the worst-case error, and for the mean at $\theta = \frac{1}{2} + \frac{f'}{2M}$ it equals the exact risk of $Q^{\mathrm{p}}_{f'}$ under $f \ge f'$ corruptions up to ties.*

Proof of Proposition 3. If climbing outputs a ball $B \not\ni y$, the aggregated mass of B is at least θ, and B lies inside a maximal wrong ball u, a sibling of some $B_j(y)$, whose aggregated mass is at least that of B because the aggregates are nonnegative. So an error requires some such u to reach mass θ under the corruption. For the mean, the corrupted sources can add at most mass 1 each to u and the adversary chooses them, so the largest reachable mean mass is $(S_{M-f}(u) + f)/M$, attained by turning the f sources with the least mass on u into point masses in u. For the coordinate-wise trimmed mean and median, replacing f of the M values of a coordinate moves each of its order statistics by at most f positions, so every aggregated coordinate lies between the average of the correspondingly shifted clean order statistics (clipped to 0 and 1 at the ends); the renormalized mass of u is increasing in the coordinates inside u and decreasing in the others, so the largest reachable mass is at most the value at these bounds. Counting the examples on which some u can reach θ gives an upper bound on the worst-case error. For the mean at $\theta = \frac{1}{2} + \frac{f'}{2M}$, the event $S_{M-f}(u) + f \ge \theta M$ reads $S_{M-f}(u) \ge (M - 2f + f')/2$, the event of Lemma 3 up to ties. □

Proposition 4 (Stability of top-down plurality; hierarchical DPA)**.** *Top-down plurality voting chooses, at each depth, the child of the current ball with most reports (ties broken towards the first child in a fixed order). At level ℓ let w be the chosen child, n_w its count and, for each sibling u, $g_u = n_w - n_u \ge 0$ and $\varepsilon_u = 1$ if a tie is resolved in favour of w and 0 otherwise. The least number of replaced reports that changes the chosen child at level ℓ while keeping the levels above is $t_\ell = \min_u \lceil (g_u + \varepsilon_u)/2 \rceil$, and the least number that changes the depth-k prefix is $\min_{\ell \le k} t_\ell$.*

Proof. A replaced report leaves one ball and enters another, so it changes the margin between two siblings by at most 2, and by 2 only if it leaves w and enters u. Reports moved within the ball chosen at level $\ell - 1$ do not change any count above it. The chosen child holds at least one report, and $g_u + \varepsilon_u \le n_w + 1 \le 2n_w$, so moving $\lceil (g_u + \varepsilon_u)/2 \rceil \le n_w$ reports from w to u is possible and closes the margin, which gives t_ℓ. Changing the depth-k prefix requires changing the first level at which it changes, whose cost is at least the corresponding t_ℓ (the levels above it are unchanged). □

This is the flat certificate of Deep Partition Aggregation (Levine & Feizi, 2021) applied along the chosen path; for $L = 1$ it is the cost of changing the plurality label. The worst cases of plurality and top-down votes that we report follow from these costs on the clean profile and are checked against brute force (Appendix C). For the majority ball the analogous stability condition is $n_s(B) - f > M/2$, the clean-anchor quota of Proposition 2, compared with $n_s(B) > (M + f)/2$ for safety.

Proposition 5 (Stability used after corruption)**.** *Consider the rule that computes the stability costs of Proposition 4 on the* observed *profile and outputs the top-down plurality path truncated at the deepest depth k with $\min_{\ell \le k} t_\ell > f$. If the $M - f$ honest sources all report the same label y and, at some level, a sibling of the path of y wins ties against the path child, this rule outputs y under every corruption of f sources if and only if $M > 4f$ (if and only if $M \ge 4f$ when every sibling loses ties). Under the same conditions Q_f outputs y if and only if $M > 3f$.*

Proof. At every level the child on the path of y holds at least $M - f$ observed reports and any sibling at most f, so the margin is at least $M - 2f$, with equality when the f corrupted reports sit in one sibling u. With the tie resolved in favour of u, the cost of Proposition 4 is $\lceil (M - 2f)/2 \rceil$, which exceeds f exactly when $M - 2f > 2f$; with ties resolved against u it is $\lfloor (M - 2f)/2 \rfloor + 1$, which exceeds f exactly when $M \ge 4f$. The statement for Q_f is Theorem 2(iii). This is also why a stability certificate of radius f on an observed profile corresponds to radius $2f$ on the clean one: the two profiles can differ in f reports. □

Proposition 6 (Supermajority safety). *For $\alpha \in [\frac{1}{2}, 1)$ let $C_\alpha(h)$ be the deepest ball holding more than $\alpha|h|$ honest reports, and call a rule (f, α)-safe if its output always contains $C_\alpha(h)$. The deepest ball with $n_s(B) > \alpha M + (1-\alpha)f$ is the finest ball-valued (f, α)-safe rule, and it keeps an answer that all honest sources share if and only if $(1-\alpha)M > (2-\alpha)f$. For $\alpha = \frac{1}{2}$ this is Theorems 1 and 2(iii); for $\alpha = \frac{2}{3}$ it tolerates $f < M/4$.*

Proof. For $\alpha \geq \frac{1}{2}$ the balls holding more than $\alpha|h|$ reports form a chain (two disjoint ones would hold more than $|h|$ reports), so $C_\alpha(h)$ is well defined and contained in every such ball. If $n_s(B) > \alpha M + (1-\alpha)f$ then $n_h(B) > \alpha M + (1-\alpha)f - |Z| \geq \alpha(M - |Z|)$, because the difference is $(1-\alpha)(f - |Z|) \geq 0$; hence $B \supseteq C_\alpha(h)$. For necessity, let B^* be the deepest ball with $n_s(B) > \alpha M + (1-\alpha)f$ and B' one of its children, so $n_s(B') \leq \alpha M + (1-\alpha)f$. Let Z be $\min(f, n_s(B'))$ sources reporting inside B'. If $|Z| = f$, then $n_h(B') = n_s(B') - f \leq \alpha(M - f) = \alpha|h|$; otherwise $n_h(B') = 0$. Either way B' holds at most $\alpha|h|$ honest reports, so $C_\alpha(h) \not\subseteq B'$. A safe ball-valued output contains $C_\alpha(h)$ for this Z, so it lies inside no child of B^*, and it contains $C_\alpha(s) \subseteq B^*$ (the case $Z = \emptyset$), so it meets B^*; being a ball, it contains B^*. An answer shared by all $M - f$ honest sources clears the threshold if and only if $M - f > \alpha M + (1-\alpha)f$, i.e. $(1-\alpha)M > (2-\alpha)f$. □

Other anchors and other rule classes. Requiring the output to contain the true label under every corruption forces the root, because every source may err. Requiring it to contain $C(h)$ only with high probability would need a probabilistic model of the sources and of the adversary; the worst-case definition avoids both, and Theorem 3 measures, on the data distribution, how often the honest-majority anchor itself is wrong. Among the anchors we consider, the honest simple majority gives the most specific finest rule: the quotas of Propositions 2 and 6, $M/2 + f$ and $\alpha M + (1-\alpha)f$, are at least $(M + f)/2$. Rules that see the whole stream of examples gain no specificity:

Proposition 7 (Stateful rules). *Let a rule see a stream of observed profiles $s'_1, s'_2, \ldots$ and output A_t for every example t as a function of the whole stream, past and future, and of the source identities. Call it f-safe under persistent corruption if $A_t \supseteq C(s'_t|_{[M]\setminus Z})$ for every stream, every t and every set Z of at most f sources corrupted on every example. Then $A_t \supseteq Q_f(s'_t)$ for every stream and every t. For probability reports, every ball-valued rule that is f-safe under persistent corruption contains $Q^{\mathrm{p}}_f(s'_t)$.*

Proof. Fix a stream and t. By Theorem 1(ii) some Z^* with $|Z^*| \leq f$ has $C(s'_t|_{[M]\setminus Z^*}) = Q_f(s'_t)$. The reports of honest and of corrupted sources are unconstrained, so every stream is consistent with Z^* being corrupted on every example, and safety at t under Z^* gives $A_t \supseteq Q_f(s'_t)$. For probability reports, the proof of Proposition 1(i) gives, for every child B' of $Q^{\mathrm{p}}_f(s'_t)$, a set $Z_{B'}$ of at most f sources with $C(s'_t|_{[M]\setminus Z_{B'}}) \not\subseteq B'$; each is admissible for the whole stream, so a safe ball-valued output lies inside no child of $Q^{\mathrm{p}}_f(s'_t)$, meets it because both contain $C(s'_t)$, and therefore contains it. □

The argument uses that honest reports are unconstrained, as in (A3): if honest sources were known to be correct on labelled examples of the stream, the admissible sets would depend on the stream and a stream-aware rule could be finer than Q_f. A rule that estimates source reliabilities across examples and discounts the sources that deviate (for example the model of Dawid & Skene (1979)) is therefore not f-safe once it commits below $Q_f(s'_t)$ on some stream: if the discounted sources are honest and another set is corrupted, its output can exclude the honest-majority ball. It can have lower error than Q_f against the truth when the honest sources behave as its model assumes; Section 5.3 measures this.

Proposition 8 (Weighted sources). *Let source m carry weight $w_m > 0$, let $W = \sum_m w_m$, and let the adversary corrupt any set Z of total weight $w(Z) \leq F < W$. Let $C_w(h)$ be the deepest ball whose honest weight exceeds half of the honest total. For a profile s and a ball B, let $n_s(B)$ be the weight reporting inside B and F_B the largest total weight at most F of a set of sources reporting inside B. Then $B \supseteq C_w(s|_{[M]\setminus Z})$ for every admissible Z if and only if $n_s(B) > (W + F_B)/2$; these balls form a chain, and a ball-valued rule is safe if and only if it contains the deepest of them, $Q^w_F(s)$, on every profile. With unit weights and an integer budget F, $F_B = \min(F, n_s(B))$ and $Q^w_F = Q_F$. The set-valued part of Theorem 1(iii) need not hold.*

Proof. As in Lemma 2, $B \supseteq C_w(h)$ if and only if the honest weight in B exceeds half of the honest total, $n_s(B) - w(Z \cap S_B) > (W - w(Z))/2$, where S_B is the set of sources reporting inside B. Writing $w(Z) = w(Z \cap S_B) + w(Z \setminus S_B)$, the condition reads $n_s(B) > W/2 + w(Z \cap S_B)/2 - w(Z \setminus S_B)/2$, which is hardest for $Z \subseteq S_B$ of the largest admissible weight, F_B. Such balls hold more than $W/2$ and are nested; the root qualifies because $F_B \leq F < W$. If a child B' of the deepest qualifying ball $Q_F^w(s)$ fails the condition, some admissible Z has $C_w(h) \not\subseteq B'$; a safe ball-valued output contains all the balls $C_w(h)$, including $C_w(s) \subseteq Q_F^w(s)$, so it lies inside no child of $Q_F^w(s)$ and contains $Q_F^w(s)$. Conversely a ball-valued rule containing Q_F^w is safe. With integer weights the admissible corrupted weights inside B' may skip the value that makes a single Z realize $C_w(h) = Q_F^w(s)$ as in Theorem 1(ii), and the union of the balls $C_w(h)$ can then be strictly smaller than $Q_F^w(s)$. For example, with three labels under the root, weights $(2, 2, 3)$, budget $F = 2$ and reports $(0, 0, 1)$, the admissible sets are $\emptyset$ and either source of weight 2, with honest-majority balls $\{0\}$, $\{1\}$ and $\{1\}$; neither $\{0\}$ ($4 \leq (7+2)/2$) nor $\{1\}$ ($F_{\{1\}} = 0$ and $3 \leq 7/2$) meets the quota, so Q_F^w is the root, while the set $\{0, 1\}$ is safe on this profile. Our enumeration finds such profiles in 245 of 10,647 cases (Appendix C). Computing F_B is a subset-sum problem, easy for small integer weights. □

Proposition 9 (Which levels survive under independent sources)**.** *If on a random example the level-k correctness of the M sources are independent Bernoulli(a_k) variables, the share of examples on which Q_f is certified correct to depth k is* $\Pr[\mathrm{Binomial}(M, a_k) \geq \tau_f + f] \geq 1 - \exp(-2M(a_k - \frac{1}{2} - \frac{3f}{2M})^2)$ *when $a_k > \frac{1}{2} + \frac{3f}{2M}$, and Q_f can be driven wrong at depth at most k with probability at most* $\exp(-2M(a_k - \frac{1}{2} - \frac{f}{2M})^2)$ *when $a_k > \frac{1}{2} + \frac{f}{2M}$.*

Proof. By Corollary 1 the level-k decision is certified correct if and only if $X = n_s(B_k(y)) \geq \tau_f + f$, and $X \sim \mathrm{Binomial}(M, a_k)$; the complementary event $X \leq (M + 3f)/2$ is bounded by Hoeffding's inequality (Hoeffding, 1963). If Q_f can be driven wrong at depth at most k, Theorem 3 gives a sibling u of some $B_j(y)$ with $j \leq k$ and $n_s(u) > (M - f)/2$; since u is disjoint from $B_k(y)$, $X < (M + f)/2$, which Hoeffding's inequality bounds in the same way. □

With correlated sources the binomial is replaced by the distribution of X_k itself; a beta-binomial with the fitted intra-class correlation matches our data (Section 5.1).

C Verification of the theory by enumeration

Every closed form used in the paper was checked against brute force or against constructive and random attacks; no check failed.

Vote reports, exhaustive. We drew 44 random trees of depth 2 or 3 with at most 9 leaves and branching at most 3, and for $M \in \{3, \ldots, 7\}$ sources and $f \in \{0, \ldots, 3\}$ drew random clean profiles concentrated around a random label (for $f = 3$ only $M \in \{4, 5\}$ and trees with at most 6 leaves, because the number of corruptions grows as $(|\mathcal{Y}|M)^f$). For each profile we enumerated *every* corrupted set Z with $|Z| \leq f$ and *every* assignment of labels to the corrupted sources, and checked: Theorem 1(i) on 999,126 (profile, corruption) pairs; that the lowest common ancestor of $C(s|_{[M]\setminus Z})$ over all $|Z| \leq f$ equals $Q_f(s)$ on 4,260 profiles; the least attacked depth of Corollary 1 and the error condition of Theorem 3(i) on 4,260 profiles each; the second claim of Corollary 1 at every depth $k \geq 1$ on 5,112 (profile, f, depth) cases (at $k = 0$ it fails exactly when $M \leq 3f$, as its proof notes); Proposition 6 for $\alpha \in \{\frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}\}$: safety on 628,050 attacked profiles, the finest ball-valued output on 8,050 profiles and the unanimity condition in 385 cases; the most severe reachable error level for Q_f, top-down plurality and flat plurality on 12,780 cases; the stability costs of Proposition 4 on 10,212 (profile, depth) pairs; Proposition 5, for the stability rule and for Q_f, by enumerating every multiset of corrupted reports on a tree with 3×3 leaves for $2 \leq M \leq 13$ and every $f < M$ (360 checks); the one-third barrier for every $2 \leq M \leq 12$ and $0 \leq f < M$ on one extremal profile each; that the construction of Theorem 1(ii) yields $C(s|_{[M]\setminus Z}) = Q_f(s)$ on 352,000 (profile, f) cases with $M \leq 9$; Theorem 3(ii) for every fixed set Z with $|Z| = f \leq 2$, against every assignment of its reports, on 5,980 (profile, Z) cases; the fixed-set no-answer condition of Table 14 (no top-level ball keeps at least τ_f honest reports) and the identity between the fixed-set worst case of Q_f and the error of the honest-majority ball of the other sources (Section 5.3), against every assignment of the reports

of Z, on 1,752 (profile, Z) cases; that Q_{2f} is the finest clean-anchored rule and that its worst case equals the clean error of the majority ball (Proposition 2) on 4,480 profiles each, enumerating every clean profile within f replacements.

Vote reports, large instances. On 7,866 clean profiles with $M \in \{9, 11, 13, 15, 21\}$ sources, every $f < M$ with $\tau_f \leq M$, and random trees of depth 3 or 4 with 6 to 120 labels, we checked that the construction of Theorem 1(ii) yields $C(s|_{[M]\setminus Z}) = Q_f(s)$; that the constructive attack of Theorem 3(i) succeeds whenever $E_f = 1$, while 300 random and targeted corruptions never make Q_f err when $E_f = 0$; that the parking attack of Theorem 2(ii) attains the least depth of Corollary 1 and no attack goes below it; and safety of Q_f under all 2,375,532 attacked profiles. The decomposition of Theorem 3(iii) was checked at every level on 666,000 (profile, f, level) cases with up to 150 labels.

Probability reports. With masses in multiples of $1/8$ (exact in floating point) on random trees with at most 9 leaves, enumerating every $|Z| \leq f$ and every point-mass corruption: safety of Q_f^{p} on 265,872 (profile, Z) pairs and its finest property on 8,700 profiles; its exact worst-case deviation and certified depth for every design f' and attack budget $f \leq 2$ on 6,156 cases (795,810 attacked profiles, including random non-point-mass corruptions); the upper bounds of Proposition 3 on 13,680 cases; Proposition 1(iii) on 202,752 (profile, Z) pairs and on the construction of its proof for 2,376 thresholds at or below θ_c; $Q_f^{\mathrm{p}} = Q_f$ on point masses with identical worst cases and certified depths (3,400 checks); and that the certified depth equals the least depth over every corrupted set and every point-mass corruption whenever the design budget is at least the attack budget (6,720 cases), while for smaller design budgets it is a strict lower bound on the point-mass worst case in 442 of 3,360 cases, and the construction of Appendix B attains it once corrupted sources may split their mass. With continuous Dirichlet reports, $M \in \{5, 7, 10, 12\}$ and up to 40 labels: safety of Q_f^{p} on 264,264 attacked profiles; Proposition 1(ii) on 2,184 profiles (constructive attack when the condition holds, 120 random and targeted attacks otherwise); the bounds of Proposition 3 on 32,760 cases; containment of Q_f^{p} in climbing with $\theta > \theta_c$ on 792,792 clean and attacked profiles; and that the parking attack drives Q_f^{p} to the root whenever its certified depth is 0 (1,862 cases).

Weighted sources. On 10,647 profiles with integer weights 1 to 3 and every budget $F < W$, enumerating every admissible Z: the characterization of Proposition 8 for all 106,895 balls, and its finest ball-valued output. The union of the honest-majority balls was strictly smaller than $Q_F^w(s)$ on 245 profiles, so the set-valued part of Theorem 1 does not extend to weights. The two counterexamples of Appendix B, for probability reports and for weights, were checked in exact rational arithmetic.

For the vote rules and Q_f^{p} the same functions produce every number in Section 5; for the climbing rules, the reported worst case is an attack lower bound bracketed by the certified upper bound of Proposition 3.

D Experimental details

CIFAR-100. The 100 classes are grouped into the 20 official superclasses ($L = 2$). We hold out 50 training images per class for validation (5,000) and use the remaining 45,000 for training; the official 10,000 test images are used once. The ten sources are linear probes on the frozen ImageNet features of ten torchvision backbones from seven architecture families (ResNet-18, MobileNetV3-Large, ShuffleNetV2, EfficientNet-B0 and B1, RegNetY-1.6GF, RegNetX-1.6GF, DenseNet-121, MNASNet-1.0 and MobileNetV2), with images resized to 160×160. Each probe is a multinomial logistic regression trained by full-batch L-BFGS (60 iterations, zero initialisation), with the weight decay chosen from $\{10^{-4}, 10^{-3}, 10^{-2}\}$ on validation accuracy, followed by temperature scaling on the validation set (Guo et al., 2017). Test leaf accuracies range from 0.68 to 0.74. For the correlated-source control, ten probes are trained on the DenseNet-121 features of ten bootstrap resamples of the training set.

FGVC-Aircraft. Manufacturer $\to$ family $\to$ variant ($L = 3$; 30, 70 and 100 nodes, a strict tree) with the official train, validation and test splits of Maji et al. (2013) (3,334, 3,333 and 3,333 images); the 20-pixel copyright banner is removed and images are resized and center-cropped to 224×224. Sources are built as for CIFAR-100 from the same ten backbones; frozen ImageNet features transfer

poorly to aircraft variants, and test leaf accuracies range from 0.23 to 0.48, which makes this the weak-source setting.

PBMC CITE-seq. The Seurat v4 reference atlas of peripheral blood mononuclear cells (Hao et al., 2021), with paired RNA and 228 surface proteins measured by CITE-seq (Stoeckius et al., 2017), and its three-level annotation l1 $\to$ l2 $\to$ l3 (8 lineages, 30 types, 57 subtypes after removing the "Doublet" label; a strict tree). We split by donor: donors P1 to P5 for training (40,000 cells sampled), P6 for validation (10,000) and P7, P8 for testing (20,000). RNA counts are library-size normalised and log-transformed; the 3,000 most dispersed genes on the training cells are randomly split into five disjoint panels of 600 genes, each reduced to 50 principal components fitted on training cells. Protein counts are transformed by the centred log-ratio, isotype controls are removed, and the remaining 224 proteins are split into five disjoint panels of 44 or 45. Each of the ten sources is a perceptron with two hidden layers ($d \to 256 \to 128 \to 57$, ReLU, dropout 0.2) trained on one panel only with Adam (learning rate 10^{-3}, weight decay 10^{-5}, batch size 512, 25 epochs, fixed seed), followed by temperature scaling.

CIFAR-10H. The 10,000 CIFAR-10 test images with 50 to 63 labels each from 2,571 human annotators (Peterson et al., 2019); attention-check trials are removed. The hierarchy is animal $\to$ {mammal: cat, deer, dog, horse; other animal: bird, frog} and vehicle $\to$ {road vehicle: automobile, truck; other vehicle: airplane, ship} ($L = 3$). The images are split once into a validation and a test half. For each source count M we draw M of the labels of each image without replacement, five times with fixed seeds, and report means. Because each image was labelled by a different subset of annotators, the M sources of an image are M sampled human reports, not M persistent annotators; regime B therefore does not apply. Human reports are 99.3%, 97.8% and 94.9% accurate at the three levels. Climbing on one-hot reports is computed in double precision, so a threshold θ keeps a ball exactly when it holds at least $\lceil \theta M \rceil$ votes.

Single-child nodes and specificity. FGVC-Aircraft has 14 manufacturers with a single family and 54 families with a single variant, and PBMC has 15 types with a single subtype. A ball with a single child contains the same labels as that child, so every rule takes such forced decisions whenever it reaches the parent, and the depth metric credits them to all rules alike. As controls we also measure specificity by $1 - \log|B| / \log|\mathcal{Y}|$ (1 at a leaf, 0 at the root) and by the share of the genuine branch decisions (those at nodes with at least two children) on the path to the true label that the output resolves.

Rules. All rules see the same ten sources. Vote-based rules use each source's most likely label. *Probability averaging* predicts the leaf with the largest mean probability; *majority vote* the most voted leaf; *top-down vote* the most voted child at each depth. *Hierarchical selective classification* (HSC) climbs from the most likely leaf of the mean probability to its deepest ancestor with mass at least θ (Goren et al., 2024); *HSC at* θ_c^+ uses $\theta = \frac{1}{2} + \frac{f}{2M} + 10^{-6}$, a threshold just above θ_c; *hierarchical conformal* uses the same climbing family with θ set by split conformal calibration at level α (Papadopoulos et al., 2002; Vovk et al., 2005; Lei et al., 2018); *trimmed-mean* and *median climbing* replace the mean by the coordinate-wise mean after removing the f largest and f smallest values, or by the coordinate-wise median, and renormalise (Yin et al., 2018); *geometric-median climbing* uses the geometric median of the probability vectors (60 Weiszfeld iterations from the mean) (Dhasade et al., 2026); *vote-share climbing* runs HSC on the mean of the one-hot reports; and *Dawid-Skene climbing* runs HSC on the posterior of a one-coin Dawid-Skene model (Dawid & Skene, 1979) fitted by EM (at most 40 iterations) to the argmax reports of the whole stream it is applied to. A rule *matched* to Q uses the smallest threshold whose clean validation error is at most that of Q. We use the coordinate-wise trimmed mean and median because they are the Byzantine-robust aggregators defined per label, and the geometric median because Dhasade et al. (2026) analyse it; by Proposition 1(i) and (iv), any aggregator followed by climbing is either unsafe on some profile or contains Q_f^{p} everywhere, so no aggregator can be both safe and more specific. Q_f^{p} is the rule of Proposition 1, Q_{2f} the clean-anchored quota of Proposition 2, and the *flat guarantee* outputs a label only if it holds more than $(M + f)/2$ reports and abstains otherwise. Thresholds are set on clean validation data only.

Attacks. For vote-based rules and Q_f^{p} we compute the worst case exactly with the closed forms verified in Appendix C. For the climbing rules the adversary replaces the probability vectors of f sources by one-hot vectors on a target label. We try as targets the 10 most likely wrong labels under the clean aggregate and the most likely label of each of the 5 most likely other top-level groups, and two choices of the corrupted sources (the f sources with the largest $p_m(y) - p_m(\tilde{y})$ for the target $\tilde{y}$, and the f sources with the least mass on the attacked ball); an example counts as an error if any attempt succeeds. This search gives a lower bound; Proposition 3 gives the certified upper bound. The no-answer share under corruption is exact for the quorum rules and the flat guarantee (Corollary 1) and for Q_f^{p} (Appendix B); for the climbing rules it comes from a diversion search that corrupts the f sources most confident in the clean prediction into one-hot reports on one of the five least likely labels, a lower bound. When the same sources are corrupted on every example, the worst case of Q_f is exact (Theorem 3(ii)) and that of HSC is the same search restricted to the fixed set.

Realistic failures. In *collusion*, the failed sources see the honest sources' mean probabilities (not the truth) and all report, as a one-hot vector, the runner-up leaf, the most likely leaf of the runner-up top-level group, or, in the *bandwagon* variant, the honest top leaf. *Coordinated poisoning*: every failed source is retrained on labels in which all training examples of one top-level group (CIFAR-100: trees; FGVC-Aircraft: Tupolev; PBMC: monocytes) are relabelled to fixed classes of another group (medium mammals, Ilyushin, NK cells); it is scored on the examples of the poisoned group. *Independent poisoning*: as above with a different random pair of groups per source. *Label noise*: 60% symmetric label noise. *Sensor failure*: the clean probe receives images with additive Gaussian noise ($\sigma = 0.25$ on $[0, 1]$ pixels) or Gaussian blur. *Panel failure* (PBMC): the panel's measurements are shuffled across cells (scramble) or 70% of its features read zero (dropout). For each f we average over all $\binom{10}{f}$ choices of the failed sources.

Statistics. Certificates use the one-sided 95% Clopper-Pearson bound of Corollary 2 on the validation split. The same split also set the temperatures and, for the image probes, the weight decay; temperatures do not change the argmax labels that the certificate of Q_f uses. For regime B we take the largest bound over the $\binom{10}{f}$ fixed sets at level $0.05/\binom{10}{f}$. The label-free warning compares the contested share $\mathbb{E}[K_f]$ between validation and test data with a two-sided two-proportion z-test; a Holm correction controls the family-wise error over the twelve machine-source tests; the false-alarm rate of each test, and of the Holm-corrected family of the four tests of a dataset, is estimated on 200 random splits of the pooled validation and test data, and its detection rate on PBMC on 200 draws of a 5,000-cell reference sample from the validation donor against 5,000 cells mixing the other half of that donor with a share π of test-donor cells. Overdispersion is fitted per dataset and level by the method of moments: with a_k the mean share of correct reports and v_k the variance of X_k, the intra-class correlation is $\kappa_k = (v_k/(Ma_k(1-a_k)) - 1)/(M-1)$, and the beta-binomial with mean a_k and correlation κ_k gives the fitted share $\Pr[X_k \geq \tau_f + f]$. Paired bootstrap intervals use 2,000 resamples of the test examples. For pool variability, every image probe is retrained on a bootstrap resample of its training set (five seeds), with weight decay and temperature chosen as before; for disjoint selection, the validation split is halved within each class, one half chooses weight decay and temperatures and the other certifies.

Compute. Feature extraction used a laptop GPU through DirectML and a laptop CPU; all aggregation experiments run on cached predictions on the CPU. Computing Q_f takes 4 to 11 μs per example with vectorised NumPy on a laptop CPU, and 13 to 75 μs with a literal pure-Python implementation. On synthetic balanced hierarchies it takes 17 μs per example with $M = 100$ sources and depth 6 (729 labels), and 55 μs with $M = 100$ and depth 8 (6,561 labels); memory is one count per ball and depth.

E Additional results

Matched comparison at other budgets, with top-level errors. Table 6 repeats Table 1 with the worst-case top-level errors and the rows omitted there (safe HSC, median climbing), and Tables 7 and 8 do so for $f = 1$ and $f = 3$. Since the quota is an integer, $Q_1 = Q_0$ and $Q_3 = Q_2$ for $M = 10$, and the clean anchor of $f = 3$ is Q_6, which some corruption of three sources drives to the root on every example ($M \leq 4f$).

Table 6: As Table 1 ($f = 2$), with worst-case top-level errors ("top").

	CIFAR-100					FGVC-Aircraft					PBMC CITE-seq				
Rule	clean	depth	worst	top	none	clean	depth	worst	top	none	clean	depth	worst	top	none
probability reports; HSC matched to Q_f^{p}															
Q_f^{p} (ours)	3.3	0.68	9.4	5.6	41.4	0.7	0.17	4.0	3.2	94.2	2.2	0.77	9.9	3.7	20.7
HSC, matched	3.4	0.69	11.7	7.0	39.9	0.6	0.17	6.5	5.4	93.5	2.5	0.78	13.9	4.7	19.6
HSC, $\theta = \theta_c^+$ (safe)	2.7	0.66	9.4	5.6	43.1	0.3	0.15	4.0	3.2	94.9	1.6	0.75	9.9	3.7	23.8
label reports; climbing on vote shares matched to Q_f															
Q_f (ours)	5.8	0.76	14.6	8.4	32.8	4.5	0.39	18.6	12.5	75.9	5.1	0.81	14.4	3.6	21.0
vote shares, matched	5.8	0.76	14.6	8.4	32.9	4.4	0.39	18.6	12.5	77.3	5.1	0.81	14.4	3.6	21.7
Q_{2f}, clean anchor	3.5	0.69	9.5	5.4	45.3	2.0	0.29	9.5	6.2	88.5	2.6	0.74	8.9	1.9	38.0
flat guarantee	4.7	0.70	12.2	6.0	48.1	2.3	0.30	9.5	4.1	85.6	3.9	0.67	10.9	1.5	51.2
averaged probabilities, matched to Q_f															
HSC	5.6	0.76	20.1	11.9	31.5	4.6	0.33	38.5	33.0	83.6	5.4	0.83	25.2	7.1	15.5
trimmed mean + climb	5.8	0.78	18.5	11.2	20.3	4.4	0.36	24.8	18.6	61.7	6.2	0.85	22.4	6.7	5.4
median + climb	5.8	0.78	17.6	10.7	20.1	4.3	0.36	22.2	16.0	62.4	6.8	0.86	22.7	6.8	4.4

Table 7: As Table 6, $f = 1$.

	CIFAR-100					FGVC-Aircraft					PBMC CITE-seq				
Rule	clean	depth	worst	top	none	clean	depth	worst	top	none	clean	depth	worst	top	none
probability reports; HSC matched to Q_f^{p}															
Q_f^{p} (ours)	4.3	0.72	7.3	4.2	27.7	1.0	0.19	2.6	2.1	84.0	3.2	0.79	7.0	2.6	8.5
HSC, matched	4.2	0.72	8.0	4.5	27.1	1.1	0.20	3.5	2.9	83.3	3.3	0.79	8.2	3.0	7.7
HSC, $\theta = \theta_c^+$ (safe)	3.9	0.71	7.2	4.2	28.4	0.8	0.18	2.6	2.1	85.7	2.7	0.78	6.9	2.5	8.5
label reports; climbing on vote shares matched to Q_f															
Q_f (ours)	9.5	0.84	14.6	8.4	17.5	9.5	0.50	18.6	12.5	50.8	8.9	0.87	14.4	3.6	7.2
vote shares, matched	9.3	0.84	14.5	8.3	17.7	9.2	0.50	18.4	12.2	51.7	8.8	0.87	14.4	3.5	7.5
Q_{2f}, clean anchor	5.8	0.76	9.5	5.4	24.2	4.5	0.39	9.5	6.2	62.8	5.1	0.81	8.9	1.9	12.0
flat guarantee	7.8	0.79	12.2	6.0	29.7	5.0	0.39	9.5	4.1	70.4	6.5	0.76	10.9	1.5	32.9
averaged probabilities, matched to Q_f															
HSC	9.7	0.85	17.9	10.4	16.1	8.9	0.44	33.3	28.1	58.4	9.0	0.88	19.8	5.6	4.1
trimmed mean + climb	9.6	0.85	16.4	9.5	12.7	9.5	0.47	23.2	17.9	43.2	9.1	0.89	17.5	5.1	3.3
median + climb	9.8	0.86	16.0	9.2	12.1	9.8	0.48	20.8	14.9	42.5	9.9	0.90	17.8	5.1	2.7

Table 8: As Table 6, $f = 3$.

	CIFAR-100					FGVC-Aircraft					PBMC CITE-seq				
Rule	clean	depth	worst	top	none	clean	depth	worst	top	none	clean	depth	worst	top	none
probability reports; HSC matched to Q_f^{p}															
Q_f^{p} (ours)	2.6	0.65	12.2	7.4	66.5	0.4	0.14	5.5	4.3	99.1	1.6	0.74	13.0	4.8	51.7
HSC, matched	2.6	0.66	16.5	9.8	60.8	0.4	0.16	14.9	12.7	97.5	2.0	0.77	24.0	7.4	49.4
HSC, $\theta = \theta_c^+$ (safe)	2.0	0.62	12.2	7.4	70.1	0.1	0.10	5.5	4.3	99.2	0.9	0.72	13.0	4.8	59.3
label reports; climbing on vote shares matched to Q_f															
Q_f (ours)	5.8	0.76	21.8	12.9	45.3	4.5	0.39	33.5	22.2	88.5	5.1	0.81	22.2	5.8	38.0
vote shares, matched	5.8	0.76	21.8	12.9	47.5	4.4	0.39	33.5	22.2	89.4	5.1	0.81	22.2	5.8	42.1
Q_{2f}, clean anchor	2.0	0.60	9.5	5.4	100.0	0.6	0.19	9.5	6.2	100.0	1.1	0.64	8.9	1.9	100.0
flat guarantee	4.7	0.70	18.6	9.7	61.0	2.3	0.30	18.9	8.5	92.5	3.9	0.67	18.1	3.0	62.9
averaged probabilities, matched to Q_f															
HSC	5.6	0.76	36.5	22.7	46.2	4.6	0.33	76.3	70.2	92.0	5.4	0.83	45.3	15.2	42.1
trimmed mean + climb	5.9	0.78	31.5	20.2	21.9	4.4	0.36	51.2	38.5	65.9	6.6	0.85	36.5	11.9	5.1
median + climb	5.8	0.78	29.7	18.9	21.9	4.3	0.36	46.9	34.7	66.5	6.8	0.86	35.8	11.5	4.5

Same report model. HSC matched to Q_f^{p} lands at $\theta = 0.572$, 0.565 and 0.557 for $f = 2$ and 0.607, 0.585 and 0.576 for $f = 3$, always below θ_c, with clean error within 0.5 points and depth within 0.03 of Q_f^{p} on the test data; its certified upper bound equals its attack value to within 0.01 points, so its worst case is exact. Climbing on vote shares matched to Q_f lands just above 0.6 on every dataset for $f = 2, 3$ (just above 0.5 for $f = 1$); it then contains Q_f (its output lies on the majority chain, whose deepest element is Q_f), nearly coincides with it and has the same worst case, and it is safe, since one-hot climbing is f-safe exactly when $\theta > \lfloor (M + f)/2 \rfloor / M$.

Safe climbing and overclaims. Table 9 compares HSC at the safe threshold $\theta_c^+ = \theta_c + 10^{-6}$ with Q_f^{p}. Its certified upper bound equals the exact worst case of Q_f^{p} to the reported precision in all nine cases (Proposition 3 with $f' = f$), its attack lower bound is within 0.1 points of it, and it is slightly coarser, as Proposition 1(iv) requires. The last two columns give the share of test examples on which the output does not contain Q_f^{p}, for HSC calibrated to the clean error of Q_f and for the vote rule Q_f itself. The rates are similar: Q_f is safe for label reports, not for probability reports, so this yardstick separates the report models rather than the two rules, whose error under attack differs (Table 1). Table 10 lists the other numbers that Section 5 quotes without a table.

Table 9: HSC at $\theta = \theta_c^+$ against Q_f^{p}: clean error (%), clean depth and worst-case error (%; HSC: attack lower bound with the certified upper bound in parentheses; Q_f^{p}: exact), and the share of test examples (%) on which the output does not contain Q_f^{p}, for HSC calibrated to the clean error of Q_f (overclaims) and for Q_f.

		HSC at θ_c^+			Q_f^{p}			not containing Q_f^{p}	
Data	f	clean	depth	worst (bound)	clean	depth	worst	HSC cal.	Q_f
CIFAR-100	1	3.9	0.71	7.2 (7.3)	4.3	0.72	7.3	21.2	18.7
	2	2.7	0.66	9.4 (9.4)	3.3	0.68	9.4	13.8	13.6
	3	2.0	0.62	12.2 (12.2)	2.6	0.65	12.2	18.9	18.0
FGVC-Aircraft	1	0.8	0.18	2.6 (2.6)	1.0	0.19	2.6	45.3	45.9
	2	0.3	0.15	4.0 (4.0)	0.7	0.17	4.0	31.7	33.2
	3	0.1	0.10	5.5 (5.5)	0.4	0.14	5.5	36.1	37.0
PBMC CITE-seq	1	2.7	0.78	6.9 (7.0)	3.2	0.79	7.0	22.0	21.0
	2	1.6	0.75	9.9 (9.9)	2.2	0.77	9.9	14.7	14.0
	3	0.9	0.72	13.0 (13.0)	1.6	0.74	13.0	19.8	17.8

Table 10: Numbers quoted in the text that no other table shows, recomputed from the result files ($f = 2$ unless noted).

Quantity	CIFAR-100	FGVC-Aircraft	PBMC CITE-seq
Q_0 stops above the leaves (% of examples)	20.9	61.0	24.4
probability-averaging error there / elsewhere (%)	52.1 / 9.7	64.0 / 12.4	44.6 / 8.6
top-level answer kept: true / permuted / no hierarchy (%)	67.2 / 52.9 / 51.8	24.1 / 16.3 / 14.4	79.0 / 52.3 / 48.8
safe climbing (θ_c^+): clean error (%)	2.7	0.3	1.6
95% conformal climbing: θ / clean error (%)	0.54 / 4.0	0.40 / 5.3	0.54 / 3.2
95% conformal climbing: worst case / overclaims (%)	14.2 / 5.8	44.1 / 34.6	16.6 / 6.1
95% conformal climbing, $f = 2$: worst case (certified bound) (%)	14.2 (14.2)	44.1 (44.5)	16.6 (16.7)
95% conformal climbing, $f = 3$: worst case (certified bound) (%)	25.3 (25.3)	82.4 (83.3)	33.1 (33.1)
at the clean depth of Q_2: worst case of the next best family / of HSC, relative to Q_2	1.09 / 1.41	1.36 / 2.91	1.07 / 1.45
HSC calibrated to the clean validation error of Q_2: θ	0.49	0.42	0.48
top-down vote: worst case (%)	33.7	69.0	35.0
collusion, $f = 3$: HSC/Q_3 at equal attacked depth, runner-up / bandwagon	1.45 / 1.01	1.95 / 1.19	1.26 / 0.99
one-backbone control, worst case of Q_2 / HSC (%)	25.0 / 34.5	n/a	n/a
other T cells: top-level worst case, validation / test (%); share of the gap	n/a	n/a	4.7 / 50.7; 73%

Upper bounds for the climbing rules. Table 11 brackets the worst-case error of the calibrated climbing rules between the attack of Appendix D and the certified upper bound of Proposition 3.

For HSC and conformal climbing the two differ by at most 0.5 points when $f \geq 2$, so their reported worst case is essentially exact; for the trimmed mean and the median the bound is up to 2.5 times their attack value, so their worst case cannot be certified.

Table 11: Worst-case error (%) of the climbing rules calibrated to the clean validation error of Q_f (conformal: split conformal at that level, not at 95% coverage): attack lower bound ("search") and certified upper bound ("bound").

		CIFAR-100		FGVC-Aircraft		PBMC CITE-seq	
f	Rule	search	bound	search	bound	search	bound
1	HSC	17.9	20.5	33.3	36.6	19.8	22.0
	conformal	17.9	20.5	33.3	36.6	19.8	22.0
	trimmed + climb	16.4	23.3	23.2	51.2	17.5	24.7
	median + climb	16.0	18.9	20.8	36.5	17.8	20.8
2	HSC	20.1	20.3	38.5	38.9	25.2	25.4
	conformal	19.9	20.1	38.4	38.8	25.1	25.4
	trimmed + climb	18.5	26.7	24.8	61.5	22.4	30.3
	median + climb	17.6	22.4	22.2	49.9	22.7	27.3
3	HSC	36.5	36.7	76.3	76.7	45.3	45.6
	conformal	36.1	36.2	76.1	76.5	45.2	45.5
	trimmed + climb	31.5	38.4	51.2	82.5	36.5	42.2
	median + climb	29.7	35.8	46.9	79.2	35.8	40.2

Geometric-median climbing. The geometric median of the probability vectors (Weiszfeld iterations) is a robust averaging rule analysed for federated inference by Dhasade et al. (2026). Matched to Q_f, its attack lower bound is 0.97 to 1.33 times the exact worst case of Q_f (Table 12). We have no certified bound for it, and the attack of Appendix D was designed for mean-like aggregates, so these values may understate its worst case.

Table 12: Climbing on the coordinate-free geometric median, matched to Q_f: threshold, clean error (%), clean depth, attack lower bound on the worst-case error and on the worst-case top-level error (%), and the exact worst case of Q_f (%).

Data	f	θ	clean	depth	worst (attack)	top (attack)	worst of Q_f
CIFAR-100	2	0.51	5.7	0.76	15.5	8.9	14.6
	3	0.51	5.7	0.76	25.3	15.1	21.8
FGVC-Aircraft	2	0.41	4.9	0.34	18.0	14.6	18.6
	3	0.41	4.9	0.34	34.7	27.8	33.5
PBMC CITE-seq	2	0.52	5.7	0.83	17.4	6.1	14.4
	3	0.52	5.7	0.83	29.5	10.8	22.2

Learning reliability under persistent corruption. One-coin Dawid-Skene (Dawid & Skene, 1979) estimates an accuracy per source and a class prior by EM on the whole test stream of argmax reports; we climb on its posterior over leaves, with the threshold set on clean validation data to the clean validation error of Q_f. For every set Z of f sources corrupted on every example we run two attacks, each followed by re-estimation on the attacked stream: *naive*, where Z reports the most-voted wrong leaf on every example, and *stealthy*, where Z reports its clean labels except on the examples where attacking makes the output exclude the true label under the current estimates (four rounds of re-estimation and re-targeting). Both are attack lower bounds. The oracle outputs the honest-majority ball of the sources outside Z; by Theorem 3(ii) its error equals the exact worst case of Q_f for that set, which Table 13 confirms. Dawid-Skene learns to discount a naive persistent pair, and then beats Q_f on CIFAR-100 and PBMC, but not a stealthy one on CIFAR-100, and with the weak aircraft sources both attacks drive it far above Q_f (39.9% against 18.3% at $f = 3$). On PBMC it lands coarser than Q_f (depth 0.70 against 0.81), which explains its lower error there.

Table 13: Persistent corruption against a rule that learns source reliability. Clean error (%) / depth of Dawid-Skene with climbing and of Q_f; under a fixed corrupted set, the worst case over all $\binom{10}{f}$ sets (mean over sets in parentheses) of the exact worst-case error of Q_f, of Dawid-Skene under the naive and the stealthy attack (attack lower bounds), and of the oracle that knows the set.

		clean		persistent set: worst over sets (mean)			
Data	f	DS	Q_f	Q_f exact	DS, naive	DS, stealthy	oracle
CIFAR-100	2	6.3 / 0.78	5.8 / 0.76	9.9 (9.4)	7.9 (7.5)	11.8 (11.1)	9.9 (9.4)
	3	6.3 / 0.78	5.8 / 0.76	13.6 (12.8)	9.6 (8.7)	16.6 (15.7)	13.6 (12.8)
FGVC-Aircraft	2	4.5 / 0.41	4.5 / 0.39	11.1 (9.9)	16.3 (13.8)	14.8 (13.7)	11.1 (9.9)
	3	4.5 / 0.41	4.5 / 0.39	18.3 (15.9)	39.9 (32.1)	34.7 (30.8)	18.3 (15.9)
PBMC CITE-seq	2	4.1 / 0.70	5.1 / 0.81	9.6 (8.9)	5.3 (4.9)	8.0 (7.3)	9.6 (8.9)
	3	4.1 / 0.70	5.1 / 0.81	14.1 (12.5)	7.0 (5.9)	12.4 (10.7)	14.1 (12.5)

Availability under a persistent set. By Theorem 2(ii), a fixed corrupted set Z can drive Q_f to the root on an example exactly when no top-level ball keeps more than $(M + f)/2$ of the reports of the other sources. Table 14 gives this share for the worst, the average and the best fixed set, next to the per-example oracle share of Corollary 1: persistence lowers the no-answer share by at most 3.1 points for the worst set (2.5 at $f = 2$).

Table 14: Share of test examples (%) that some corruption drives Q_f to the root: per-example oracle (regime A) and a fixed corrupted set (regime B, exact) for the worst, the average and the best of the $\binom{10}{f}$ sets.

Data, f	per-example oracle	worst fixed set	mean over fixed sets	best fixed set
CIFAR-100, 2	32.8	30.3	29.1	27.9
CIFAR-100, 3	45.3	42.3	40.9	39.5
FGVC-Aircraft, 2	75.9	74.1	70.2	62.3
FGVC-Aircraft, 3	88.5	87.4	83.8	73.6
PBMC CITE-seq, 2	21.0	19.9	17.3	14.3
PBMC CITE-seq, 3	38.0	36.2	32.3	26.5

Uncertainty. The sources are fixed (L-BFGS from zero initialisation, fixed seeds), so the variability of the headline numbers comes from the finite test sets and from the choice of sources; all uncertainty below is conditional on the trained source pool. Table 16 gives paired bootstrap confidence intervals over test examples for the differences in worst-case error between HSC matched to Q_f^{p} and Q_f^{p}, and between the baselines matched to Q_f and Q_f; every interval excludes zero. Table 17 repeats the matched comparison on five random subsets of $M' \in \{5, 7, 9\}$ of the ten sources, including odd source counts; the ratio of the worst-case errors of HSC and Q_f stays between 1.16 and 2.31 and varies by at most 0.27 across subsets of the same size. To vary the pool itself, we retrain all ten probes of each image dataset on five bootstrap resamples of its training set, with the same backbones and protocol (Table 15). In every pool, HSC matched to Q_f^{p} crosses θ_c and has 1.19 to 1.42 times the worst case of Q_f^{p} on CIFAR-100 and 1.60 to 4.30 times on FGVC-Aircraft, and 95% coverage sets θ below θ_c for $f = 2, 3$ (at $f = 1$, where $\theta_c = 0.55$, three of the five CIFAR-100 pools land just above it); the PBMC networks were not retrained. Choosing weight decay and temperatures on one half of the validation split and certifying Q_f on the other changes the certificate by at most 0.7 points (at $f = 2$: 17.0% against 16.8% on CIFAR-100, 19.6% against 20.3% on FGVC-Aircraft, both certified on the same held-out half), and every bound still covers the exact test worst case.

Table 15: Five bootstrap-retrained source pools per image dataset (mean $\pm$ standard deviation over pools): ratio of the worst-case error of HSC matched to Q_f^{p} to that of Q_f^{p}, and of HSC matched to Q_f to that of Q_f; the threshold matched to Q_f^{p} ($\theta_c = 0.6$ and 0.65 for $f = 2$ and 3); and the 95% split-conformal threshold.

Data	f	HSC/Q_f^{p}	HSC/Q_f	θ at Q_f^{p}	95% conformal θ
CIFAR-100	2	1.20 ± 0.01	1.40 ± 0.04	0.574 ± 0.002	0.555 ± 0.007
	3	1.37 ± 0.04	1.66 ± 0.04	0.608 ± 0.004	0.555 ± 0.007
FGVC-Aircraft	2	2.03 ± 0.37	2.19 ± 0.17	0.561 ± 0.013	0.384 ± 0.007
	3	3.65 ± 0.39	2.40 ± 0.11	0.578 ± 0.004	0.384 ± 0.007

Table 16: Worst-case error (%) of a safe rule (Q_f^{p} in the first two rows of each dataset, Q_f otherwise) and of a baseline matched to it (HSC, or HSC and the trimmed mean matched to Q_f), and their difference with a paired bootstrap 95% confidence interval over test examples. Collusion errors are averaged over all $\binom{10}{3}$ choices of the colluders before resampling.

Data	Comparison	safe rule	baseline	difference (95% CI)
CIFAR-100	worst, $f = 2$, Q_f^{p} vs HSC matched to it	9.4	11.7	+2.2 [1.9, 2.5]
	worst, $f = 3$, Q_f^{p} vs HSC matched to it	12.2	16.5	+4.2 [3.8, 4.6]
	worst, $f = 2$, vs HSC	14.6	20.1	+5.5 [5.0, 6.1]
	worst, $f = 2$, vs trimmed + climb	14.6	18.5	+3.9 [3.4, 4.4]
	worst, $f = 3$, vs HSC	21.8	36.5	+14.8 [14.1, 15.5]
	worst, $f = 3$, vs trimmed + climb	21.8	31.5	+9.7 [9.1, 10.3]
	collusion, $f = 3$, vs HSC	3.7	13.2	+9.5 [9.1, 9.9]
FGVC-Aircraft	worst, $f = 2$, Q_f^{p} vs HSC matched to it	4.0	6.5	+2.6 [2.0, 3.1]
	worst, $f = 3$, Q_f^{p} vs HSC matched to it	5.5	14.9	+9.4 [8.4, 10.4]
	worst, $f = 2$, vs HSC	18.6	38.5	+19.9 [18.5, 21.5]
	worst, $f = 2$, vs trimmed + climb	18.6	24.8	+6.2 [4.8, 7.5]
	worst, $f = 3$, vs HSC	33.5	76.3	+42.8 [41.1, 44.4]
	worst, $f = 3$, vs trimmed + climb	33.5	51.2	+17.7 [16.1, 19.4]
	collusion, $f = 3$, vs HSC	5.3	26.9	+21.5 [20.6, 22.4]
PBMC CITE-seq	worst, $f = 2$, Q_f^{p} vs HSC matched to it	9.9	13.9	+4.0 [3.8, 4.3]
	worst, $f = 3$, Q_f^{p} vs HSC matched to it	13.0	24.0	+11.0 [10.5, 11.4]
	worst, $f = 2$, vs HSC	14.4	25.2	+10.7 [10.3, 11.2]
	worst, $f = 2$, vs trimmed + climb	14.4	22.4	+8.0 [7.6, 8.4]
	worst, $f = 3$, vs HSC	22.2	45.3	+23.1 [22.5, 23.7]
	worst, $f = 3$, vs trimmed + climb	22.2	36.5	+14.3 [13.8, 14.8]
	collusion, $f = 3$, vs HSC	3.8	18.0	+14.2 [13.8, 14.5]

Table 17: Matched comparison on random subsets of M' sources (five subsets per M'): mean worst-case error (%) of Q_f and of HSC calibrated to the clean validation error of Q_f, and the mean and range of their ratio.

Data	M'	f	Q_f worst	HSC worst (matched)	ratio (range)
CIFAR-100	5	1	13.8	18.0	1.30 (1.27 to 1.34)
	7	1	12.8	15.9	1.24 (1.19 to 1.29)
	7	2	22.4	36.2	1.62 (1.57 to 1.67)
	9	1	12.0	14.3	1.20 (1.16 to 1.24)
	9	2	18.9	27.7	1.47 (1.43 to 1.50)
FGVC-Aircraft	5	1	19.4	39.0	2.01 (1.87 to 2.14)
	7	1	15.8	29.8	1.88 (1.76 to 1.99)
	7	2	35.7	80.8	2.27 (2.22 to 2.31)
	9	1	13.9	23.4	1.68 (1.58 to 1.73)
	9	2	27.6	59.3	2.15 (2.12 to 2.20)
PBMC CITE-seq	5	1	13.9	23.0	1.65 (1.60 to 1.76)
	7	1	12.5	18.7	1.50 (1.45 to 1.57)
	7	2	23.0	45.7	1.99 (1.95 to 2.02)
	9	1	11.6	16.3	1.40 (1.37 to 1.43)
	9	2	19.1	34.6	1.81 (1.77 to 1.84)

The certificate out of sample, and the label-free warning. Table 18 extends Table 3 to every $f \leq 4$ and to Q_f^{p}. The bound covers the test worst case in every case except the top-level errors on PBMC (bold), where the test donors differ from the validation donor (Section 5.4); three quarters of the gap come from other T cells, whose share is nearly unchanged (3.2% against 3.8%) but whose worst-case lineage error rises from 5% to 51%. Table 19 gives the label-free contested share of Theorem 3(iii) on both splits: its change is significant exactly on PBMC, at the top level for $f = 2$ and 3 and for all errors at $f = 3$ (with $M = 10$, $f = 1$ and $f = 4$ give the same indicators as $f = 2$ and $f = 3$), and a Holm correction over the twelve machine-source tests keeps exactly these three. Table 20 calibrates the test: over 200 random splits of the pooled validation and test data, which are exchangeable by construction, it fires at the nominal rate, and on PBMC its detection rate grows with the share of cells from the shifted donors. The warning is computed from the reports and cannot say why they changed: a shift of the data, a changed or recalibrated source pool and corrupted sources all move the contested share, so it is meant for uncorrupted deployment reports, not for detecting attacks. It warns of a shift but does not repair the certificate: the contested share is only a lower bound on the new risk (0.6% against a test worst case of 3.6% at $f = 2$), so when it fires the certificate must be re-estimated on labelled data from the new population, here the test donors. Under persistent corruption the exact worst case of Q_f^{p} over fixed sets is 6.0%, 2.5% and 6.2% at $f = 2$ and 6.4%, 2.9% and 7.0% at $f = 3$; the trimmed mean and the median were not evaluated in this regime.

Table 18: Certificate estimated on validation data, its one-sided 95% Clopper-Pearson bound, and the exact worst case on test data (%), for all errors and for top-level errors. Bold: test value above the bound.

			all errors			top-level errors		
Data	Rule	f	valid.	bound	test	valid.	bound	test
CIFAR-100	Q_f	1	15.3	16.1	14.6	8.4	9.1	8.4
		2	15.3	16.1	14.6	8.4	9.1	8.4
		3	21.7	22.6	21.8	12.4	13.2	12.9
		4	21.7	22.6	21.8	12.4	13.2	12.9
	Q_f^{p}	1	8.9	9.5	7.3	4.9	5.4	4.2
		2	10.7	11.5	9.4	5.9	6.5	5.6
		3	13.2	14.0	12.2	7.4	8.0	7.4
		4	15.9	16.8	15.2	9.0	9.7	9.1
FGVC-Aircraft	Q_f	1	18.0	19.1	18.6	12.6	13.6	12.5
		2	18.0	19.1	18.6	12.6	13.6	12.5
		3	32.5	33.8	33.5	22.7	23.9	22.2
		4	32.5	33.8	33.5	22.7	23.9	22.2
	Q_f^{p}	1	2.7	3.2	2.6	2.3	2.7	2.1
		2	3.8	4.3	4.0	3.1	3.7	3.2
		3	5.4	6.1	5.5	4.5	5.1	4.3
		4	7.7	8.5	7.1	6.3	7.1	5.7
PBMC CITE-seq	Q_f	1	16.1	16.8	14.4	1.1	1.3	**3.6**
		2	16.1	16.8	14.4	1.1	1.3	**3.6**
		3	21.6	22.3	22.2	1.8	2.0	**5.8**
		4	21.6	22.3	22.2	1.8	2.0	**5.8**
	Q_f^{p}	1	9.8	10.3	7.0	0.8	0.9	**2.6**
		2	12.2	12.7	9.9	1.1	1.3	**3.7**
		3	14.9	15.5	13.0	1.4	1.6	**4.8**
		4	17.8	18.5	16.9	2.0	2.2	**6.1**

Table 19: Label-free warning. Certificate bound and exact test worst case of Q_f (%, as in Table 3), whether the bound covers the test value (CIFAR-10H: number of the five draws), and the contested share (%) on validation and test data with the two-sided two-proportion p-value (CIFAR-10H: mean over draws). It needs no test labels.

			certificate			contested share (no labels)		
Data	f	errors	bound	test	transfers	valid.	test	p
CIFAR-100	2	all	16.1	14.6	yes	1.5	1.7	0.36
		top level	9.1	8.4	yes	0.9	1.0	0.60
	3	all	22.6	21.8	yes	8.2	8.5	0.56
		top level	13.2	12.9	yes	4.8	5.2	0.29
FGVC-Aircraft	2	all	19.1	18.6	yes	0.8	1.0	0.60
		top level	13.6	12.5	yes	0.8	0.7	0.78
	3	all	33.8	33.5	yes	8.9	9.5	0.42
		top level	23.9	22.2	yes	7.6	7.4	0.82
PBMC CITE-seq	2	all	16.8	14.4	yes	1.3	1.4	0.67
		top level	1.3	3.6	**no**	0.3	0.6	$2 \cdot 10^{-4}$
	3	all	22.3	22.2	yes	7.4	9.7	$7 \cdot 10^{-11}$
		top level	2.0	5.8	**no**	1.2	3.5	$4 \cdot 10^{-30}$
CIFAR-10H	2	all	1.7	1.2	5/5	0.5	0.4	0.35
		top level	0.2	0.1	4/5	0.1	0.1	0.68
	3	all	2.7	2.1	5/5	1.8	1.6	0.33
		top level	0.3	0.2	4/5	0.1	0.2	0.31

Table 20: Calibration of the label-free warning (two-sided two-proportion test at level 5%, 200 draws). Top: false-alarm rate (%) on random splits of the pooled validation and test data, per test and for any of the four tests with a Holm correction. Bottom: PBMC detection rate (%) with a reference sample of 5,000 validation-donor cells and a deployment sample of 5,000 cells, a share π of which come from the test donors.

Data	$f = 2$, all	$f = 2$, top	$f = 3$, all	$f = 3$, top	any of four (Holm)
CIFAR-100	6.0	3.5	5.5	4.0	2.0
FGVC-Aircraft	4.0	4.5	4.0	7.0	4.0
PBMC CITE-seq	8.0	4.5	6.0	4.0	3.5
PBMC detection rate, share π of cells from the test donors					
$\pi = 0.00$	5.0	6.0	4.5	6.0	4.0
$\pi = 0.10$	3.0	8.5	9.5	18.5	10.0
$\pi = 0.20$	3.5	7.5	14.5	53.5	33.0
$\pi = 0.30$	3.0	9.5	21.0	75.0	57.0
$\pi = 0.50$	3.0	27.0	55.5	100.0	99.0
$\pi = 0.75$	1.5	47.0	88.0	100.0	100.0
$\pi = 1.00$	1.5	72.5	100.0	100.0	100.0

All realistic failures. Table 21 reports every non-adaptive failure for $f = 2$ and $f = 3$ and Table 22 repeats Table 2 for $f = 2$. Across the failed-source subsets the standard deviation of the error of Q_f and HSC is at most 1.3 points.

Table 21: Non-adaptive failures, error (%) / depth, averaged over all choices of the f failed sources; poisoning is scored on the examples of the poisoned group.

Data	Failure	f	Q_f	Q_f^{p}	HSC	trimmed	prob. avg.	majority	top-down
CIFAR-100	coordinated poisoning	2	8.1 / 0.66	2.7 / 0.52	4.8 / 0.62	8.1 / 0.70	30.9 / 1.00	32.0 / 1.00	32.2 / 1.00
		3	6.2 / 0.54	1.2 / 0.30	3.8 / 0.55	8.8 / 0.69	32.0 / 1.00	33.0 / 1.00	32.9 / 1.00
	independent poisoning	2	5.7 / 0.76	3.0 / 0.66	5.2 / 0.75	5.4 / 0.77	18.6 / 1.00	19.6 / 1.00	19.8 / 1.00
		3	5.7 / 0.76	2.2 / 0.62	4.9 / 0.74	5.3 / 0.76	18.6 / 1.00	19.6 / 1.00	19.8 / 1.00
	label noise 60%	2	4.8 / 0.72	2.3 / 0.62	4.4 / 0.72	5.6 / 0.77	18.8 / 1.00	20.0 / 1.00	20.2 / 1.00
		3	4.3 / 0.70	1.4 / 0.55	3.8 / 0.69	5.5 / 0.76	18.9 / 1.00	20.2 / 1.00	20.5 / 1.00
	Gaussian noise	2	3.1 / 0.65	1.4 / 0.55	3.1 / 0.67	5.8 / 0.77	19.0 / 1.00	20.3 / 1.00	20.5 / 1.00
		3	1.8 / 0.54	0.3 / 0.35	2.0 / 0.60	5.7 / 0.76	19.5 / 1.00	21.1 / 1.00	21.1 / 1.00
	blur	2	5.0 / 0.71	2.6 / 0.62	4.7 / 0.71	5.5 / 0.75	19.4 / 1.00	20.7 / 1.00	21.0 / 1.00
		3	4.6 / 0.67	1.7 / 0.53	4.3 / 0.68	5.4 / 0.74	20.1 / 1.00	21.5 / 1.00	21.8 / 1.00
FGVC-Aircraft	coordinated poisoning	2	4.9 / 0.13	0.0 / 0.01	4.5 / 0.07	5.6 / 0.14	40.0 / 1.00	45.5 / 1.00	52.0 / 1.00
		3	4.1 / 0.07	0.0 / 0.00	4.4 / 0.05	5.4 / 0.11	48.9 / 1.00	53.6 / 1.00	56.7 / 1.00
	independent poisoning	2	4.3 / 0.38	0.6 / 0.16	4.3 / 0.32	4.2 / 0.35	44.0 / 1.00	46.6 / 1.00	47.6 / 1.00
		3	4.2 / 0.38	0.3 / 0.12	4.2 / 0.31	4.1 / 0.35	44.0 / 1.00	46.6 / 1.00	47.6 / 1.00
	label noise 60%	2	2.9 / 0.30	0.1 / 0.08	2.6 / 0.25	2.8 / 0.28	44.8 / 1.00	47.4 / 1.00	48.7 / 1.00
		3	2.2 / 0.24	0.0 / 0.03	1.8 / 0.20	1.9 / 0.23	45.2 / 1.00	48.2 / 1.00	49.6 / 1.00
	Gaussian noise	2	1.9 / 0.25	0.1 / 0.08	2.4 / 0.23	4.1 / 0.34	47.2 / 1.00	47.5 / 1.00	48.7 / 1.00
		3	1.0 / 0.16	0.0 / 0.03	1.7 / 0.18	4.0 / 0.33	49.8 / 1.00	48.4 / 1.00	50.0 / 1.00
	blur	2	2.7 / 0.27	0.3 / 0.10	3.5 / 0.25	4.5 / 0.34	47.2 / 1.00	47.9 / 1.00	49.4 / 1.00
		3	2.0 / 0.20	0.1 / 0.05	3.2 / 0.21	4.5 / 0.33	49.4 / 1.00	49.0 / 1.00	50.9 / 1.00
PBMC CITE-seq	coordinated poisoning	2	0.1 / 0.97	0.1 / 0.97	0.3 / 0.99	0.6 / 1.00	0.8 / 1.00	0.9 / 1.00	1.0 / 1.00
		3	0.0 / 0.94	0.0 / 0.91	0.1 / 0.98	0.7 / 1.00	0.8 / 1.00	0.9 / 1.00	0.9 / 1.00
	panel scramble	2	2.3 / 0.70	0.7 / 0.65	2.5 / 0.76	6.9 / 0.85	18.5 / 1.00	20.4 / 1.00	20.2 / 1.00
		3	1.2 / 0.59	0.2 / 0.47	1.7 / 0.71	7.9 / 0.87	20.1 / 1.00	21.2 / 1.00	21.0 / 1.00
	panel dropout	2	2.9 / 0.73	0.8 / 0.68	2.9 / 0.78	5.7 / 0.83	17.9 / 1.00	20.2 / 1.00	19.9 / 1.00
		3	1.9 / 0.67	0.3 / 0.55	1.9 / 0.75	5.6 / 0.83	18.3 / 1.00	20.8 / 1.00	20.5 / 1.00

Table 22: As Table 2, $f = 2$ failed sources out of 10 (all 45 choices).

Data	Failure	Q_f	Q_f^{P}	HSC	trimmed	prob. avg.	majority
CIFAR-100	collusion, runner-up label	3.8 / 0.66	2.0 / 0.57	6.3 / 0.70	5.5 / 0.73	26.5 / 1.00	23.4 / 1.00
	collusion, runner-up group	3.0 / 0.64	1.2 / 0.54	4.0 / 0.66	5.3 / 0.74	24.5 / 1.00	22.4 / 1.00
	collusion, bandwagon	8.8 / 0.83	5.5 / 0.75	11.4 / 0.87	12.2 / 0.90	18.9 / 1.00	19.0 / 1.00
	coordinated poisoning	8.1 / 0.66	2.7 / 0.52	4.8 / 0.62	8.1 / 0.70	30.9 / 1.00	32.0 / 1.00
FGVC-Aircraft	collusion, runner-up label	4.0 / 0.30	1.1 / 0.13	11.2 / 0.34	7.2 / 0.37	62.8 / 1.00	51.4 / 1.00
	collusion, runner-up group	2.3 / 0.25	0.0 / 0.06	7.9 / 0.22	5.0 / 0.33	63.6 / 1.00	51.3 / 1.00
	collusion, bandwagon	8.9 / 0.48	1.7 / 0.23	14.3 / 0.54	14.3 / 0.56	44.8 / 1.00	45.0 / 1.00
	coordinated poisoning	4.9 / 0.13	0.0 / 0.01	4.5 / 0.07	5.6 / 0.14	40.0 / 1.00	45.5 / 1.00
PBMC CITE-seq	collusion, runner-up label	3.5 / 0.74	1.9 / 0.69	7.2 / 0.80	6.7 / 0.82	29.4 / 1.00	23.8 / 1.00
	collusion, runner-up group	1.9 / 0.68	0.4 / 0.63	3.2 / 0.72	6.1 / 0.82	24.4 / 1.00	22.5 / 1.00
	collusion, bandwagon	8.3 / 0.86	4.9 / 0.82	12.4 / 0.92	13.4 / 0.95	17.9 / 1.00	18.1 / 1.00
	coordinated poisoning	0.1 / 0.97	0.1 / 0.97	0.3 / 0.99	0.6 / 1.00	0.8 / 1.00	0.9 / 1.00

The whole trade-off. Figure 3 plots clean depth against worst-case error for every rule family over its whole parameter range. At the permissive operating points where calibration lands, vote quorums have the lowest worst-case error: at the clean depth of Q_2, the attack lower bound of the next best family is 1.07 to 1.36 times the exact worst case of Q_2, and that of HSC 1.4 to 2.9 times (1.27 to 1.53 and 1.7 to 2.6 at $f = 3$). Between clean depths of about 0.57 and 0.67 on CIFAR-100, 0.70 and 0.78 on PBMC and 0.08 and 0.17 on FGVC-Aircraft, the exact worst case of the probability quorum is below every certified bound of the climbing rules (the trimmed mean and median reach lower attack values on CIFAR-100 and PBMC), and at still more conservative points HSC with a high threshold has a lower certified bound. $Q_{f'}^{\mathrm{P}}$ with $f' \le f$ attains exactly the certified bound of HSC at threshold $\frac{1}{2} + \frac{f'}{2M}$ with more depth, and the clean anchor Q_{2f} caps the worst case at the clean error of the majority ball (9.5%, 9.5%, 8.9%).

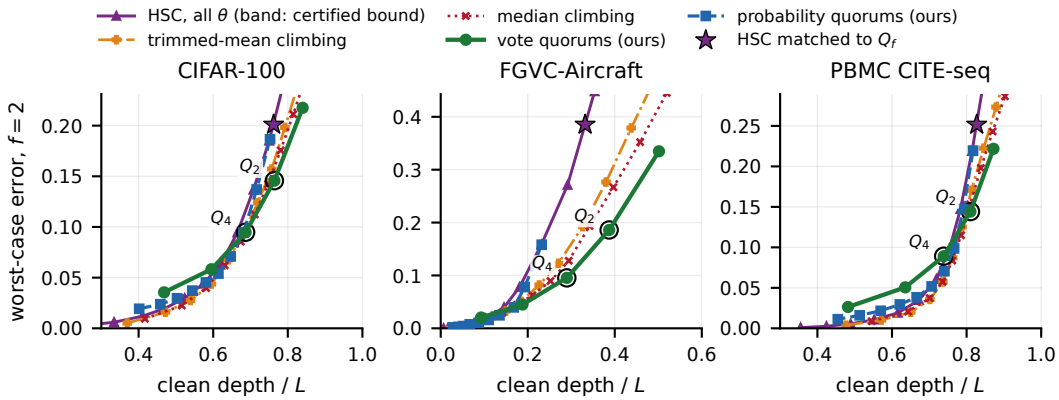

Figure 3: Clean depth against worst-case error, $f = 2$, for every rule family over its whole parameter range: vote quorums (every quota; Q_2 and the clean anchor Q_4 circled), probability quorums $Q_{f'}^{\mathrm{P}}$ (every design budget f'), and climbing on the mean (HSC), trimmed mean and median. Quorums are exact; climbing rules show the attack lower bound (HSC band: up to the certified bound). Star: HSC matched to Q_2.

Frontier with three corrupted sources and with other specificity measures. Figure 4 repeats Figure 3 for $f = 3$: on all three datasets vote quorums have the lowest worst case over a slightly wider range of depths, since their crossover with the probability quorum moves from about 0.67, 0.17 and 0.78 at $f = 2$ to 0.65, 0.14 and 0.75. Figure 5 replaces the normalised depth by information-based specificity and by the share of genuine branch decisions resolved (Appendix D). The two measures differ from the normalised depth by at most 0.04 and 0.03 for every rule family and parameter, and the frontiers keep their shape and ordering.

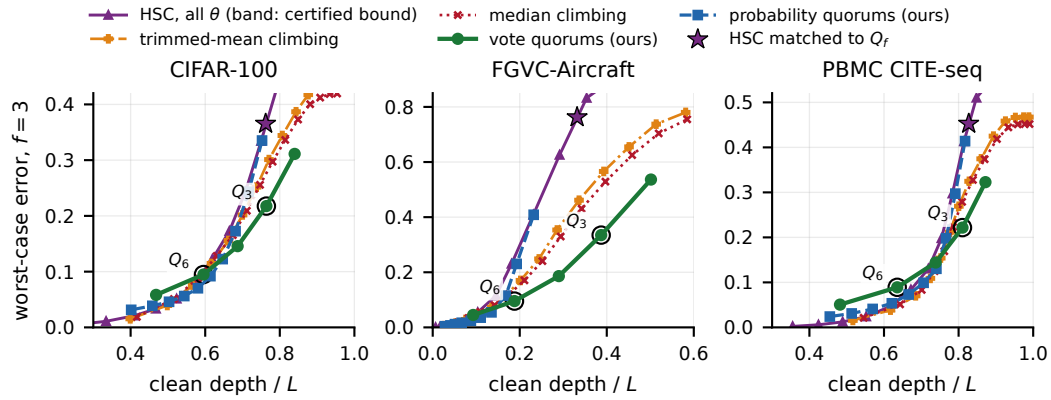

Figure 4: As Figure 3, $f = 3$ corrupted sources ($Q_3 = Q_2$ and the clean anchor Q_6 circled).

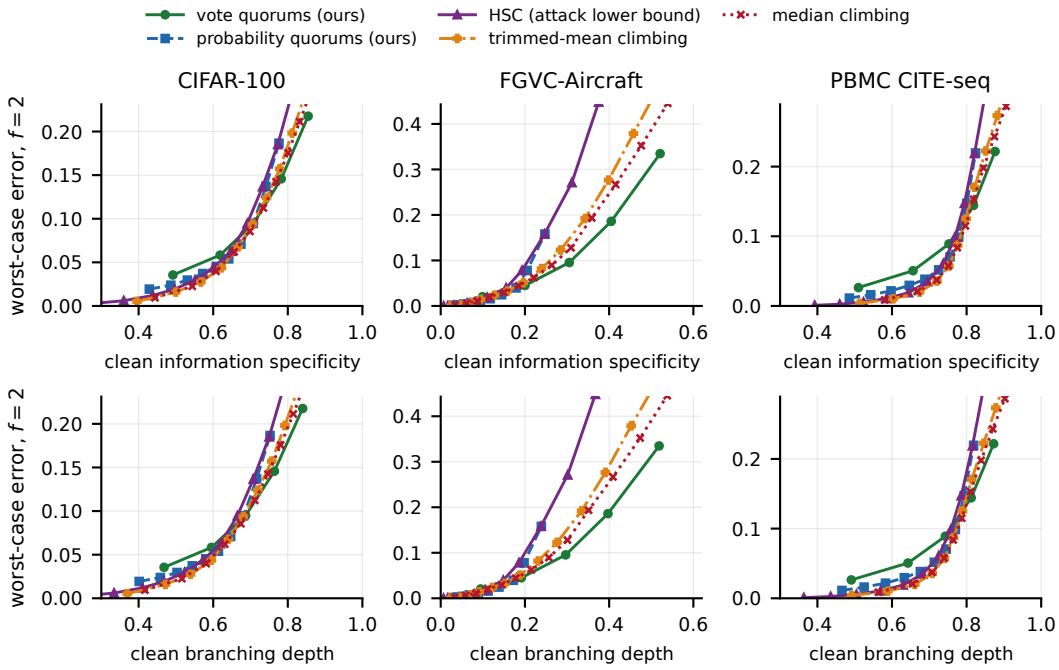

Figure 5: As Figure 3, $f = 2$, with clean specificity measured by $1 - \log|B|/\log|\mathcal{Y}|$ (top) and by the share of genuine branch decisions resolved (bottom). Climbing rules: attack lower bound.

Worst case as a function of f. Figure 6 plots the matched comparison for every f, and the worst-case depth of Q_f with the true hierarchy, with a randomly permuted hierarchy and without a hierarchy. The worst-case error of Q_f stays flat from $f = 1$ to $f = 2$ and from $f = 3$ to $f = 4$ because, with $M = 10$, a wrong ball needs the same number of clean reports, more than $(M - f)/2$, in both cases, and from $f = M/3$ on its worst-case depth is zero, as Theorem 2(iii) states. Under the worst corruption the calibrated HSC loses depth as well: by the diversion search its depth falls to 0.59, 0.13 and 0.68 at $f = 2$, against 0.60, 0.19 and 0.64 for Q_2.

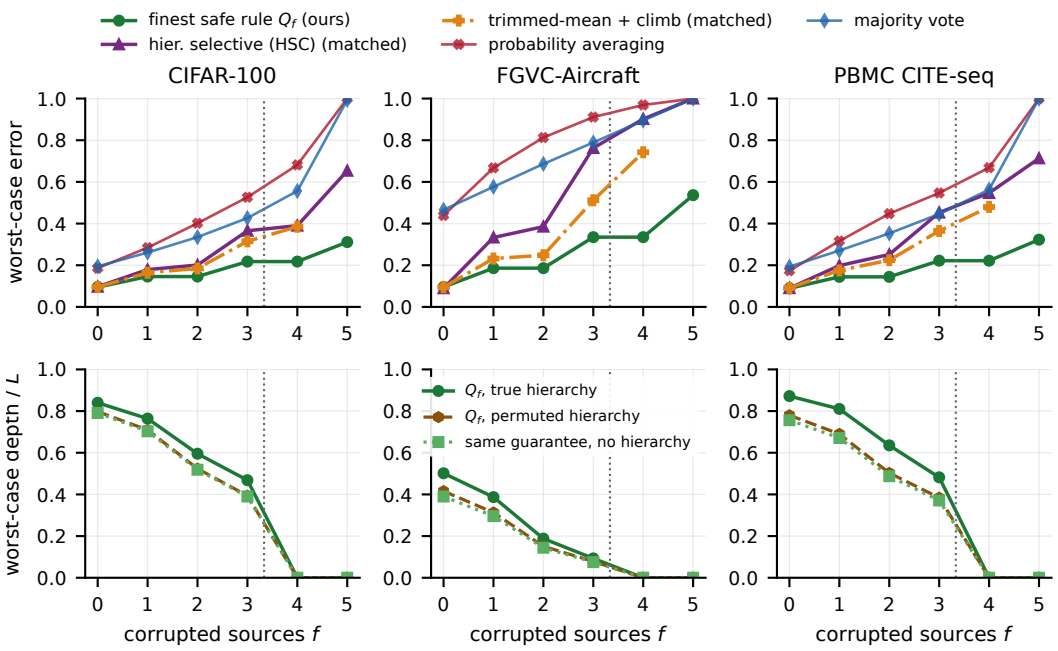

Figure 6: *Top:* worst-case error when f of the ten sources are corrupted. HSC and trimmed-mean climbing are re-calibrated at every f to the clean error of Q_f; probability averaging and majority vote have no operating point. The trimmed mean is shown for $2f < M$. *Bottom:* worst-case depth of Q_f with the true hierarchy, with a randomly permuted hierarchy (mean of five permutations), and of the same guarantee without a hierarchy. Dotted line: $f = M/3$.

The one-third barrier across source counts. Figure 7 draws random subsets of $M' \in \{3, \ldots, 10\}$ sources and plots the worst-case depth of Q_f against f/M'. Every point with $f \geq M'/3$ is at depth zero, as Theorem 2(iii) requires, and every point below keeps a positive mean depth, consistent with its guarantee for unanimous honest reports.

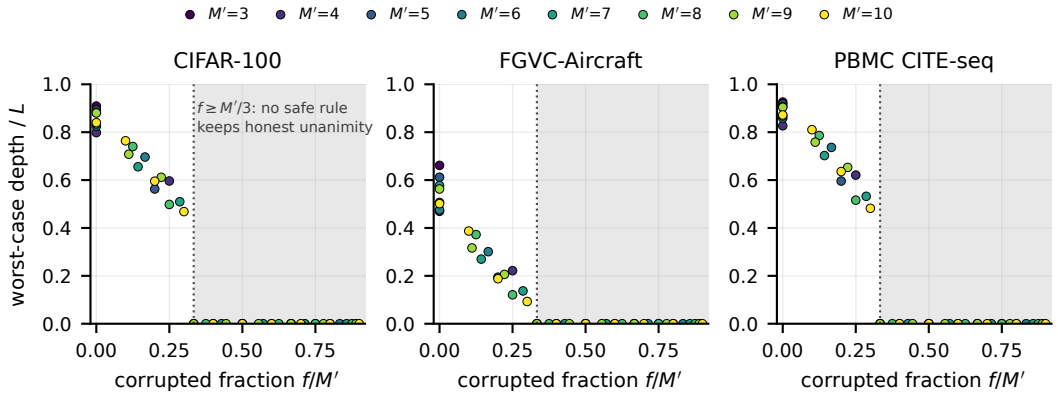

Figure 7: Worst-case depth of Q_f for random subsets of M' of the ten sources (five subsets per $M' < 10$), against the fraction f/M' of corrupted sources. In the shaded region $f \geq M'/3$ no safe rule can keep an answer that all honest sources share.

Human reports. Table 23 and Figure 8 report CIFAR-10H for several numbers M of sampled human reports per image. Plurality and top-down votes never abstain, so their worst case is larger, increasingly so as f/M grows; Q_f keeps nearly full depth until f approaches $M/3$ and then drops to the root. With one-hot reports, climbing on vote shares with a threshold above $\frac{1}{2}$ is a vote quorum

along the plurality path, and it is f-safe if and only if $\theta > \lfloor (M+f)/2 \rfloor / M$, which is 0.6 for both $f = 2$ and $f = 3$ when $M = 10$ (Proposition 1(iii) needs fractional reports for its converse). Matched to Q_2 with $M = 10$, it lands just above 0.6 in every draw, so it keeps a ball with at least 7 of the 10 votes; it then contains Q_2 and nearly coincides with it (clean error 0.37% against 0.38%, depth 0.988 for both, and the same worst case of 1.2%).

Table 23: CIFAR-10H: Q_f with M sampled human reports per image (mean of five draws), and the worst case of plurality and top-down votes. Errors in %.

M	f	Q_f clean	Q_f depth	Q_f worst	Q_f worst depth	plurality worst	top-down worst
5	0	1.3	1.00	1.3	1.00	1.4	1.5
	1	0.3	0.98	1.3	0.90	4.4	4.4
	2	0.3	0.98	3.9	0.00	17.4	17.4
10	0	0.7	0.99	0.7	0.99	1.0	1.0
	1	0.7	0.99	1.2	0.99	1.7	1.7
	2	0.4	0.99	1.2	0.95	3.2	3.2
	3	0.4	0.99	2.1	0.84	6.1	6.0
	4	0.2	0.98	2.1	0.00	18.2	17.4
20	0	0.7	1.00	0.7	1.00	0.8	0.9
	1	0.7	1.00	0.9	0.99	1.2	1.2
	2	0.5	0.99	0.9	0.99	1.5	1.6
	3	0.5	0.99	1.2	0.98	2.0	2.1
	4	0.4	0.99	1.2	0.96	2.7	2.8
50	0	0.6	1.00	0.6	1.00	0.7	0.7
	1	0.6	1.00	0.7	1.00	0.9	0.9
	2	0.5	1.00	0.7	0.99	1.0	1.1
	3	0.5	1.00	0.9	0.99	1.2	1.3
	4	0.5	0.99	0.9	0.99	1.3	1.4

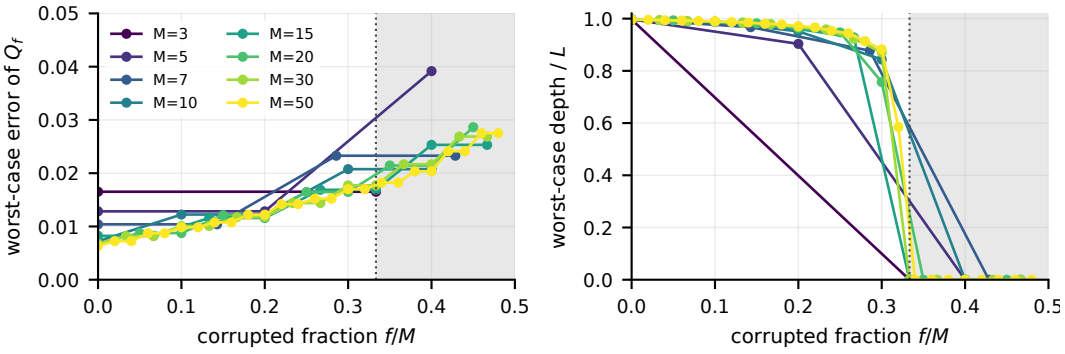

Figure 8: CIFAR-10H: worst-case error (left) and worst-case depth (right) of Q_f against the corrupted fraction f/M, for M sampled human reports per image.

Level survival and overdispersion. By Corollary 1, the share of examples whose depth-k decision is certified correct under every corruption of f sources is $\Pr[X_k \geq \tau_f + f]$. Figure 9 and Table 24 compare it with the binomial prediction for independent sources of mean accuracy a_k (Proposition 9) and with a beta-binomial fit whose intra-class correlation κ_k is estimated on the same data. The binomial errs by up to 41 points in both directions: real sources keep more certified-correct decisions than independent ones when the required count exceeds the mean number of correct sources, and fewer when it is well below it. The beta-binomial with one extra parameter per level is within 3.2 points everywhere (mean 0.8). Human reports are much less correlated (κ_k from 0.09 to 0.19) than machine sources (0.30 to 0.48).

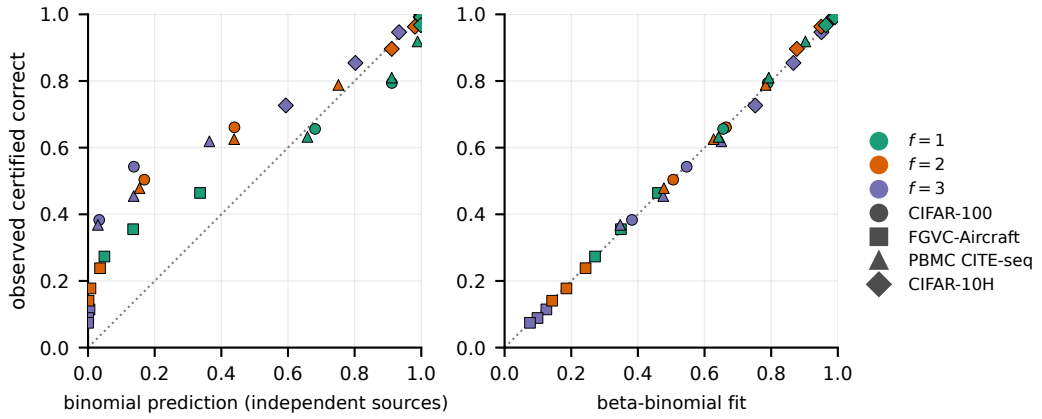

Figure 9: Observed share of examples certified correct at level k under every corruption of f sources, against the binomial prediction for independent sources of the same accuracy (left) and a beta-binomial fit with the estimated intra-class correlation (right); one point per dataset, level and f.

Table 24: Per level: mean source accuracy a_k, fitted intra-class correlation κ_k of the correctness of the sources, and the share (%) of examples certified correct at that level: observed, binomial prediction and beta-binomial fit. CIFAR-10H: one draw of $M = 10$ reports per image.

				$f = 2$			$f = 3$		
Data	level	a_k	κ_k	observed	binomial	beta-bin.	observed	binomial	beta-bin.
CIFAR-100	1	0.820	0.45	66.1	43.9	66.5	54.3	13.8	54.6
	2	0.712	0.48	50.4	16.9	50.6	38.3	3.3	38.2
FGVC-Aircraft	1	0.581	0.30	23.8	3.6	24.3	11.5	0.4	12.5
	2	0.476	0.38	17.8	0.7	18.6	8.9	0.1	9.9
	3	0.392	0.42	14.1	0.1	14.3	7.4	0.0	7.6
PBMC CITE-seq	1	0.904	0.30	78.8	75.1	78.3	61.9	36.4	65.1
	2	0.820	0.35	62.6	43.8	62.7	45.5	13.7	47.6
	3	0.703	0.44	47.9	15.5	47.8	36.8	3.0	34.7
CIFAR-10H	1	0.993	0.09	99.1	99.8	98.6	94.6	93.3	95.1
	2	0.978	0.12	96.3	98.1	95.0	85.4	80.2	86.7
	3	0.949	0.19	89.6	91.1	87.7	72.7	59.4	75.2

Stability versus safety. Table 25 reports the share of examples whose depth-k decision is certified correct under f corruptions: by safety (Q_f keeps depth k and cannot be driven wrong at depth k or above), by stability of the top-down vote on the clean profile (its prediction cannot be changed by f replacements), and by the stability rule of Proposition 5 applied to the observed profile, which certifies depth k exactly when the clean profile is stable at radius $2f$. Stability of the plain top-down vote certifies more examples, but that rule never abstains, and its worst-case error is 34%, 69% and 35% at $f = 2$, against 15%, 19% and 14% for Q_2. Applied to observed profiles, stability certifies fewer examples than safety for $f \geq 2$ and none for $f = 3$, as Proposition 5 predicts.

Table 25: Certified accuracy (%) of the depth-k decision under f corruptions, for the top level and the leaves: safety of Q_f, stability of the top-down vote on the clean profile, and stability used on the observed profile (radius $2f$ on the clean profile).

		top level			leaves		
Data	f	safety	stab. (clean)	stab. (obs.)	safety	stab. (clean)	stab. (obs.)
CIFAR-100	1	79.4	84.4	79.4	65.6	73.7	66.3
	2	66.1	79.4	60.2	50.4	66.3	44.2
	3	54.3	72.0	0.0	38.3	57.1	0.0
FGVC-Aircraft	1	46.4	63.2	49.7	27.3	41.9	31.0
	2	23.8	49.7	17.6	14.1	31.0	10.5
	3	11.5	34.3	0.0	7.4	20.8	0.0
PBMC CITE-seq	1	91.9	94.2	91.1	63.2	73.3	65.0
	2	78.8	91.1	73.0	47.9	65.0	43.7
	3	61.9	85.1	0.0	36.8	55.9	0.0