STATISTIC PRESERVATION DOES NOT IDENTIFY ATTACK-SUPPRESSION PRESERVATION IN COMPOSED FEDERATED-LEARNING DEFENSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning composes defenses never designed together. Does the downstream defense still suppress the attack? **A preserved statistic is not preserved suppression.** The natural answer asks whether the downstream defense's own statistic is preserved under composition, and it fails structurally rather than statistically. Admitting a composition because one constituent already suppresses the attack selects on the outcome to be explained. The counterfactual identification requires is then absent, so the comparison is missing at any n. The fix is a design, not a better statistic: hold the downstream defense and attack fixed, and intervene upstream. Ignoring this costs not power but the sign. On one `coord_median` cell the outcome-gated design moves -0.273 where the within-defense one moves $+0.125$ (n=20, both 95% intervals exclude zero). Against ourselves: with the aggregate magnitude held fixed the $+0.125$ is -0.172 and both agree in sign. The reversal replicates on the same cell on CIFAR-100; a composition-level replication there was attempted and did not carry. Decision change is neither necessary nor sufficient for suppression to move. One aggregator's selection flips repeatedly while admitting no adversary; another's is bit-identical while attack success falls. Oracle-free, a Krum decision flip alone raises ASR $+0.322$, so both need an oracle no server has. This is identification, deliberately scoped to committed attacks and controlled interventions, and not a deployable security test or a predictor of adaptive robustness.

1 INTRODUCTION

A composition of two defenses suppresses an attack. **Is the downstream defense's mechanism what suppressed it?** Everything below is an answer to that one question.

Federated learning (McMahan et al., 2017) is vulnerable to data poisoning (Chen et al., 2017; Bagdasaryan et al., 2020), and an arms race has produced defenses (Fung et al., 2020; Pillutla et al., 2022; Yin et al., 2018; Blanchard et al., 2017) with accuracy/robustness trade-offs and no dominant member (Appendix F). *Can we compose them for stronger robustness?* Composition does not even commute: FoolsGold→CoordMedian suppresses a pixel backdoor to ASR 0.099 while the reverse reaches 0.732, that order collapsing to FoolsGold alone bit-identically, since a rank aggregator emits one vector rather than per-client updates.

We propose no new defense. Preserving what the downstream defense reads leaves open whether the composition still suppresses the attack. The natural test admits a composition because one constituent already suppresses the attack, and so cannot answer the question at any sample size: it removes the counterfactual support identification requires.

The statistic is not the estimand; the effect of the upstream intervention is. What replaces it is a design, not a better test: hold the downstream defense and the attack fixed and intervene on the upstream transform rather than compare across a menu of defenses. The transform can weaken the attack itself, so an outcome-gated comparison credits the downstream defense with a fall in ASR that attenuation upstream produced. On one cell that change reverses the *sign*.

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.

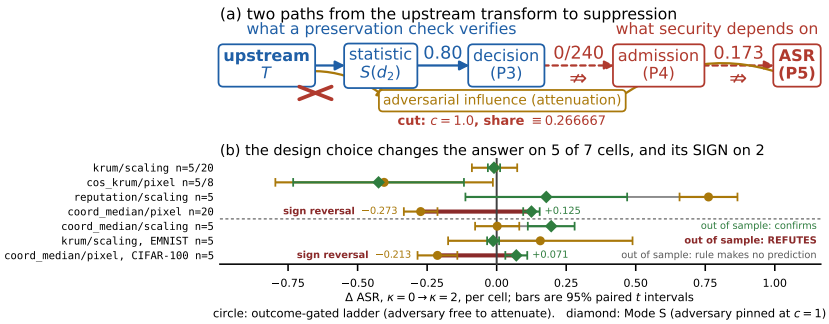

Figure 1: **The paper in one figure. A preserved statistic is not preserved suppression. (a)** What a preservation check verifies, left of the divide, against what security depends on, right of it: transform, statistic, decision, admission, ASR. Solid links hold by definition; dashed ones (⇏) do not, and the lower arc is the attenuation channel Mode S cuts. **(b)** Seven cells, one contrast each, estimated both ways (App. D.3). Row labels carry the n *frozen here* (App. I for the one row since topped up), outcome-gated/Mode S where the two differ, and the four above the dashed rule are training data for a since-withdrawn rule, not evidence for it.

ΔASR, one cell	outcome-gated	within-defense
CIFAR-10, n=20	-0.273 $[-0.334, -0.213]$	$+0.125$ $[+0.095, +0.155]$
CIFAR-100, n=5	-0.213 $[-0.284, -0.141]$	$+0.071$ $[+0.031, +0.110]$

Both paired 95% intervals exclude zero and are disjoint on the first row; the rows hold `coord_median`, attack and rung grid fixed and differ in dataset at the n shown (§5, App. D.3). Both columns leave the aggregated update's magnitude free to move, and a pre-registered control that closes it reverses the within-defense leg's own sign (App. D.10). **Both instruments are identification devices, not deployable defenses or diagnostics**: they read adversary identity, which no server has, and pre-registered oracle-free versions refuted us at n=20 (App. D.7, App. D.8).

Two measured facts put the question there, and they fail the criterion separately. Two defenses that each fail alone can jointly suppress a committed (non-adaptive) attack: FoolsGold→RFA reaches ASR 0.093 where on the pixel backdoor FoolsGold alone reaches 0.732 and RFA alone 0.845. No gate of this shape can certify it: a *coverage* failure, independent of the broken link that follows (Prop. 2). And the link an evaluator leans on breaks in both directions: Krum's selection flips in 0.733 of rounds while its admission is unchanged and its suppression does not follow ($|\Delta|$=0.026), while `cos_krum`'s selection is bit-identical at every rung and its ASR nonetheless *falls* 0.173 (both drawn in Fig. A7, App. I).

Why the question is sharper without a trusted coordinator. A contributor clips, a scheduler reweights by reputation, and whoever merges aggregates robustly: three independently developed mechanisms, nobody owning the stack, so the only check at each seam is that the next stage still reads what it expects. We claim no transfer beyond the controlled setting we study (page 4).

Contributions. The question is scientific rather than operational. *Our claim is not that preservation is uninformative, nor that admission predicts attack success*: it is the narrower one that no outcome-gated cross-defense comparison can attribute a composition's suppression to the downstream mechanism, because the gate removes the contrast such an attribution needs. The evaluation criterion that result implies is priced rather than validated, and we do not claim it certifies security. Table 1 gives each claim's level and status; supp. §S12 answers the objections we expect.

In the appendix: every proof, cited where needed; the literature audit is in the supplement (§S16).

2 What the Gated Test Cannot Identify, and What Can

Comparing many defense pairs is where identification is lost. Screen is the shorthand the formal statements and the appendix keep for what the narrative calls the *evaluation criterion*, or *criterion*

Table 1: **What this paper claims, at what level, and with what status.** Prop. 3's two counterexamples are exact invariances, not small mean differences; Table A15 is the full ledger.

Claim	Level	Status
Outcome-gated evaluation cannot attribute suppression	Formal, general	proved; exact (Prop. 1)
A criterion of that shape cannot certify emergence	Formal, general	proved (Prop. 2)
Positive per-client rescaling preserves lower levels	Formal, scoped	proved (Prop. 11, Thm. 9)
Neither decision state determines suppression	Measured	exact invariances (Prop. 3)
The same, without adversary identity	Measured, n=20	refuted, by us (App. D.8)
The sign disagreement is not a magnitude artifact	Measured, n=20	refuted, by us (App. D.10)
Our criterion predicts a menu's best compositions	Retrospective	not established

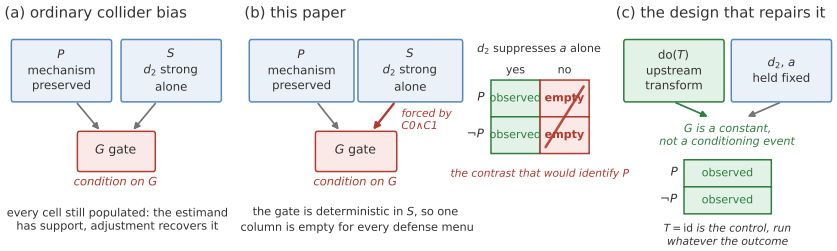

Figure 2: **Not ordinary collider bias but a structural *positivity* violation, which is why the remedy is a design and not an estimator. (a)** Ordinary collider bias, where every cell stays populated and adjustment recovers the estimand in principle. **(b)** The inherited-suppression gate $G{=}E$ is *deterministic* in $S{=}M(d_2, a)$, so the column the identifying contrast needs is empty and no sample size restores it (Prop. 1). **(c)** The repair: hold d_2 and a fixed, intervene on T, and run $T{=}\mathrm{id}$ as the control whatever its outcome (Cor. 1; App. C.2).

for short: a rule for choosing which compositions are worth running, never a predictor of security. Also fixed: *gate* is the condition below, C0–C3 the composability conditions, *protocol* §4's clauses, *instruments* Mode S (statistic-only) and score-only Krum.

Intuition. An admitted pair's downstream defense is guaranteed to suppress the attack alone, so no admitted pair exists in which that defense fails alone while the composition succeeds. **That empty cell is the whole of it**, and it is empty by the gate rather than by bad luck: it is the one comparison that would credit the suppression to the mechanism rather than to the downstream defense's own strength, and comparing the cells the gate does leave occupied varies the downstream defense, which changes the conditioning event instead of intervening on the transform.

Proposition 1 (When mechanism attribution is identifiable, and when it is not)**.** *Let a screen certify* $(d_1 \to d_2)$ *against attack* a *only if both (i) an* ***eligibility gate*** $E(d_1, d_2, a) := \min_{d \in \{d_1, d_2\}} M(d, a) < \theta$ *holds, where* M *and* θ *are the* same *outcome metric and threshold the screen predicts for the composition, and (ii) a* ***mechanism condition*** $P(d_1, d_2)$ *holds, determined by the two defenses' definitions alone. Suppose the screen is applied only where the downstream mechanism is load-bearing,* $M(d_1, a) \geq \theta$*. Consider estimating* P*'s contribution to* $M(d_1 \to d_2, a)$ *from contrasts across a defense menu* $\mathcal{D}$*. Then every certified pair satisfies* $M(d_2, a) < \theta$*, so of the four cells of* $\{P, \neg P\} \times \{M(d_2, a) < \theta,\ M(d_2, a) \geq \theta\}$ *the certified set occupies at most two, and the cell needed to separate* P *from* d_2*'s standalone effectiveness is empty by the gate, for every* $\mathcal{D}$*. No contrast varying* d_2 *identifies* P*'s contribution. Conversely, lift the gate's reading of* M *and the same contrasts identify it, so on this class of screens the boundary is exact. Identification therefore requires holding* d_2*,* a *and therefore* E *fixed while intervening on the upstream transform:* ***the admissible designs are within-defense.***

What this characterizes is a shape of criterion, not federated learning. M is any outcome metric, θ any threshold, P any condition on the two mechanisms' definitions, and $\mathcal{D}$ any menu; App. C.2 substitutes our protocol into each, term by term. The proof is short (App. C.1).

What is general, what is FL theory, what is FL evidence. The violation is standard; the three layers around it are not, and they are Table 1's three levels: general, one named class of FL mechanisms, and measured. What is general is the estimand, an intervention on the upstream transform: which *level* of the hierarchy it preserves, and that no outcome-gated test identifies it (Prop. 1). What is FL evidence is that on one cell the two admissible estimands do not differ in precision, they give contrasts of opposite sign (§5).

What we estimate instead. Fix a downstream defense d_2 and a committed attack a. The **treatment** T is the upstream per-client transformation applied before d_2 aggregates, with $T{=}\mathrm{id}$ the control that is bit-identical to d_2 standalone. The **outcome** Y is $\mathrm{ASR}(T, d_2, a)$. The **mechanism-preservation variable** $P \in \{0, 1\}$ records whether T preserves the statistic d_2 reads, at whichever level is at issue. Which of the five levels is at issue is §3's subject; the boundary below holds for any such condition. P is a deterministic function of T and d_2's definition, not a measured covariate. The causal quantity is $\tau(T) = \mathbb{E}[Y \mid \mathrm{do}(T), d_2, a] - \mathbb{E}[Y \mid \mathrm{do}(T{=}\mathrm{id}), d_2, a]$.

We do not estimate the causal effect of P itself. P is not manipulable apart from T, being a property of the pair, so "the effect of preservation" names no intervention. We estimate the effect of intervening on the upstream transformation while holding d_2 and a fixed, and what we report about P is whether $\tau(T) \neq 0$ occurs while P holds, and whether $\tau(T){=}0$ occurs while P fails. Both occur (§5), which is what refutes P as a surrogate for Y.

Only half of this is ordinary collider bias. We claim no credit for it: that half is textbook (Elwert & Winship, 2014; Pearl, 2009). What the gate adds is a structural *positivity* violation rather than a confounding one (Fig. 2b), while reading as a reasonableness filter. The gate also earns its keep: it stops the criterion certifying pairs no constituent suppresses, so wherever our conditions certify a pair, a low composition ASR is attributable to d_2's standalone strength rather than to preservation, the reading we held ourselves (§6).

A criterion of this shape fails a second, independent way, bounding what it can find rather than what it can be validated on: no emergent composition can pass the gate, so recall is capped structurally.

Proposition 2 (Coverage: inherited-suppression screens cannot certify emergent synergy)**.** *$(d_1 \to d_2)$'s suppression of a is **emergent** if $M(d_1 \to d_2, a) < \theta$ while $\min_{d \in \{d_1, d_2\}} M(d, a) \geq \theta$. No emergent composition satisfies the eligibility gate of Prop. 1. Hence, writing $\mathcal{S}$ for the compositions with $M < \theta$, $\mathcal{E} \subseteq \mathcal{S}$ for the emergent ones and $\mathcal{S} \setminus \mathcal{E}$ for the inherited ones, any such screen's recall obeys* $\mathrm{recall} \leq 1 - |\mathcal{E}|/|\mathcal{S}|$*, on every menu and every attack.*

Our own lowest-ASR composition is the witness: §1's FG→RFA, at max-committed ASR 0.093 ($n{=}30$, CI $[0.062,\ 0.123]$) against constituents that each fail alone, is emergent and a false negative of our own criterion, which is what puts the bound above the recall we report (App. C.1). It splits a menu into *inheritance*, which we test for, and *emergence*, a search problem we do not solve. A criterion reaching $\mathcal{E}$ would have to read the composition's own suppression, the outcome it exists to avoid conditioning on: Prop. 1's boundary met from the other side. Our blind spot is one condition, C1, so that recall is an *operating point*, not a ceiling (§6).

What this paper establishes, and what it does not.

Formally. Neither proposition above contains a dataset, architecture or scale, so what carries beyond this paper is the design they license and not our criterion. Thm. 9 (App. C.3) and the invariance classes are proved for positive per-client rescalings into named rank and geometric aggregators, and bound no ASR. In the other direction they also say when preservation does hold: on that class the downstream discriminative mechanism survives, given a bounded weight ratio for the rank aggregators and a separation margin for the geometric one, and Cor. 3 puts the latter in a form checkable before the composition is run. What survives is the suppression ordering rather than any security property, and the conditions need a short pilot (L5).

Empirically. We identify the effect of intervening on the upstream transform at fixed downstream defense and attack, and offer no causal model of FL; on one cell that change flips the *sign* of the estimated effect. The experiments are four aggregators of three structural kinds under two committed attacks on CIFAR-10 with a 4-layer CNN, $N{=}10$ clients and $f{=}0.2$ adversarial, three arms replicated, on EMNIST-byclass ($n{=}3$ frozen, $n{=}5$ pooled), `resnet18` ($n{=}3$) and CIFAR-100, at $n{=}5$, $n{=}20$ on the CIFAR-10 sign-reversal cell

and Krum/scaling's Mode S leg, n=8 for `cos_krum`, plus part of the 42-pair menu re-run on CIFAR-100 (§S24.3).

Not established. We do *not* establish that admission predicts security, we claim nothing about adaptive robustness, our instruments read adversary identity (§H), and we claim nothing about confidentiality or availability. This is a methodology for identifying what a composition preserves, not for choosing a defense to deploy: the criterion prioritizes what to evaluate, and nothing here is a deployment recommendation. Claims do not transfer upward either: scope conditions in Table A2, evidence tiers in Table A1, limits in §G, (L1)–(L6), and every withdrawn or refuted claim in Table A14. The criterion's validation is exploratory (L4).

3 Which Level a Composition Preserves

"Preserved" has five senses. The condition P the boundary quantifies over is not one condition but a graded family, and **which member of it a claim asserts decides whether the claim is checkable at all**. An upstream transform can leave a downstream defense "the same" in five senses, ordered from what the defense computes to what the attacker gets: three about the defense's *internal* behaviour, then two about *security* behaviour, what the adversary gets.

The algebra cuts one place later, everything up to admission being *observable* and attack suppression *mechanistic*. These are five different claims, and preservation at one level does not imply preservation at the next. **Where that chain breaks is this paper's subject**: the first two implications are logical and need no experiment, while both failures start at admission, whose preservation implies neither a preserved decision (Krum, §5) nor preserved suppression (`cos_krum`, §6).

To instantiate the hierarchy in FL we characterize the invariance classes, which is where the levels stop being a taxonomy and start being decidable: every per-client defense we consider acts as a positive scalar rescaling $T(u_i) = c_i u_i$, so the algebra fixes which level a given composition sits at, before any run: cosine similarity is exactly invariant, coordinate ordering only under a bounded-ratio condition, and consensus distance and residual magnitude not at all (Prop. 11, App. C.3). That class is also the theory's whole reach: nothing we prove speaks to a transform outside it, and two composed defenses are outside it, Bulyan (El Mhamdi et al., 2018) zeroing the clients its selection drops and FLAME (Nguyen et al., 2022) adding noise to the aggregate.

Definition 1 (Preservation levels)**.** *For an upstream transform T and a downstream defense d_2 reading statistic S:*

- *(P1)* ***Value invariance:*** *$S(T(U)) = S(U)$.*
- *(P2)* ***Ordering invariance:*** *$S(T(U)) = \varphi(S(U))$ for a φ preserving S's client ordering.*
- *(P3)* ***Decision invariance:*** *the client or coordinate d_2 actually selects is unchanged.*
- *(P4)* ***Adversarial contribution (admission) invariance:*** *the adversarial input d_2 admits into its aggregate is unchanged.* Admission *is the pre-registered term throughout for whether adversarial input survives d_2: the support of the admitted mass, not the mass itself. The mass is influence Λ_a, defined at the end of this section and in App. J, which a preserved support leaves free to move (Table 2). This level is instantiated defense-specifically and is not a single metric, so we never compare its magnitude across aggregators.*
- *(P5)* ***Attack-suppression preservation:*** *ASR stays below threshold for the arm's committed attack.*

The levels induce the screening conditions C0–C3, stated in App. C. Where each sits on this ladder is the argument in miniature: the two preservation conditions ask whether the transform leaves admission intact, while the two applicability conditions gate on standalone ASR, which is suppression itself, so the criterion conditions on a descendant of what it predicts.

Six words, fixed here and used in one sense throughout, in the order the chain runs.

term	level	what it names	bears on security
statistic	(P1)	what the defense computes from the updates it reads	no
ordering	(P2)	the ranking that statistic induces over clients	no
decision	(P3)	which client or coordinate it selects	no
The implication breaks at this rule: no level above it implies any level below it.			
admission	(P4)	whether adversarial input enters its aggregate at all	yes
influence	—	the weight the adversary's direction carries in the aggregate (Λ_a)	yes
suppression	(P5)	whether the attack still succeeds, i.e. max-committed ASR	yes

Influence appears there with no level of its own, because Λ_a is not a step in the chain but the quantity crossing from admission to suppression; Figure 1(a) draws that chain, and the arc beneath it is the one channel these rows do not name. *Attenuation* is the adversary's own update shrunk upstream, reaching ASR without passing through the statistic, and it is the channel an intervention on T must close before either of the other two can be read.

Admission (P4) is the first level at which adversarial influence enters the causal graph, which is why we intervene there rather than on the statistic. It is privileged for its semantics, not for sufficiency. It is also instantiated per aggregator (App. J), so our admission result is four within-family statements and we never difference or rank across them.

4 A Causal Protocol: Intervene Upstream, Not Across Defenses

The inadmissible comparison is replaced by a different design, not a better criterion. Prop. 1 determines it: do not select across defenses; hold the downstream defense and the attack fixed, intervene on the upstream transformation, and measure the channels it opens separately. Writing T_1 for it, the estimand is $\Delta_{\text{mech}} = \text{ASR}(T_1^{\text{int}}, d_2, a) - \text{ASR}(T_1^{\text{base}}, d_2, a)$, at fixed d_2, attack a and adversarial coefficient share; the third clause below is why this is a mechanism contrast, not a dose.

(1) **Hold d_2 and the attack fixed.** *Forced by* the gate $E(d_1, d_2, a)$: vary d_2 or a and the two sides of a contrast are conditioned on different events, so they are not comparable at any sample size.
(2) **Intervene on the upstream transform.** *Forced by* the empty cell: variation in d_2 is exactly what the gate turns into missing support, so the intervention has to live on d_1.
(3) **Pin the adversarial coefficient channel.** *Forced by* what the proposition does not constrain: nothing prevents T from reaching the outcome without passing through d_2's statistic at all, by attenuating whichever client carries the poison (Lem. 2). Mode S pins it by setting every adversary's coefficient to c=1, which requires knowing which clients are adversarial (§5).
(4) **Measure the channels, not only the outcome.** *Forced by* what P asserts: a claim about a mechanism, so a contrast reporting only ASR cannot say which channel carried the effect, and every channel the intervention opens must be reported, including the accidental ones. Report the arm as ΔP_1–ΔP_5 and classify it by the highest level that survives, five outcomes; Table 2 is that report for seven arms.

Operationally, before any of that, name d_2's invariance class and the level of Def. 1 the claim asserts (§3; Table A1's Formal row names the transforms the algebra does not reach).

Threat model: committed, not adaptive. Every clause holds the attack fixed, so the protocol establishes attack-suppression preservation under a fixed committed attack and nothing wider, which is why we do not call the object "security". One whitebox composition-aware *adaptive* stress test measures that boundary: solving the joint FoolsGold/RFA constraint knowing both stages raises FG→RFA to ASR 0.157 ± 0.073 at n=30, $2.4\times$ its seed-matched 0.064 (paired $+0.093$, $[+0.066, +0.120]$). **Inherited robustness is not adaptive robustness**, and a protocol stated against an adaptive adversary is the sharpest thing this framework lacks (supp. §S19).

5 Statistic Preservation Does Not Identify Attack-Suppression Preservation

Disturbing what a defense reads need not cost suppression, and **on one cell the two designs disagree in sign**: the outcome-gated design gives ΔASR$=-0.273$ where the within-defense design

Table 2: **The four channels, all four aggregators, under Mode S**; endpoint change $\kappa{=}0 \to \kappa{=}2$. Δ agg. is the relative displacement of the aggregate the defense *emits*, so it is what enters training and not the value of a scoring statistic; Δ adm. and $\Delta\Lambda_a$ are §3's admission and influence, never interchangeable. The magnitude channel stays open here; row 6 closes it, where a *larger* number is a *narrower* channel (App. D.2.1). **Row 6's Δ ASR is at** 20 **seeds and row 1's at** 5**, so the two are not differenced here**; App. D.2 pairs both at 20. The table does not order the arms, $\Delta\Lambda_a$ and ΔASR disagreeing in rank. Row 7 varies architecture alone (ResNet18). n is seeds, not rounds; provenance, round counts and the emitter are in App. D.2.

		magnitude	statistic	attenuation		total	
Aggregator	Kind	Δ agg.	Δ dec.	Δ adm.	$\Delta\Lambda_a$	Δ ASR	n
Krum	selector (pairwise distance)	0.892	0.733	0.000	0.000	-0.026	5
Reputation	weighted averager (cross-round state)	0.659	1.000	0.000	0.020	$+0.178$	5
Cosine-Krum	selector (cosine distance)	1.297	0.000	0.000	0.000	-0.425	8
Coord. median	coordinate-wise order statistic	0.670	0.482	0.000	0.033	$+0.098$	5
Krum (EMNIST-byclass)	selector (pairwise distance)	0.842	0.643	0.000	0.000	-0.017	3
Krum, score-only	selector, magnitude closed	0.986	0.733	0.000	0.000	-0.009	20
Krum (ResNet18)	selector (pairwise distance)	0.824	0.556	0.000	0.000	-0.041	3

gives $+0.125$ at $n{=}20$, with disjoint intervals both excluding zero, so the two admissible estimands give opposite contrasts here, not one at two precisions.

Proposition 3 (Empirical non-identifiability of preservation levels)**.** *Within the positive per-client rescaling class, on the defenses and attacks we run, there are transforms that change the statistic and the* decision *at d_2 arbitrarily at* unchanged *admitted adversarial mass and suppression, and transforms that change suppression at unchanged statistic, decision and admitted mass. Hence no predicate of the statistic identifies admitted mass and none of admitted mass identifies suppression: exhibited counterexamples, not a general law (Mode S into Krum and* `cos_krum` *under the pixel backdoor, both below; proof, Prop. 12).*

One cell is carried end to end, and it is Krum under model scaling: row 1 of Table 2, with row 6 its magnitude-controlled twin.

The instrument, as a counterfactual. The clean question asks what happens when the downstream statistic changes but the adversary cannot become weaker. **Mode S** enforces exactly that: it pins every adversary's coefficient at $c{=}1.0$ and spreads only the benign clients over $\rho = e^{2\kappa}$, holding the adversarial coefficient share constant at 0.266667 to within float32 read-back noise, closing Figure 1(a)'s attenuation path, so a rise in ASR here cannot come from it.

Two readings of preservation, frozen against each other. *The statistic reading*: what matters is disturbance of d_2's statistic, read off its *decisions*. *The admission reading*: what matters is the *adversarial input d_2 admits*. Mode S disturbs Krum's selection heavily at $\rho{=}54.6$ while leaving that admission untouched, so that arm adjudicates (Table A7).

Here a non-increase is established, not merely consistency: against that frozen ± 0.15 margin, an evaluation-design tolerance conjoined with every rung below 0.5, not a claim that 15 ASR points are negligible, only the upper TOST leg is a test, and the pre-registered top-up to $n{=}20$ bounds any rise at $+0.012$ ($\Delta{=}-0.010$; supp. §S18 scores every arm, us included). *Margin sensitivity: the margin is not load-bearing* for any reading in Table S10, each surviving the whole ± 0.05–± 0.20 ladder at a binding $m^*{=}0.034$; `resnet18`'s $n{=}3$ replication is the one arm that needs it (App. D.2).

The statistic reading is refuted: statistic and decision disturbance need not change suppression once the adversarial coefficient share is held fixed. *The zero is a floor here*, Krum admitting no adversary at any rung, so unchanged admission is not by itself decisive, and we claim no per-channel decomposition of ΔASR (§D.2.1).

Krum's decision channel has room, changing on 0.733 of rounds, so the refutation just stated does not rest on the floor. Its admission channel has none, nor do the paper's three other Krum rows; but that floor is the attack's, not Krum's, since under the pixel backdoor the same Krum admits baseline mass 0.333 in 4 of 12 adversary rounds, where the same masked dose moves ASR -0.2857 at $n{=}20$,

not the -0.010 above (App. D.9). The admission reading rests on the two arms with the most room, reputation and coord. median, at baseline 0.0014 and 0.286, whose admitted support holds across all 48 adversary rounds while suppression moves in both.

The table contains its own positive control, which tells "we moved something and nothing happened" from too blunt an instrument: suppression moves in exactly the two arms whose *influence* channel moves, reputation and coord. median. Krum changes the decision at $\Delta\Lambda_a{=}0.000$ and suppression does not follow, so a decision change is not *sufficient*; `cos_krum`'s selection is bit-identical at every rung, also at $\Delta\Lambda_a{=}0.000$, yet ASR *falls*, so it is not *necessary* either. Neither reading rests on Krum's admission floor; both are taken across all four kinds (App. D.2).

What Mode S does not hold fixed, and what closing it costs on each aggregator. The instrument identifies the intervention's effect at a fixed adversarial coefficient share, not statistic disturbance in isolation, so what it yields is a *dissociation*, not proof that disturbance is inert. It pins the adversarial coefficients alone, so the dose still moves the aggregated update's magnitude and geometry, a third channel reading Δ agg.$= 0.892$ here. Score-only Krum, frozen before any score-only ASR existed, closes it by aggregating the *untransformed* selected update at the same Δ dec. of 0.733: $\Delta = -0.009$ at $n{=}20$ against -0.010 uncontrolled, so on this cell the negative is instrumented against both confounds (App. D.2).

On the reversal cell the same control reverses the controlled leg's own sign: a coordinate-wise analogue, frozen likewise, takes the within-defense leg from $+0.125$ to -0.172 at $n{=}20$, so the disagreement between the two designs there was carried by magnitude and not by the *statistic* (App. D.10). The channel is inert on a selector and decisive on an order statistic, and neither result transfers.

A pre-registered replication that refutes the positive half. `coord_median` under the pixel backdoor is *rank-discordant*: of the interpretable arms it has the smallest decision change (0.482) and the largest admission-level change ($\Delta\Lambda_a{=}0.033$), so the two readings predict opposite rank positions for its rise, and the thresholds were frozen before its runs existed. We find $\Delta = +0.098$ at $n{=}5$, which **refutes the admission *ordering***: the arm with the largest admission-level change rose less than one with a smaller change, and the pooled decision reading does not predict it either.

What replicates is the negative on a coordinate-wise order statistic: that decision change costs no suppression by the frozen margin, at accuracy 0.698. Equivalence here is to this cell's own identity rung and not to a low absolute ASR, and the rise is positive in $5/5$ seeds (post hoc): small and consistent, not zero (App. D.2). A pre-registered intervention at Dirichlet $\alpha \in \{0.1, 1.0, 10.0\}$, none of them pooled with the $\alpha{=}0.5$ reading above, reproduces the rise at all three (App. D.5).

Two admissible estimands, opposite conclusions, one cell. This section's opening sign disagreement sits on this identical `coord_median` cell (matched aggregate displacement, decision change and clean accuracy, Table A3): the -0.273 is §6's outcome-gated ladder, whose adversary may attenuate, and the $+0.125$ the Mode S intervention above, pinning the adversary at $c{=}1$; both intervals are paired. The $n{=}5$ above is the same contrast at its frozen threshold's own n, differing in n alone. Collider bias is usually taught as attenuation; **here the confounded design does not lose power, it points the other way**: a criterion calibrated against that contrast is calibrated against the confound.

The two designs are not the same intervention: Λ_a moves in *both*, 0.081 against 0.033 ($2.4\times$), so neither is a clean statistic-only intervention and no p-value attaches to their difference. What the pair establishes, *within the rescaling class and defense family measured here* and by counterexample, not by law, is that **statistic preservation is not sufficient to identify attack-suppression preservation**, which is not the claim that a preserved statistic is uninformative (Prop. 3, App. C.3; Table A2).

Seven cells, not one (App. D.3). To separate a property of the design from one of this aggregator and attack, we ran seven cells. They disagree on five, across two attacks, three datasets and two CNN configurations, and **the reversal replicates**: the same cell on CIFAR-100 ($n{=}5$) reads -0.213 against $+0.071$, both intervals excluding zero. A mechanism frozen for when they disagree is refuted out of sample and withdrawn, so §4 says to intervene rather than consult a rule.

The oracle-free version, attempted three times. Norm clipping upstream on CIFAR-100 needed no mask; its premise failed, so that second-dataset replication did not carry (App. D.6). A second,

RFA into Krum at n=20, refuted our flatness prediction ($+0.297$), attenuation left open (App. D.7). A third closes it, pinning the adversarial share to 10^{-7}: a decision flip alone carries $+0.322$, so the negative is not established without an oracle (App. D.8).

6 What the Boundary Licenses, and What We Found

The regime, with its attacks and seed counts, is page 4's; the protocol is fixed in the supplement (§S6). The core evidence is one cell's seed-matched intervention, not a mean comparison across cells; the four kinds of answer are unequal and never pooled (Table A1).

The inadmissible design still told us something. Our earlier suites tried to validate ordering preservation by choosing a different downstream defense, the move Prop. 1 rules out. Varying only the *dose* at fixed d_2 and attack refutes that suite's pooled prediction that disturbance predicts each arm's rise (§D.1). **The fault is ours and is now measured**: fixing $\mathrm{mean}(c)$=1 leaves relative weights free, so a large spread mostly *dilutes* the poisoned client: attenuation, not statistic disturbance (§S7). Under the instrument, our one confirming arm was mostly attenuation too: what remains of reputation's rise tracks the small *admission* change $\Delta\Lambda_a$ rather than the large decision change (App. D.2).

The Krum dissociation replicates on a second dataset and on a second architecture family: pre-registered reruns on EMNIST-byclass/`simple_cnn` and CIFAR-10/`resnet18` change the decision in 0.643 and 0.556 of rounds, the adversary never admitted, and neither n=3 rise reaches the frozen 0.15. Neither 95% interval is a plain null: EMNIST's excludes zero, and `resnet18`'s is wider than the margin, so equivalence there is indeterminate, not established (supp. §S18, App. D.2).

The dissociation also replicates outside the rescaling family the theory covers: masking a random 0.8 fraction of each *benign* client's coordinates at fixed L_2 norm leaves adversarial updates untouched, so nothing we prove predicts it: the decision changes in 0.467 of rounds at admitted adversarial mass 0.000 and suppression does not follow, and neither pre-registered refuting branch fired (App. D.2.1). One mask is not the class of all non-rescalings.

Two arms support the account rather than carry it. `cos_krum`'s selection is bit-identical at every rung yet ASR falls 0.173 at the first accuracy-gate-passing rung, from an identity 0.584 that never suppresses (one-sided p=0.044, n=8, $5/8$ seeds), a pre-registered counterexample in the tested regime; and pinning payload share at fixed dispersion separates the theorem's two regimes, confirmed for `coord_median`, refuted for reputation (App. D.2, supp. §S24, Table A9).

Two pre-registered regimes ask whether the disagreement needs a non-adaptive attacker or ten clients (App. D.4). Against an adversary that adapts to d_2 and not d_1 on `coord_median` the two designs still part: D, the seed-paired difference of the Δs, is -0.324 $[-0.543, -0.105]$. At N=100, K=20 the disagreement reproduces and the sign reversal does not: D is -0.126 $[-0.217, -0.035]$, with both designs now reading the dose as protective and neither Δ powered at n=3.

The criterion fails exactly where Prop. 2 says it must: it discards the menu's lowest-ASR composition, so to find a menu's strongest, do not select on inherited suppression at all. Across the nine variants an evaluator could run, it sits at the high-precision, low-recall end (supp. §S22). **Its own label does not survive more seeds**, where C1 fails and it certifies nothing (supp. §S24.1, §S23).

7 Conclusion

Preserving a defense statistic does not suffice to identify preserved attack suppression, the reason structural: an outcome-gated comparison conditions on what it explains, which no data repairs (Prop. 1), and misses emergent suppression (Prop. 2). *What transports is one rule, and it is not our criterion*: hold the downstream mechanism and the threat fixed, and intervene upstream. Its cost: a diagnostic under a committed attack, not a guarantee against an adapting adversary (§6).

Neither a preserved decision nor a disturbed one determines suppression: it is a property of the intervention, not of the aggregator (App. D.8). **What changes for an evaluator**: *do not* read a preserved statistic or decision as evidence suppression survived (Table 2); *do not* select compared pairs on the outcome you mean to attribute; *do* run §4's four clauses and report where the chain broke. Both instruments read adversary identity, leaving open (App. H): **can admission (P4) be estimated without knowing which clients are malicious?**

ETHICS STATEMENT

This paper studies poisoning attacks in order to evaluate defenses against them, so its subject matter is dual-use. Four things bound the risk. Every attack we run is already published (model-scaling (Bagdasaryan et al., 2020), pixel-pattern backdoors (Chen et al., 2017), DBA (Xie et al., 2020)) and we introduce no new attack or new attack capability. All experiments are simulations on public benchmark datasets (CIFAR-10, CIFAR-100, EMNIST-byclass, UCI Spambase) with no human subjects, no personal data and no deployed system involved. Our headline result is defensive and cautionary rather than offensive: it tells practitioners that a preserved downstream statistic is *not* evidence that a composed defense stack still suppresses an attack, which removes a false assurance rather than supplying an attacker with one. The instruments we build (Mode S, the payload-share instrument, score-only Krum, the coordinate-wise magnitude control of §D.10, the boundary blend of §D.8) are measurement devices, not deployable defenses, for two different reasons: the first two require knowing which clients are malicious, and the other three need no such knowledge but reach inside the aggregator, score-only Krum because it then aggregates an update its own scoring stage never saw, the coordinate-wise control because it emits at each coordinate a value its own ordering stage never ranked, and the boundary blend because it rewrites the per-client coefficients on which the selection is scored; we state this at each use so that no reader mistakes them for a recommendation.

REPRODUCIBILITY STATEMENT

Every claim in the paper is tagged *general*, *formal* or *empirical* at the point it is made, and Table A2 states the scope conditions on each. For the formal claims, Propositions 1, 2 and 3 are stated in full in the body and proved in the appendix, and Proposition 11 and Theorem 9 are stated and proved in the appendix and referenced from the body. For the empirical claims: each pre-registered arm has a decision rule git-committed before the corresponding runs existed, with the commit hash recorded in the text and a runner that refuses to start unless its own recorded commit matches; every table cell is emitted by a generator that reads only frozen artifacts rather than being transcribed; sample sizes, seed ranges and per-seed values are given for every arm whose contrast we report, with per-seed detail in the appendix and the supplement, the two arms that close the aggregate-magnitude channel in supplement §S10; the one pre-registered arm that reports no contrast is the exception, and it says so at its own site (the supplement §S15): it computes no Δ at any n and reports no single leg, so there is no per-seed table for it to give; and the figures are recomputed per seed by the same scripts. Where a number was revised we report the revision and its cause rather than the corrected value alone, and where a derived condition turned out to be empty on our data we report it as empty rather than omitting it (§C.3). **No code or data package accompanies this submission**, and throughout this paper "the supplement" means the companion PDF only. What will be released on publication is specific rather than a general promise: every experiment runner and measurement script, the `results/` JSON artifacts every reported number is read from, the simulator itself, the generators that emit the figures, and the 38 pre-registration documents fixing the decision rules for every pre-registered arm, seed top-up and audit, which is the count of `experiments/pre_registration_*.md` and is checkable as one rather than taken on trust. Everything this paper's tables and figures are computed from runs from the repository root on Python 3 with `numpy`, `scipy`, `matplotlib`, `torch` and `torchvision`. The analysis entry points behind them are **idempotent**: they read the frozen artifacts and rewrite them byte-identically, so every number we report is a deterministic function of a frozen artifact rather than of a re-run, and no table cell can drift between builds. What the paper itself provides is the pre-registration commit hashes, the frozen decision rules they fix, and per-seed values for every arm we report a contrast from, which is what is needed to check the reported numbers against a reimplementation.

AI USE STATEMENT

In this work, we used generative AI tools as a writing and analysis assistant: for prose editing and revision of the manuscript, for LaTeX restructuring, and for writing verification and analysis scripts that check the manuscript against the frozen result artifacts. We did not use generative AI tools to generate research ideas, to formulate the hypotheses, to design the experiments or pre-registrations,

or to produce any experimental result, and no claim, number, figure or table in this paper was authored by an AI tool. We have reviewed all AI-assisted work: AI-assisted prose was checked against the underlying artifacts sentence by sentence, and AI-written analysis scripts were run against frozen inputs whose outputs are reproducible independently of the tool. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Fatima Z. Abacha, Sin G. Teo, Yuanxiang Wu, Lucas C. Cordeiro, and Mustafa A. Mustafa. FedSurrogate: Backdoor defense in federated learning via layer criticality and surrogate replacement. *arXiv preprint arXiv:2605.11122*, 2026.

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.

Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*. PMLR, 2019.

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*, 2017.

Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. FLTrust: Byzantine-robust federated learning via trust bootstrapping. In *Network and Distributed System Security Symposium (NDSS)*, 2021.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: A standardized adversarial robustness benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in byzantium. In *International Conference on Machine Learning*. PMLR, 2018.

Felix Elwert and Christopher Winship. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40:31–53, 2014.

Lucas Fenaux, Zheng Wang, Jacob Yan, Nathan Chung, and Florian Kerschbaum. Hammer and anvil: Toward a theory of backdoors in federated learning. *arXiv preprint arXiv:2509.08089*, 2025.

Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. The limitations of federated learning in sybil settings. In *23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, pp. 301–316, 2020.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.

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

Harsh Kasyap, Anil Kumar Pradhan, Ugur Ilker Atmaca, Graham Cormode, and Carsten Maple. PRoVeFL: Private robust and verifiable aggregation in federated learning. *arXiv preprint arXiv:2607.06612*, 2026.

Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. In *European Conference on Computer Vision (ECCV)*, 2018.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pp. 1273–1282. PMLR, 2017.

Arnab Nath and Harsh Kasyap. CQSA: Byzantine-robust clustered quantum secure aggregation in federated learning. *arXiv preprint arXiv:2602.22269*, 2026.

Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, and Thomas Schneider. FLAME: Taming backdoors in federated learning. In *USENIX Security Symposium*, 2022.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

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

Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *Network and Distributed System Security Symposium (NDSS)*, 2021.

Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H Brendan McMahan. Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*, 2019.

Nawrin Tabassum and Yanzhao Wu. STAR-FL: Secure federated learning with spatial-temporal analysis and robust aggregation. *arXiv preprint arXiv:2608.14861*, 2026.

Adrita Rahman Tory, ABM Shawkat Ali, Md Abu Layek, and Khondokar Fida Hasan. Privacy, robustness, and fairness trade-offs in federated intrusion detection: Geometric indistinguishability at the aggregation interface. *arXiv preprint arXiv:2609.03420*, 2026.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018.

Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jyyong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. DBA: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2020.

Chulin Xie, Minghao Chen, Pin-Yu Chen, and Bo Li. CRFL: Certifiably robust federated learning against backdoor attacks. In *International Conference on Machine Learning (ICML)*, 2021.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations (ICLR)*, 2018.

Jiahao Xu, Zikai Zhang, and Rui Hu. Achieving byzantine-resilient federated learning via layer-adaptive sparsified model aggregation. *arXiv preprint arXiv:2409.01435*, 2024.

Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter L. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael W. Mahoney, Prateek Mittal, Kannan Ramchandran, and Joseph E. Gonzalez. Neurotoxin: Durable backdoors in federated learning. In *International Conference on Machine Learning (ICML)*, 2022.

HOW TO READ THIS APPENDIX

If you are verifying rather than reading, three sections are the minimum, and one of the three is in the supplement. App. D carries the intervention evidence and every seed count behind it; the supplement §S24 prices what the screen buys against what it costs in recall; App. H lists claim by claim what is proved, what is merely supported, and what is withdrawn. Three is the minimum and not the whole: the per-section index below marks, one line each, which of the remaining sections become necessary if you doubt a particular claim, and which nothing in the body rests on at all.

Every pointer from the body into this appendix attaches to a claim the body already makes, and none defers a claim to be found here. That is a checked property rather than an intention: `experiments/measure_self_containment.py` resolves every pointer the body makes into this appendix, and fails if a target is missing, if the index below drops a section or lists one out of document order, or if a body sentence hands its content to the appendix instead of stating it. **So the appendix is where each of the body's claims keeps its provenance, not where any of them is kept.** Four groups, in the order a sceptical reader would want them. **(1) What licenses a claim**: the evidence tiers, the levels in full, the proofs, and the identification argument term by term (App. A–C.3). **(2) What the screen buys, priced**: all of it is in the supplement, the 42-pair menus with the prospective test and the metric-swap ablation (§S24, §S24.1, §S24.2) together with the frontier and the closed census (§S22, §S23). **(3) The intervention evidence**: the confounded ladder, the instrument that replaced it, seven cells, and the composition the screen cannot certify (App. D.2, D.3, E). **(4) What we do not claim**: limitations, the evidence-status ledger and notation (App. G, H, J), and the equivalence testing at the existing seed counts, which is in the supplement (§S18).

(5) What is auxiliary, and the one block that only looks it. Skippable without touching any claim in the body: the scale and architecture checks (the supplement §S20.1, §S20.2), which bound an external-validity statement and establish no boundary of their own; the composition and adaptive-attack details (the supplement §S19); notation (App. J); the per-cell out-of-sample detail behind §S24.4 (§S4); and the threshold-tolerance sweeps, whose result is that no published verdict moves (§S13). **The prospective suite (the supplement §S24.1) is not auxiliary, and we say so because its headline reads like a weakness**: it is the only out-of-sample prospective test here, it is where the criterion's shortfall is priced at $13/20$, and its FLTrust near-identity confound is the reason six of its arms cannot falsify anything. Cutting it would remove the evidence against us. **Nothing has been deleted to shorten this appendix**, and that is a considered choice rather than an oversight: every scope condition, limitation and withdrawal keeps exactly one home, since a disclosure relocated into a subordinate clause is one a later reader cannot find.

(6) One question asked and answered on its own: whether randomizing *which* composition is applied each round is itself costly at matched defense computation (the supplement §S17, with the composition audit's rubric and its 59 coded rows at the supplement §S16). Neither is a claim this appendix defers: the first is a new pre-registered experiment with its own controls, not a reinstatement of the randomized-menu comparison this paper withdraws, and the second is the artifact behind a rate §F already states and a claim it already withdraws. **They are in the supplement because nothing here rests on either**, which is the only ground on which anything was moved.

Per top-level section, since the groups above cross section boundaries. One line each, in document order; *skippable* means no claim in the body rests on it.

- **App. A**: the three tier words every empirical claim in the paper is tagged with, and Table A1's orthogonal strength axis (seed count, bit-identical, replicated). Read first if you want to know what any one sentence rests on.
- **App. B**: the federated-learning, poisoning and game-theoretic notation the rest of the appendix uses. Reference; skippable if §3's notation was enough.
- **App. C**: the five preservation levels stated in full (Def. 2), the screen C0–C3, the identification boundary, and every proof. Not skippable: Prop. 7 and Thm. 9 are stated and proved here, and the boundary is the paper's first contribution.
- **App. D**: the intervention evidence, from the outcome-gated ladder through Mode S to the seven cells. Not skippable: the sign reversal the abstract leads with is measured, disclosed and replicated here, and this is where its seed counts are stated.
- **App. E**: the composition the screen cannot certify, which is our own flagship pair. Not skippable, since it is the evidence against us.
- **App. F**: related work, and the closed citation census with its $6/59$ result. Reference.
- **App. G**: limitations L1–L6 in full, each with the claim it bounds. Not skippable.
- **App. H**: the evidence-status ledger, claim by claim, including what is withdrawn and what is merely supported. Not skippable if you are deciding what to believe.
- **App. I**: Figure 1(a)'s three numbers, which are three scopes rather than one experiment. Supporting; read it before quoting any of the three.
- **App. J**: notation. Reference, skippable.

Nine blocks are in the supplement rather than here, and this is where a reader looking for them finds them. The supplement §S16 carries the composition audit's frozen rubric, its retrieval frame and its 59 coded rows; §S17 answers one self-contained pre-registered question, whether randomizing *which* composition is applied each round is itself costly; §S18 is the equivalence testing at the existing seed counts, that is, where the frozen ± 0.15 margin binds and where it does not; §S19 holds the composition and adaptive-attack construction details; §S20 holds the scale and architecture checks, which bound an external-validity statement and establish no boundary of their own; §S21 states the one use we claim for the screen, evaluation prioritization, together with the two designs an evaluator whose object is discovery should run instead; §S22 prices that use over the nine variants an evaluator could run, precision against recall against evaluation cost; §S24 is the subordinate boundary study that prices the screen itself, carrying the 42-pair development and held-out menus, the prospective suite's $13/20$, the metric-swap ablation, the CIFAR-100 composition menu and the out-of-sample PASS battery; and §S23 closes the region Prop. 11(a) settles algebraically, which is where the one falsifiable prospective positive sits. **What moved them is that no claim in the body is stated only there, not that any of them is unimportant**, and all nine keep every number they had. Five are *skippable* in the sense above. **The equivalence testing, the boundary study, the frontier and the census are not**: read the first if you doubt the margin, since it is where each arm's binding m^* and the one-sided bound behind every non-increase are computed, the second for what the screen buys and the two boundary conditions on it, and the other two for what the screen costs and for the single prospective positive the paper claims.

A THE THREE TIERS OF EVIDENCE

Three tiers of evidence, of very unequal weight, never pooled. Every empirical claim about the preservation hierarchy or the criterion is tagged with one of the three words of Table A1 at the point it is made, so a reader can look up what licenses it instead of reconstructing it from hedges.

Table A1: **The three tiers of evidence, of very unequal weight and never pooled**, and the *orthogonal* strength axis. Every empirical claim in the paper carries one of the three tier words at the point it is made; the tiers are never combined into a single verdict. The lower panel is a second, independent axis: what a claim's *design* licenses (upper panel) and how strongly the *statistics* support it are different questions, and a Causal-tier claim can still sit in the weakest band.

tier	design	what it licenses
Causal	within-defense intervention at fixed d_2, attack and C1 input: Mode S, Mode A (§D.2)	statements about the *mechanism*, scoped to these cells
Replication	EMNIST-byclass, a second dataset and a differently configured CNN ($\Delta{=}-0.017$, $n{=}3$); `resnet18`, a second architecture family with architecture varying alone (-0.041, $n{=}3$); `coord_median`, a second invariance class ($+0.098$, $n{=}5$); and CIFAR-100, a third dataset, where the sign disagreement of §D.3 replicates (-0.213 against $+0.071$, $n{=}5$)	"replicated on a further dataset or architecture, one arm each", **never** "generalizes"; one composition suite now leaves CIFAR-10, inside wave 1 only and at the coverage §S24.3 states, with wave 2's 24 held-out pairs deliberately unrun
Screening	the outcome-gated 42-pair suite	**evaluation prioritization only**; by Proposition 7 it cannot identify the mechanism's contribution at *any* sample size
Statistical strength, an axis orthogonal to the tiers above. Read this before quoting any number.		
Strong	the invariance results (Prop. 11, Thm. 9) and the identification boundary (Prop. 7), which are *proofs* and carry no n; `cos_krum`'s exact score-only identity, which is bit-identical rather than estimated; the pre-registered Mode S intervention, frozen before any of its ASR existed; and the *direction* of the negative, which replicates on a second dataset, on a second architecture family and outside the rescaling family	
Moderate	the ASR practical-equivalence results at $n{=}5$, one of which was topped up to $n{=}20$ under a pre-registered clause and held ($\Delta{=}-0.010$); at $n{=}5$ these are TOST verdicts against a frozen margin, not tight estimates	
Weak, exploratory	any *cross-arm quantitative ordering*, which the admission reading predicted and which Table 2 refutes; and the exact (P4)→(P5) *magnitudes*, for which we claim the implication and not the size. Nothing in the paper's argument rests on this band	

B PRELIMINARIES: FEDERATED LEARNING, POISONING AND GAME-THEORETIC NOTATION

Federated learning. N clients each hold a local dataset $\mathcal{D}_i \sim P_i$; the server learns $w^* = \arg\min_w \sum_i \frac{|\mathcal{D}_i|}{|\mathcal{D}|} F_i(w)$. In FedAvg (McMahan et al., 2017), each round the server broadcasts w^t, selected clients run local SGD, and the server aggregates $w^{t+1} = w^t + \frac{1}{|S_t|} \sum_{i \in S_t} \Delta_i^t$.

Poisoning and robust aggregation. An adversary controlling a fraction f of clients can flip labels (Chen et al., 2017), inject backdoor triggers (Bagdasaryan et al., 2020; Xie et al., 2020), or scale malicious updates to dominate aggregation (Bagdasaryan et al., 2020). The server may replace FedAvg with a Byzantine-robust rule: Krum (Blanchard et al., 2017), Trimmed Mean, Coordinate-wise Median, Norm Clipping (Sun et al., 2019), or RFA (geometric median) (Pillutla et al., 2022), each imposing an accuracy/robustness trade-off.

Game theory. A normal-form game is $G = (\mathcal{N}, \{\mathcal{S}_i\}, \{u_i\})$; an equilibrium is *pure* when every player's strategy puts all its mass on one action and *mixed* otherwise. **No claim of this paper is game-theoretic**: this is needed only to read the residual scale check of the supplement §S20.1, and the type-space and Bayes-Nash machinery an earlier version used is not needed here.

C THE SCREEN, THE IDENTIFICATION BOUNDARY, AND THE PROOFS

One level, two names, and which is primary. The title and the body call this level *attack-suppression preservation*, which is (P5) of Definition 1. This section and the proofs below call the same level *mechanism preservation*, because here it is stated as an invariance property of the upstream transformation rather than as an outcome about the attack. **The two names denote the same level**, the body's is the primary one, and nothing below depends on the difference.

The paper's central practical question is *which* defense pairs compose well under deterministic joint application. Specifically, the criterion screens for *mechanism-preserving composition*: when one constituent defense is already individually effective (C1), can the other defense be added without breaking that suppression? This is the operationally relevant question (the criterion is not a general characterization of all composable pairs, but a sufficient condition for *mechanism preservation* among pairs where at least one component has known effectiveness), never a sufficient condition for low ASR. We answer it first, with a checkable criterion, and treat the alternative, temporal mixing, as outside the scope of this paper. **But the practical question is not what the paper establishes**: what it establishes is that preservation of the statistic the screen checks is not causally tied to the security the screen predicts, so the criterion below should be read as the application the supplement's §S21 scopes it to (an ordering over an evaluation budget) and not as a deployment recommendation.

Observe that each robust aggregation defense d operates through an *admission signal* $\sigma(u)$ applied to client update u: it accepts, clips, or down-weights u based on whether $\sigma(u)$ exceeds a threshold. Crucially, the admission signal varies by defense family:

- **Weighting defenses** (reputation, FoolsGold): $\sigma(u) =$ distance from consensus (ℓ_2) or pairwise similarity.
- **Clipping defenses** (NormClip): $\sigma(u) = \|u\|_2$.
- **Rank-based defenses** (TrimmedMean, CoordMedian): $\sigma_k(u) =$ coordinate-wise rank of u_k.
- **Geometric defenses** (RFA): $\sigma(u) =$ distance from geometric median.

When two defenses are applied sequentially (d1 transforms the updates, d2 aggregates the result), d1's transformation may *corrupt* d2's signal or vice versa.

Criterion 4 (Signal-Preservation Composability: an evaluation prioritizer, never a security screen (L2))**.** *Let d_1 and d_2 be defense mechanisms. The following sufficient conditions guarantee **preservation of d_2's suppression mechanism under composition, conditional on the standalone suppression C1 asserts**. They are **not** an independent guarantee of low composition ASR: Theorem 9, which proves them for one class, bounds no ASR. Stated once so the scope cannot be misread:*

$$C1 \wedge C2 \wedge C3 \;\Rightarrow\; d_2\text{'s suppression mechanism survives composition, levels (P1)–(P3) of Def. 2.}$$
$$C1 \wedge C2 \wedge C3 \;\not\Rightarrow\; \text{low composition ASR.}$$

*The conditions are of **two different kinds**, and we group them accordingly, because conflating an empirical screen with a mechanistic property is the main source of confusion about what the criterion claims.*

Stage 0: applicability. *Before either kind of condition is informative, the question the criterion asks has to be live.*

(C0) The downstream mechanism is load-bearing. $\mathrm{ASR}(d_1, a) \geq 0.5$ *for at least one committed attack a. C2 and C3 concern preservation of d_2's admission signal, so they say nothing when d_1 alone*

already suppresses everything: the composition inherits that suppression, destroying d_2's signal costs nothing, and a correct prediction would be correct for a reason the criterion did not supply. The asymmetry is deliberate: a strong d_2 is the intended case, since asking whether upstream processing breaks an effective downstream defense is the framework's whole question, whereas a strong d_1 makes the question moot. C0 was ***formulated post-hoc*** *and its provenance is in Remark 6.*

The empirical screening condition. *This one is measured rather than read off definitions, and it is the gate the identification boundary is about.*

(C1) Standalone suppression. *Some constituent suppresses the committed attack on its own:* $\min_{d\in\{d_1,d_2\}} \mathrm{ASR}(d,a) < 0.5$*, quantified* per attack. *With C0 this forces the suppressing constituent to be* d_2 on the attack a^* that witnesses C0*, and there only (Lemma 1), which is exactly what makes the gate a descendant of the quantity the screen is trying to estimate (Proposition 7).* ***On a committed attack that*** d_1 ***happens to suppress, the constituent doing the work can be*** d_1 ***instead****, because C0 is existential over attacks while C1 is universal over them; both of our certified pairs are of that kind, and §S24 reports what that costs.*

Mechanism-preservation conditions. *These are what Theorem 9 proves for the weighting→aggregator class, and they are determined by d_1's and d_2's definitions alone, no experiment required.* ***What we recommend is an*** **evaluation-prioritization criterion*****, C1∧C2, because that is the whole of what our evidence supports.*** *We do call it a* screen *throughout, in the ordinary sense of a cheap test run before an expensive one.* ***We disclaim predictive validity explicitly, here at first use****: the criterion orders candidate compositions for evaluation, and it does* not *predict whether a composition is secure. The pooled two-channel model over* 26 *cells is unconfirmed (R^2=0.239), and our own pre-registered replication* ***refuted*** *the cross-arm admission ordering the criterion would need in order to predict (§D.2); recall is* 40% *(Proposition 8), so the negative direction carries no information either.* ***It is never a security screen, and a PASS is an evaluation ordering rather than a deployment decision*** *(Table A2). C3 is retained as a* ***margin diagnostic*** *rather than a gate, because on our data it never binds on a pair C1∧C2 already admits and shows no incremental predictive value (supplement §S21).* The pre-registered labels of §S24 were frozen with C3 in the gate and are not relabelled*; the demotion is a recommendation going forward, supported by a post-hoc factorial ablation, and it changes no reported number: C1∧C2 selects exactly the same two in-distribution pairs.*

(C2) Signal preservation. *d_1's per-client transformation T_1 does not destroy d_2's admission signal: $\sigma_2(T_1(u_a)) \approx \sigma_2(u_a)$ for malicious updates u_a.*

(C3) Boundary preservation. *d_1's transformation does not compress the adversarial–benign separation below d_2's discrimination threshold. Formally: if d_2 suppresses a by detecting $\tau_{d_2}(u_a) > \theta$ (triggering trimming/rejection), then $\tau_{d_2}(T_1(u_a)) > \theta$ must hold, i.e., d_1 must not map adversarial updates from above to below d_2's rejection boundary.*

When C1 holds but C2 fails (signal destruction), composition is no better than d_2 alone. When C1 and C2 hold but C3 fails (boundary compression), d_2 retains its admission signal direction but its threshold no longer fires: the attack passes because d_1's rescaling places adversarial magnitudes within d_2's acceptance region.

Writing C2 as preservation of d_2's admission signal elides a distinction our own experiments later force, so we name the levels before relying on the condition. They are the paper's main conceptual instrument, and every claim below is stated at one of them.

Definition 2 (Preservation levels)**.** *For an upstream transform T_1 and a downstream defense d_2 reading admission signal σ_2:*

- *(P1)* ***Value invariance:*** *$\sigma_2(T_1(U)) = \sigma_2(U)$.*
- *(P2)* ***Ordering invariance:*** *$\sigma_2(T_1(U)) = \varphi(\sigma_2(U))$ for some φ preserving the ordering σ_2 induces on clients.* This is the level C2 is stated at *(made explicit in* (2)*); the distinction from (P1) matters, because exact values are sufficient but never necessary.*
- *(P3)* ***Decision invariance:*** *the client or coordinate d_2 actually selects is unchanged.*
- *(P4)* ***Adversarial contribution (admission) invariance:*** *the* adversarial *input d_2 admits into its aggregate is unchanged.* Admission *is the term we use throughout for whether adversarial input survives d_2, and it is a* ***level, not one measurable quantity****: for a selector it is the indicator that an adversarial update is selected at all; for a weighted averager, the adversarial share of the weight mass; for a coordinate-wise order statistic, the adversarial share of selected coordinates, which is why Table A8's influence column is aggregator-specific by necessity. The committed pre-registrations say "admission", so that word is retained in every frozen rule name and no scored quantity is relabelled.*
- *(P5)* ***Attack-suppression preservation:*** *ASR stays below threshold.*

(P1)⇒(P2)⇒(P3) hold, and hold by definition*: equal values induce the same ordering, and a rule reading only that ordering cannot decide differently.* ***These three are logical implications under the definitions***

above, not empirical claims about how defenses behave**, and nothing is asserted by proving them: their content is that they stop at (P3). **Below (P3) the chain breaks, and it breaks in both directions.** Both breaks are measured.* Holding *decision* and *admission invariance does not deliver suppression: `cos_krum`'s selection is bit-identical at every dose and its admitted adversarial mass is* 0.000 *throughout, yet its ASR falls* 0.173 *by the first rung alone through the magnitude channel, at* ρ=2.72*, where clean accuracy moves only* $0.599 \rightarrow 0.567$ *and our accuracy gate passes (paired* t=-1.976*, one-sided* p=0.044*, falling in* $5/8$ *seeds). Across the whole ladder it falls* 0.425*, but that endpoint sits below our own accuracy floor, so **the gate-passing rung is the witness we rest on** (§D.2).* Losing *decision invariance does not cost suppression: Krum's selection flips in up to* 80% *of rounds under an intervention that pins the adversarial mass, and its ASR stays inside the pre-registered practical-equivalence margin (*$|\Delta|$=$0.026 < 0.15$*) (§D.2). A criterion stated at (P1)–(P2) therefore constrains security in* neither *direction. **(P4) is where the security content sits, but we do not claim it is sufficient for (P5), and it is a **level** ***rather than one universal scalar security quantity.** It is instantiated per aggregator by the measure that aggregator's own rule exposes, Krum's and `cos_krum`'s selected-client indicator, `coord_median`'s share of selected coordinates, reputation's adversarial weight share (`experiments/build_channel_table.py`), so **(P4) names what to measure, not a single number comparable across defenses**. It is the first level that refers to* adversarial influence on the aggregate *rather than to the downstream defense's internal decision, and `cos_krum` is our own witness that (P4)* $\not\Rightarrow$ *(P5): admission identically* 0.000 *at every rung while suppression moves. **So the hierarchy answers "which level does a security claim have to be stated at?" and not "which level predicts ASR?"**: no level in it is offered as a predictor.*

Remark 5 (Which level C2 asserts, and which one is load-bearing)**.** *At* $\rho = 54.6$ *the* uncontrolled *dose ladder of §D, whose adversary is free to attenuate, changes Krum's* decision *in* 86.7% *of rounds while changing its* admission *of adversarial input in* 13.3%*, and Krum's suppression does not deteriorate at all. **Those two figures are the uncontrolled arm's and no other's**: Mode S, which pins the adversary at the same rung, reads* 0.733 *and* 0.000 *instead (Table 2, `dose|krum|2.0|*` against `doseS|krum|2.0|*` in `results/admission_measurement.json`), which is why Table 2 can say admission never moves in any round of any arm it reports. A selection that flips from one benign client to another disturbs the statistic maximally and the mechanism not at all. **Terminology, once and for all: "**mechanism**" always means the downstream defense's** discriminative operation **(what separates adversarial from benign updates) and never its security behaviour, since that implication is precisely what this paper refutes.** We reserve "mechanism preservation" for the formal property Theorem 9 establishes, saying* statistic *or* decision *preservation where only the weaker property holds. C2 above is the (P2) condition (which is what makes it checkable from definitions alone) but the load-bearing reading is (P4). **This is the paper's central tension: our theorem operates at (P1)–(P3), while the experiments indicate security behaviour is decided at (P4)**; §D.2 puts the two against each other in an experiment where they predict opposite outcomes, and Figure A7 is that argument in one picture.*

Remark 6 (C0's provenance: an applicability condition formulated post-hoc)**.** *C0 sits in Criterion 4 as a stage-0 applicability condition because Lemma 1 is stated over C0∧C1, but **it does not have the same evidential status as C1–C3**, and we record that here rather than let its promotion obscure it. It is **vacuous on the 7-defense menu of §S24**, where the strongest single defense reaches only* 0.519*, and we formulated it only after the prospective suite of §S24.1 introduced a defense for which it binds. It is a lesson from prospective evaluation, applied there retrospectively; it is* not *part of the criterion validated on the 42 pairs, and none of the accuracy figures in §S24 depend on it.*

A scope condition on the criterion itself, not a limitation of our runs. C1 and every prediction built on it are defined against a *committed* adversary, one whose attack is fixed before the composition is chosen. An adversary who observes the composition can re-target whichever channel the screen certifies, so **what the screen bounds is *inherited* robustness under a fixed attack, never robustness against an adaptive one**. Our single whitebox composition-aware stress test is consistent with that reading ($0.064 \rightarrow 0.157$ seed-matched at n=30 on the pixel arm the attack is built from, $2.4\times$; §E).

C.1 The testability boundary

C0 and C1 together constrain what any validation of this criterion can look like. We did not anticipate this; it turns out to explain the central weakness of our own prospective suite (§S24.1, L4), and (once stated in general form) to *derive* the class of within-defense designs in which C2's contribution is identifiable (§D). Its content is a boundary on one class of screens rather than on evaluation at large, and naming the class is what makes it *exact* rather than merely negative: **under inherited-effect gating**, cross-defense contrasts cannot identify the mechanism's contribution, because the gate conditions on a descendant of both the mechanism and the outcome; lift that one condition and the same contrasts identify it. What we offer as the transferable contribution is that characterization together with **the admissible design it derives**, set out below and run in §D.2.

Lemma 1 (Testability boundary)**.** *Let* $(d_1 \rightarrow d_2)$ *satisfy C0 and C1, and let* a^* *be a committed attack witnessing C0, so* $\mathrm{ASR}(d_1, a^*) \geq 0.5$*. Then* $\mathrm{ASR}(d_2, a^*) < 0.5$*: the downstream defense suppresses* a^* on its own.

Table A2: **Scope of claims: three kinds, the oracle disclosure that qualifies the empirical one, and the one place each scope condition is stated. Claims do not transfer upward**: the formal guarantee does not certify the general framing, and the empirical figures certify neither level above them. This table is the conservation record for every caveat in the paper; §G restates the same conditions individually as (L1)–(L6).

kind	content	status, and what it does *not* license
General	An evaluation gated on the outcome it predicts cannot identify the mechanism's contribution, and admissible designs are within-defense; an inherited-suppression gate cannot certify emergent synergy (Prop. 7, Prop. 8).	**Proved, and not specific** to FL, to backdoors, or to our conditions. Licenses no claim that any composition *is* secure, and none that mechanism preservation implies suppression. Recall is bounded, not achieved: the bound is 80% from $\lvert\mathcal{E}\rvert/\lvert\mathcal{S}\rvert{=}1/5$, and 40% is what we measure (L2).
Formal	Mechanism preservation as *invariance* under **positive per-client rescalings** $T(u_i) = c_i u_i$, $c_i > 0$, into specific rank and geometric aggregators, with explicit margins (Thm. 9, Lem. 2, Cor. 4, Prop. 11, Prop. 10).	**Proved on that class only**: a *laboratory* for composition, not a model of FL defenses in general. A d_1 that alters participation, carries cross-round state, mixes clients, or aggregates first admits no such T and is **covered by nothing here**. **Bounds no ASR.** C0 was formulated post-hoc and applied retrospectively to arms frozen before it (Rem. 6, L4). Regime A's condition is not operationally checkable, $r^\star$ not being a stable per-attack constant (L5). Cor. 4's bound goes vacuous as $m \to 1$, which heterogeneity realizes.
Empirical	§D.2's interventions and the supplement §S24's 42-pair menu: CIFAR-10, CifarCNN/ResNet18, $N \in \{10, 100\}$, $K \in \{5, 10\}$, $f \in [0.05, 0.6]$, Dirichlet $\alpha \in \{0.1, 0.5\}$, **star topology**, three attack families; one arm on EMNIST-byclass/`simple_cnn`, two on CIFAR-100/`cifar_cnn` and one on CIFAR-10/`resnet18` varying architecture alone.	**Measured** at $n{=}5$ seeds per cell ($n{=}8$ for `cos_krum`, $n{=}3$ frozen and $n{=}5$ pooled post-hoc on EMNIST-byclass, $n{=}3$ on the `resnet18` arm, $n{=}30$ for the FG→RFA characterization), all fixed before the runs, so intervals are wide and **no null is read as flatness** without a pre-registered practical-equivalence margin. ASR is thresholded at 0.5, the level at which the trigger fires on the majority of triggered inputs; a cell below the 0.35 clean-accuracy floor is *uninterpretable*, not suppressed. Dataset transfer is tested on *three* arms and no more: the flagship Mode-S arm on EMNIST-byclass/`simple_cnn`; the sign-reversal cell of Table A10 on CIFAR-100, where the reversal replicates, the confounded leg falling and the controlled leg rising (§D.3); and wave 1 of the 42-pair composition menu on CIFAR-100 at the coverage §S24.3 states, the only one of the three that composes two defenses. All three are "replicated on a further dataset", **never** "generalizes" (L1). The EMNIST arm is a *channel* replication and carries no screen: no single defense measured on EMNIST-byclass suppresses the committed pixel backdoor alone (minimum 0.609; the menu measured there excludes `foolsgold`, `reputation` and the cosine mirrors), so C1's per-attack precondition fails on that arm and the strategy comparison is **out of scope over the measured menu** rather than merely untested (L1). **"Refuted"** always means a pre-registered *directional* hypothesis was rejected against a rule frozen for it, **never** that a mathematical claim was disproved. The $\gamma{\approx}0.66$ boundary is architecture-conditional (L1); the strongest development-set figures are *fit*, not prediction (L2).
Both instruments are oracles	Mode S and Mode A build c_i from adversary identity, as does Δ_c (§D.2, supplement §S11).	They **identify a causal channel and are not screens**: no deployed defense knows which clients are adversarial, and estimating (P4) without that knowledge is the open problem of §H. Admission is *a direction to measure*, not a validated predictor: the pooled two-channel model is unconfirmed over 26 cells ($R^2{=}0.239$) and its cross-arm ordering was refuted by our own replication (L6). C1 and everything built on it assume a **committed** adversary, so what is bounded is *inherited* robustness under a fixed attack, **never adaptive robustness**.

Proof. C1 at a^* gives $\min_{d\in\{d_1,d_2\}} \mathrm{ASR}(d, a^*) < 0.5$. By C0 the minimum is not attained at d_1, hence it is attained at d_2. □

Nothing in the argument is specific to federated learning, to backdoors, or to our defense menu. Lemma 1 is **the enabling observation for the FL instance**; the substantive claim is the general one, and we state it as such.

Proposition 7 (Why constituent-effect contrasts cannot identify composition effects under inherited-effect gating; Proposition 1 of the body, restated with proof)**.** *Let a screen certify $(d_1 \to d_2)$ against attack a only if both (i) an **eligibility gate** $E(d_1, d_2, a) := \min_{d\in\{d_1,d_2\}} M(d, a) < \theta$ holds, where M and θ are the* same *outcome metric and threshold the screen predicts for the composition, and (ii) a **mechanism condition** $P(d_1, d_2)$ holds, determined by the two defenses' definitions alone. Suppose the screen is applied only where the downstream mechanism is load-bearing, $M(d_1, a) \geq \theta$. Consider estimating P's contribution to $M(d_1 \to d_2, a)$ from contrasts across a defense menu $\mathcal{D}$. Then every certified pair satisfies $M(d_2, a) < \theta$, so of the four cells of $\{P, \neg P\} \times \{M(d_2, a) < \theta,\ M(d_2, a) \geq \theta\}$ the certified set occupies at most two, and the cell needed to separate P from d_2's standalone effectiveness is empty by the gate, for every $\mathcal{D}$. No contrast varying d_2 identifies P's contribution. Conversely, lift the gate's reading of M and the same contrasts identify it, so on this class of screens the boundary is exact. Identification therefore requires holding d_2, a and therefore E fixed while intervening on the upstream transform:* ***the admissible designs are within-defense.***

Proof. The load-bearing condition excludes d_1 from attaining the minimum in (i), so certification forces $M(d_2, a) < \theta$; both cells with $M(d_2, a) \geq \theta$ are therefore unpopulated whatever $\mathcal{D}$ is, and P is never observed against a downstream defense that fails on its own. Any estimator of P's contribution from such contrasts is unidentified.

For the converse, drop (i) and let the rule deciding which pairs are run be a function of the design alone rather than of M. Then $\{M(d_2, a) < \theta\}$ and $\{M(d_2, a) \geq \theta\}$ are both assignable, and on a menu realizing both at each value of P the conditional means $\mathbb{E}[M(d_1 \to d_2, a) \mid P, M(d_2, a)]$ are all estimable, so P's contribution is identified by differencing in P at fixed $M(d_2, a)$. The gate's reading of M is thus the *only* obstruction within this class of screens, which is what makes the boundary exact rather than merely negative. Two conditions do the work and are worth separating, because a design can satisfy either alone: selection that does not read the outcome to be explained, and support at both levels of $M(d_2, a)$ for each level of P. Corollary 1 is the instance of the converse this paper runs, and it obtains the second condition by construction rather than by luck, since it fixes d_2 and runs every assigned arm including the control. □

Corollary 1 (The design that restores identification)**.** *Fix d_2 and a. Let $\mathcal{T}$ be a set of upstream transforms with* $\mathrm{id} \in \mathcal{T}$*, let the experimenter assign $T \in \mathcal{T}$ by a rule that is a function of the design alone, and let $Y := M(T \to d_2, a)$ be run and recorded for every assigned T, the control $T{=}\mathrm{id}$ included. Then for every $t \in \mathcal{T}$ the contrast*

$$\tau(t) \ := \ \mathbb{E}\big[Y \mid \mathrm{do}(T{=}t)\big] - \mathbb{E}\big[Y \mid \mathrm{do}(T{=}\mathrm{id})\big]$$

is identified, and equals the difference of the two arms' means. ***Its scope is narrow, and stating it is half the point.*** *$\tau(t)$ is the total effect of the assigned transform at that one (d_2, a): it separates into a statistic channel and an admission channel only when one of them is pinned by construction (§S24.2), and it transfers to no other downstream defense, no other attack, and no adversary that adapts to T.*

Proof. Assignment is a function of the design, so T is independent of the potential outcomes $\{Y(t)\}_{t\in\mathcal{T}}$, and every arm is run, so $\Pr(T{=}t) > 0$ on $\mathcal{T}$: ignorability and positivity hold by construction rather than by assumption, and $\mathbb{E}[Y \mid \mathrm{do}(T{=}t)] = \mathbb{E}[Y \mid T{=}t]$. The cell that Proposition 7 empties is emptied by conditioning on E, a descendant of Y; here nothing is conditioned on, and the control arm is run whatever its outcome. The scope clause is a restatement of $\mathrm{do}(T{=}t)$: the intervention is on the transform, not on one channel of its effect. □

This is the elementary converse of the proposition and we claim no depth for it. Its content is that the repair is **available and cheap, and is not what cross-defense comparison does**, including in our own first two suites: §5 runs it, and Table A3 is the CIFAR-10 cell where both designs exist, where they agree numerically at $\kappa{=}0$ and disagree in *sign* at the top rung; §D.3 reports a second such cell, on CIFAR-100, where the disagreement replicates.

A second boundary: what such a screen cannot find. The gate that limits what a screen can be *validated* on also limits what it can *cover*, which converts the most obvious objection to our criterion, that its recall is poor, into a scope result.

Proposition 8 (Inherited-suppression screens cannot certify emergent synergy)**.** *$(d_1 \to d_2)$'s suppression of a is **emergent** if $M(d_1 \to d_2, a) < \theta$ while $\min_{d\in\{d_1,d_2\}} M(d, a) \geq \theta$. No emergent composition satisfies the eligibility gate of Proposition 7. Hence, writing $\mathcal{S}$ for the compositions with $M < \theta$, $\mathcal{E} \subseteq \mathcal{S}$ for the emergent*

Table A3: **The confounded ladder against the Mode-S instrument, on the CIFAR-10 arm that has both**; §D.3 adds a CIFAR-100 arm that has both and reproduces the sign disagreement. Same downstream defense, same attack, same seeds, same nominal dose, and the two designs share their κ=0 cell *numerically* over all 20 seeds (mean and worst per-seed gap 0.00e+00, asserted at 10^{-12}). **Each row carries its own n, because this table is mixed.** A pre-registered top-up (`experiments/pre_registration_reversal_seed_topup.md`) added seeds 47–61 to both legs at the two *endpoint* rungs, so the outcome rows are n=20; the mechanism rows are the frozen n=5 measurement on seeds 42–46, which is the only n they were ever recorded at. At the top rung the two designs are matched on aggregate displacement and decision change (at n=5) and on mean clean accuracy (at n=20), so **the ladder's fall is not an accuracy collapse**; they differ chiefly in the **adversarial coefficient channel**, and the ASR effect **changes sign**. **The one non-match, disclosed:** aggregate adversarial influence Λ_a moves in *both* designs, 0.081 against 0.033 (2.4×), so neither column is a clean statistic-only intervention, which is why the contrast is "differ chiefly in the coefficient channel", never "only in it". *Vocabulary:* this row is the $\Delta\Lambda_a$ column of the channel table (the adversarial share of selected coordinates, the same field in the same artifact) and **not** its Δ adm. column, which counts rounds where the *support* of the adversarial mass changes and reads 0.000 on every row: what moves here is how much weight the adversary's direction carries, not whether it enters at all. **Which p this is, since the top-up changed it.** Each effect's p is the one-sided *endpoint* paired t at that effect's own n=20, on the same per-seed differences as the interval beneath it. It is *not* the four-rung Jonckheere–Terpstra trend test: the interior rungs were not topped up, so that test stays at n=5 where it was frozen, and its values ($p_{\downarrow}$=0.0015, $p_{\uparrow}$=0.069, both n=5) are reported in §D.3 rather than mixed into this column. This is a two-column contrast on one arm: it locates where the two designs differ, and attaches no p-value to the difference between the columns. Every cell recomputed from per-seed and per-round rows by `build_comparability_table.py`, which also asserts that the published n=5 pair, −0.272 against +0.098, still reproduces on the frozen seed block; nothing here is retyped.

	Confounded ladder `dose_kappa`	Instrument (Mode S) `doseS_kappa`
downstream d_2	`coord_median`	`coord_median`
attack	committed pixel	committed pixel
seeds	42–61	42–61
nominal dose	κ 0.0→2.0, ρ 1→54.6	κ 0.0→2.0, ρ 1→54.6
benign coefficients	mean 1 over *all* clients	mean 1 over *benign* clients
adversarial coefficients	drawn from the same spread	pinned $c = 1.0$ exactly
adv. coefficient share (n=5)	0.267 → **0.187**	0.266667, constant
Δ_c (n=5)	+0.000 → **−0.336**	0 by construction (read back 9e−08)
aggregate displacement (n=5)	0.000 → 0.722	0.000 → 0.670
decision change (n=5)	0.000 → 0.487	0.000 → 0.482
adversarial influence $\Delta\Lambda_a$ (n=5)	0.000 → **0.081**	0.000 → **0.033**
mean clean accuracy (n=20)	0.761 → 0.692	0.761 → 0.703
ASR (n=20)	0.478 → 0.205	0.478 → 0.603
ASR effect (n=20)	**−0.273** ($p_{\downarrow} = 6.4\times10^{-9}$)	**+0.125** ($p_{\uparrow} = 1.9\times10^{-8}$)
paired 95% interval	[−0.334, −0.213]	[+0.095, +0.155]

ones and $\mathcal{S} \setminus \mathcal{E}$ for the inherited ones, any such screen's recall obeys recall $\leq 1 - |\mathcal{E}|/|\mathcal{S}|$ *on every menu and every attack.*

Proof. Emergence is the negation of the gate, so each member of $\mathcal{E}$ is rejected while belonging to $\mathcal{S}$: a false negative. The bound is the resulting recall. □

Corollary 2 (The cap reaches 0, so it is not a reassurance)**.** *If $\mathcal{E} = \mathcal{S}$, every screen carrying that eligibility gate has recall exactly 0 on that menu, at every sample size: it certifies no low-M composition at all. And $1 - |\mathcal{E}|/|\mathcal{S}|$ is a quantity the screen's own user cannot report, since $|\mathcal{E}|$ is measurable only by running the compositions the gate rejects.*

Proof. Every member of $\mathcal{S}$ is then emergent, hence rejected by Proposition 8, so the certified set meets $\mathcal{S}$ in $\emptyset$ and recall is 0; recall is also non-negative, so the bound is attained. For the second claim, deciding membership in $\mathcal{E}$ requires $M(d_1 \to d_2, a)$ for a pair with $\min_d M(d, a) \geq \theta$, and that is a pair the gate rejects. □

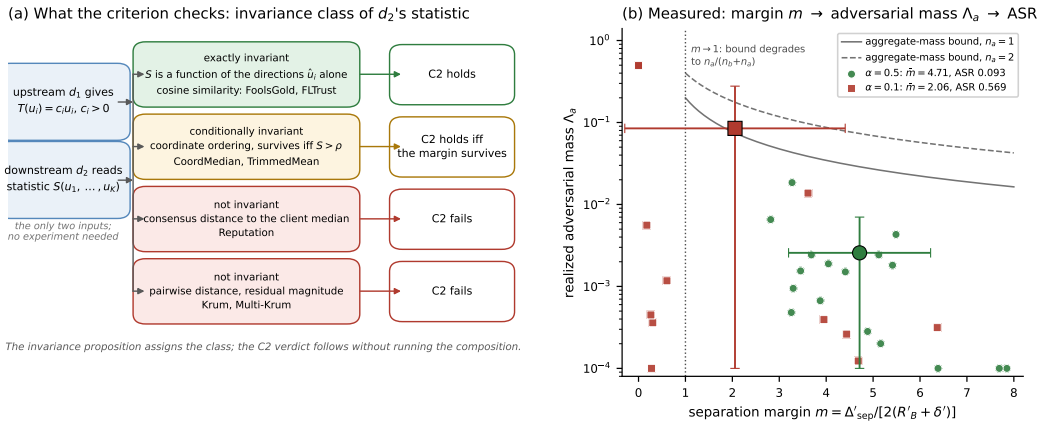

Figure A1: **The criterion is a mechanism test, and the mechanism is measurable.** **(a)** What C2 checks: the upstream transform T and the statistic S that d_2 reads determine an *invariance class* (Proposition 11), and the class determines the verdict: no composition needs to be run. This is why norm clipping does not blind FoolsGold but does blind Reputation. **(b)** The measured chain for FG→RFA, per round, on the transformed points RFA actually receives. Each marker is one round with $w_a > 0$ (rounds where FoolsGold zeroes every adversary are Lemma 2 and carry no payload); large markers are mean ± s.d. Corollary 4's bound holds in every round, and the α=0.1 collapse is visible as a collapse in the theorem's own quantity: m falls from 4.71 ± 1.51 (n=17 rounds) to 2.06 ± 2.35 (n=12), with several rounds at m<1 where the bound degrades to $n_a/(n_b{+}n_a)$, and realized Λ_a rises from 0.003 to 0.085 (max 0.50) as composition ASR goes $0.093 \to 0.569$ (that α=0.1 cell is below our accuracy floor and so uninterpretable as ASR; the margin collapse plotted here is not). Values below 10^{-4} are drawn at 10^{-4}.

Our own menu witnesses $\mathcal{E} \neq \emptyset$: FoolsGold→RFA reaches 0.093 where on the pixel backdoor FoolsGold alone reaches 0.732 and RFA alone 0.845 (§E). One emergent pair among five low-ASR pairs puts the cap here at 80% against a measured 40%, so emergence explains part of the gap; had this menu's low-ASR set been emergent throughout, the same criterion would have certified nothing. That is why we report recall as an operating point rather than as a residual (supplement §S22).

Mechanism–effectiveness confounding: Proposition 7 in our instance. The boundary has an uncomfortable consequence for validation, and it is worth spelling out concretely. In *every* in-scope pair the criterion certifies, d_2 already suppresses the attack that d_1 fails. A certified composition's low ASR can therefore always be attributed to d_2's standalone strength (C1) rather than to d_1 having preserved d_2's mechanism (C2, C3): the two are confounded **by the conditions themselves**, not by an unlucky choice of defense menu. We name this **mechanism–effectiveness confounding** and offer it as a methodological observation about validating composition criteria of this shape, applicable beyond our own experiments.

In particular it rules out the experiment we previously named as decisive: a d_2 whose admission signal is invariant under positive rescaling but which is *weak in isolation* cannot satisfy C1 in scope, so no such certified pair exists to test. The achievable substitute is **per-attack complementarity**: a d_2 that suppresses the attack d_1 fails while remaining weak on a *different* committed attack, so its max-committed ASR stays high and it cannot suppress the composition single-handedly. Reputation is such a defense (0.017 on model-scaling, 0.842 on the pixel backdoor), and §S24.2 builds the test on that observation.

Argument sketch. (C1 sufficient) If at least one of d_1, d_2 individually suppresses a, composition inherits that suppression provided C2∧C3 preserve the signal chain. **That covers two routes and only one of them is this paper's subject.** If the suppressing constituent is d_2, the inheritance is exactly what C2∧C3 are for. If it is d_1, the suppression is inherited *provided d_2 is not invariant to the transformation class d_1 suppresses through*; subject to that proviso d_1 sits upstream and nothing downstream undoes it, so C2∧C3 are not required on that attack and the preservation argument contributes nothing to it. **We had stated this route without the proviso, and it is not vacuous** (below). Both routes count toward a max-committed label, which is why the label can be right on an attack the mechanism has no purchase on (§S24). C1 is sufficient, not necessary: two individually-weak defenses may suppress synergistically through complementary mechanisms (e.g., partial downweighting + geometric rejection), but predicting such synergy requires the quantitative margins below. The criterion is therefore *conservative* in the direction this sketch argues for: it may reject working compositions. **It is not a guarantee in the other direction, and nothing here establishes one.** That no clean C1∧C2∧C3 PASS is a false positive is a *measured* property of 42 pairs (supplement §S21), not a consequence of the argument

above; the criterion has no established predictive validity ((L2), §G), and against an adversary that solves both constituents' constraints at once a certified pair's ASR rises 2.4× (§E).

(C2 role) Suppose d_2 individually suppresses a via signal $\sigma_2(u_a) > \theta$. If d_1's transformation T_1 destroys this signal ($\sigma_2(T_1(u_a)) \not\approx \sigma_2(u_a)$), d_2 cannot detect the attack. Example: NormClip as d_1 maps every client norm above τ to τ. A σ_2 keyed on the update norm is then client-independent, and C2 fails outright. Reputation's consensus distance $\sigma_2(u) = \|u - \hat{\mu}\|$ does not collapse that way: at $\|u_i\| = \tau$ for every i, $\|u_i - \hat{\mu}\|^2 = \tau^2 + \|\hat{\mu}\|^2 - 2\langle u_i, \hat{\mu}\rangle$, so client dependence survives through the inner product. Clipping removes the radial part of the separation and no more. For a purely radial attack that is the whole of it, and model-scaling is one, since it multiplies the update by a constant that clipping undoes exactly; where the adversarial direction also differs, the angular part survives and C2 fails by margin compression rather than by annihilation. This is measured, not hypothetical: NormClip→reputation leaves model-scaling ASR at 0.807 ± 0.048 and the pixel backdoor at 0.842 ± 0.047 over 5 seeds, against reputation alone at 0.017 (per-seed values in the supplement §S19.1). Quantitatively: C2 holds when relative signal distortion $\Delta_{\text{signal}} := |\sigma_2(T_1(u_a)) - \sigma_2(u_a)|/\sigma_2(u_a) < \delta_2$ for a defense-specific tolerance δ_2.

(C3 role) Even when d_1 preserves the *direction* of d_2's signal (C2 holds), it may compress magnitudes below d_2's threshold. Example: FoolsGold as d_1 downweights adversarial clients by $\alpha \ll 1$. TrimmedMean's coordinate-wise rank signal still distinguishes adversarial coordinates (C2 holds), but after rescaling by α, per-coordinate values $\alpha \cdot u_{a,k}$ fall within the trimming range: the coordinate-wise extreme that would trigger trimming is no longer extreme. Quantitatively: C3 holds when the margin ratio $\Delta_{\text{margin}} := (\tau_{d_2}(T_1(u_a)) - \theta)/(\tau_{d_2}(u_a) - \theta) > 0$, i.e., adversarial scores remain above d_2's rejection boundary after d_1's transform.

(Sufficiency) If C1 holds, C2 holds ($\Delta_{\text{signal}} < \delta_2$), and C3 holds ($\Delta_{\text{margin}} > 0$), then d_2 detects and suppresses u_a even after d_1's transformation. For the weighting→aggregator class, Theorem 9 instantiates these margins explicitly: $\rho < r^\star$ for the rank aggregators, and the transformed-point separation condition $\Delta'_{\text{sep}} > 2(R'_B + \delta')$ for the geometric one, which Corollary 3 restates in a form that can be checked without running the composition. □

The proviso on C1's d_1 route, and why this menu cannot test it. Every d_1 we run acts as a positive per-client rescaling $T(u_i) = c_i u_i$ (Proposition 11), so the proviso asks only whether d_2 reads that rescaling at all. Two of our own downstream choices do not, and they fail it to different depths. *Partially:* `cos_krum` scores a rescaling-invariant statistic, so its *decision* ignores d_1's downweighting entirely, which is measured rather than argued. Under one upstream RFA stage on live update stacks, Krum's selection moves in 0.533 of rounds and `cos_krum`'s in $\mathbf{0.000}$, at a nonzero aggregate displacement of 0.133 (App. D.7). The transform acted and the decision did not move. But the emitted update is still the downweighted one, so on that d_2 the magnitude channel carries the inheritance the decision channel has stopped carrying. *Completely:* FLTrust gates on the scale-invariant $\cos(u_i, g_{\text{srv}})$ and then renormalizes every update to $\|g_{\text{srv}}\|$, so a positive per-client rescaling cancels *exactly* on both channels, which §S24.1 measures on the six pairs placing it downstream. **The proviso bounds the argument and corrects no measured row**, and Lemma 1 says why it cannot be otherwise here: the only d_2 in our menu that cancels both channels also suppresses both committed attacks alone (0.048), so in scope the d_1 route is never the *sole* route to C1 and no certified pair isolates it. The five measured reputation→ d_2 pairs inherit reputation's model-scaling suppression at 0.0146 to 0.0181 against its own 0.017 (n=3 each, seeds 42–44, in `results/all_compositions/summary.json`) precisely because none of those five downstream defenses renormalizes. Testing the proviso needs a d_2 that is rescaling-invariant on *both* channels and weak in isolation, and this paper does not contain one.

C.2 From the federated protocol to the abstract statement, term by term

Proposition 7 is stated over an abstract screen precisely so that nothing in it rests on federated learning, and Lemma 1 is its instance under our criterion. That correspondence is what carries the result into this paper and out of it again, so we display it rather than assert it. Figure 2 in the main text is the same object drawn as a graph, including the design that repairs it; Table A4 is the same object written as a substitution.

Two features of the substitution carry the whole argument, and both are legible in the table rather than argued for. **First, E and the quantity the screen predicts are the same measurement.** C1 is an ASR comparison at 0.5 and the PASS label is an ASR prediction at 0.5, so the gate is not an unlucky covariate that a larger menu would decorrelate: it is a function of the outcome. **Second, P is decided without running anything.** C2 and C3 are read off d_1's and d_2's definitions, which is what Prop. 11 makes algebraic, so no version of the design recovers identification by measuring P more carefully. The two together are what force the within-defense design of §D.2: with d_2, a and the seeds fixed, E is fixed too, and the intervention on T moves P alone.

What the mapping does not license. It shows our criterion is *an* instance, not that every composition screen is one. A screen whose gate reads something other than the outcome it predicts, a per-client norm statistic say, has a different E and falls outside Proposition 7 entirely; the proposition is a statement about outcome-gated

Table A4: **The abstract statement and our criterion are the same object under one substitution.** Left: the symbols of Proposition 7, which mentions no dataset, architecture, attack family or defense. Middle: what each becomes under Criterion 4. Nothing in the middle column is new here; the table is a reading aid for a correspondence the two statements already have.

Proposition 7	Our federated instance	Stated at
outcome metric $M(d, a)$	$\mathrm{ASR}(d, a)$, the committed attack's success rate under defense d	Criterion 4
threshold θ	0.5, the level at which the trigger fires on the majority of triggered inputs	Table A2
load-bearing condition $M(d_1, a) \geq \theta$	(C0), $\mathrm{ASR}(d_1, a) \geq 0.5$ for some committed a	Rem. 6
eligibility gate E: $\min_d M(d, a) < \theta$	(C1), $\min_{d \in \{d_1, d_2\}} \mathrm{ASR}(d, a) < 0.5$, quantified per attack	Criterion 4
mechanism condition $P(d_1, d_2)$	(C2)∧(C3): $\sigma_2(T_1(u_a)) \approx \sigma_2(u_a)$, and the margin survives T_1	Prop. 11, Thm. 9
menu $\mathcal{D}$	the pairs $(d_1 \to d_2)$ a suite draws from a defense library	§S24
certification $M(d_1 \to d_2, a) < \theta$	the criterion's PASS label on a pair	§S24.1
the empty cell of the 2×2	$\neg P$ against a d_2 that fails alone: no admitted pair has one	Lem. 1
holding d_2, a, E fixed	Mode S and Mode A: same d_2, attack and seeds, intervene on T	§D.2

screens, ours included. Nor does the mapping transfer any *empirical* claim: the middle column names measured quantities, and their scope is Table A2's, not the proposition's.

C.3 A RESTRICTED THEOREM: C2∧C3 FOR WEIGHTING→RANK AND WEIGHTING→GEOMETRIC COMPOSITIONS

What follows is a boundary theorem, and deliberately not a security theorem. It bounds what an *invariance* argument about composition can establish: preservation of the downstream defense's discriminative mechanism, levels (P1)–(P3) of Definition 2, and it says nothing whatever about suppression, which lives at (P4)–(P5). We state that in the same breath as the theorem because the whole point of the paper is the gap between the two: the theorem is the best case for statistic-level reasoning, proved with explicit margins, and §D.2 then shows that even exact invariance neither implies nor preserves suppression. Read it as a *formal* claim in the sense of Table A2: proved, on one class, bounding no ASR.

The criterion above is an operational heuristic. For the class of compositions that matter most in practice, a per-client *weighting* defense d_1 (reputation, FoolsGold) followed by a *rank-based* or *geometric* aggregator d_2 (TrimmedMean, CoordMedian, RFA), which contains the flagship FoolsGold→RFA, C2 and C3 can be *proven* rather than asserted, with an explicit tolerance on how aggressively d_1 may reweight. The key structural fact is that d_2's admission score is scale-equivariant, so it commutes with d_1's reweighting up to a bounded ratio; the composition fails only when d_1 compresses that ratio past a margin set by d_2's separation.

Theorem 9 (Bounded reweighting preserves the downstream discriminative mechanism)**.** *Consider a round with n_b benign and n_a adversarial updates under honest majority $n_b > n_a$. Let d_1 be a per-client reweighting $T_1(u_i) = w_i u_i$ with $w_i \geq 0$, and define the ratio over* strictly positive *weights, $\rho := \max_i w_i / \min_{i:\, w_i > 0} w_i$. Suppose d_2 suppresses attack a through a client admission score $\sigma_2(\cdot\,; U)$ that is scale-equivariant, $\sigma_2(c\,u_i; c\,U) = c^k \sigma_2(u_i; U)$ for all $c > 0$ and some degree $k \geq 0$. Assume $w_a > 0$ for adversarial clients; the $w_a = 0$ case is Lemma 2. The theorem establishes preservation of d_2's suppression* mechanism*; it does not bound ASR. Then:*

1. ***(Rank aggregators; C2∧C3 hold up to a margin.)*** *If $d_2 \in \{$TrimmedMean, CoordMedian$\}$, so σ_2 is a coordinate-wise rank, then d_2's per-coordinate suppression of a is preserved when two conditions hold.* ***(i) Count majority:*** *$n_a < K/2$ for CoordMedian, and $n_a \leq \lfloor \beta K \rfloor$ for TrimmedMean with β-trimming.* ***(ii) Signed ordering:*** *on each suppression-relevant coordinate k, the adversarial transformed value lies strictly beyond every benign one, $w_a u_{a,k} > \max_b w_b u_{b,k}$ (upper tail) or $w_a u_{a,k} < \min_b w_b u_{b,k}$ (lower tail).* Sufficient magnitude condition for *(ii): if $\mathrm{sign}(u_{a,k})$ is common to all adversaries and $|u_{a,k}| > \rho \cdot \max_b |u_{b,k}|$, then (ii) holds; writing $r^\star := \min_k \min_a \min_b |u_{a,k}|/|u_{b,k}|$ for the coordinate-wise separation margin, which minimizes over adversaries as well so that the bound binds on the least extreme one, this is $\rho < r^\star$. Magnitude alone is* not *sufficient without the sign assumption, and the reason is not that a large adversarial coordinate could sit between the benign values: since $\rho \geq 1$, the magnitude condition gives $|u_{a,k}| > \max_b |u_{b,k}| \geq \max(\max_b u_{b,k}, -\min_b u_{b,k})$, which forces $u_{a,k} > \max_b u_{b,k}$ or*

$u_{a,k} < \min_b u_{b,k}$. *What magnitude leaves open is* which *of the two. Mixed adversarial signs split the adversaries across both tails, while (ii) is a one-tail condition and (i)'s count budget is written against a single tail; a common sign is what puts them all in one. With* $n_a{=}1$ *the sign assumption is therefore vacuous. At* $\rho = 1$ *ranks are preserved exactly, recovering Proposition 10.*

2. ***(Geometric median / RFA; C2∧C3 hold up to a margin.)*** *The composition hands RFA the* transformed points $u'_i := w_i u_i$ *and RFA runs its standard* unweighted *geometric median on them; it is not a weighted geometric median, and* w_i *enters only by moving the points. All quantities must therefore be defined on the transformed set. Let* μ'_B *be the geometric median of the benign* u'_b *alone,* μ' *that of all of* U'*,* $R'_B := \max_b \|u'_b - \mu'_B\|$ *the transformed benign radius, and* $\delta' := \|\mu' - \mu'_B\|$ *the displacement caused by including the adversarial points. Suppression is preserved when*
$$\Delta'_{\text{sep}} := \min_{a,b} \|u'_a - u'_b\| \;>\; 2\,(R'_B + \delta'),$$
under which every adversarial residual $\|u'_a - \mu'\|$ *strictly exceeds every benign residual* $\|u'_b - \mu'\|$*, so RFA's inverse-residual weight on* a *stays below the benign floor.*

3. ***(Norm-equalizing*** d_1***; C2 fails or is compressed: the degenerate case.)*** *If instead* d_1 *maps every client norm to a common value (NormClip with* τ *below all client norms), the consequence depends on what* σ_2 *reads. A* σ_2 *keyed on the update norm becomes constant across clients, so it stops separating* a *and C2 fails outright. A* σ_2 *keyed on consensus distance does* not *become constant: at* $\|u_i\|{=}\tau$ *for all* i*,* $\|u_i - \hat{\mu}\|^2 = \tau^2 + \|\hat{\mu}\|^2 - 2\langle u_i, \hat{\mu}\rangle$*, so client dependence survives through the inner product. Equalization removes the* radial *part of the separation and compresses the rest to its angular part, and C2 then fails once that residual drops below* d_2*'s threshold, which is a margin condition of the same form as parts (1) and (2) rather than an algebraic collapse. Nor is the weight ratio infinite:* $w_i = \tau/\|u_i\|$ *is finite for every non-zero update, so* ρ *is a finite scalar and a* ρ*-margin formally applies. What fails is that* ρ *is set by the realized norms, so no margin fixed* before *the round bounds it; weight exactly zero is the genuinely unbounded case, and that is Lemma 2.*

Parts (1) and (2) are sufficient conditions for d_2 *to* preserve the downstream suppression ordering*, not to exclude adversarial contributions outright. Neither is a sufficient condition for low ASR, and part (2) orders residuals per client only; it constrains the* aggregate *adversarial coefficient solely through Corollary 4.*

Proof. (1) TrimmedMean/CoordMedian act on coordinate k through the rank of $w_i u_{i,k}$ among $\{w_j u_{j,k}\}_j$; a is suppressed on k because its transformed coordinate is extreme (trimmed, or not selected by the median). Given (ii), at most n_a of the extreme order statistics are adversarial, so with (i) the selected value is benign. For the magnitude condition, $w_a > 0$ gives $\rho \geq \max_b w_b/w_a$, so $|u_{a,k}| > \rho \max_b |u_{b,k}|$ implies $|w_a u_{a,k}| > w_a \rho \max_b |u_{b,k}| \geq (\max_b w_b) \max_b |u_{b,k}| \geq \max_b |w_b u_{b,k}|$; with a sign common to all adversaries this places $w_a u_{a,k}$ strictly beyond every benign value on that side. Scale-equivariance ($k{=}0$ for pure rank) makes the uniform case $\rho{=}1$ exact. (2) All quantities are on the transformed points $u'_i = w_i u_i$, which is what RFA receives. Every benign residual obeys $\|u'_b - \mu'\| \leq R'_B + \delta'$ by the triangle inequality. For adversarial u'_a, $\|u'_a - \mu'_B\| \geq \Delta'_{\text{sep}} - R'_B$, hence $\|u'_a - \mu'\| \geq \Delta'_{\text{sep}} - R'_B - \delta'$. Demanding that this exceed the benign bound gives $\Delta'_{\text{sep}} - R'_B - \delta' > R'_B + \delta'$, i.e. the stated condition. No weighted-majority assumption is needed, and none would be appropriate: RFA weights every input point equally and the w_i act only by relocating the points. Our implementation runs Weiszfeld for 50 iterations at tolerance 10^{-6}, so the argument transfers to it. (3) Clipping to a common τ sends $\|u_i\| \mapsto \tau$ for all i. For $\sigma_2(u_i) = \|u_i\|$ (NormClip-as-d_2) the score is then a client-independent constant, separation vanishes, and that is C2-failure outright. For the consensus distance, expand $\|u_i - \hat{\mu}\|^2 = \tau^2 + \|\hat{\mu}\|^2 - 2\langle u_i, \hat{\mu}\rangle$: the first two terms are common to all clients, so every trace of magnitude separation is gone, but the inner product is not, and it carries the angular separation between u_i and the consensus direction. Equalization therefore compresses the separation onto its angular component instead of annihilating it, and C2 fails once that residual falls below d_2's threshold θ, a margin condition of the same shape as parts (1) and (2). The norm-keyed score is the special case in which the residual is identically zero. □

The condition as stated cannot be checked before running the composition, and that is fixable. Part (2) is stated on the transformed points, and $\delta' = \|\mu' - \mu'_B\|$ is by definition the displacement that including the adversarial points causes: evaluating it requires executing the very composition the criterion is meant to screen. In a screening criterion that is a real defect. It is also removable, because the displacement is controlled by the benign spread and the client counts alone.

Corollary 3 (Checkable separation margin)**.** *In the setting of Theorem 9(2), write* $S'_B := \sum_b \|u'_b - \mu'_B\|$ *and let* $\bar{R}'_B := S'_B/n_b$ *be the* mean *transformed benign residual. Under honest majority* $n_b > n_a$*,*
$$\delta' \;\leq\; \frac{2S'_B}{n_b - n_a} \;=\; C_A\,\bar{R}'_B, \qquad C_A := \frac{2n_b}{n_b - n_a},$$
and therefore the separation condition of Theorem 9(2) holds whenever
$$\Delta'_{\text{sep}} \;>\; 2R'_B + 2\,C_A\,\bar{R}'_B. \tag{1}$$
No quantity in (1) *is computed from the composed aggregate.*

Proof. μ' minimizes $F(x) := \sum_i \|x - u'_i\|$, so $F(\mu') \le F(\mu'_B)$. Write $t := \delta'$. By the triangle inequality each benign term obeys $\|\mu' - u'_b\| \ge t - \|u'_b - \mu'_B\|$ and each adversarial term obeys $\|\mu' - u'_a\| \ge \|\mu'_B - u'_a\| - t$, so $F(\mu') \ge n_b t - S'_B + \sum_a \|\mu'_B - u'_a\| - n_a t$. Since $F(\mu'_B) = S'_B + \sum_a \|\mu'_B - u'_a\|$, the inequality $F(\mu') \le F(\mu'_B)$ gives $(n_b - n_a)\, t \le 2S'_B$, which is the displacement bound. Substituting it into $\Delta'_{\text{sep}} > 2(R'_B + \delta')$ yields (1). □

Honest majority is exactly what the geometric case needs, and it enters on the displacement, not on the weight ratio. C_A diverges as $n_a \to n_b$, so Corollary 3 is the precise point at which $n_b > n_a$ becomes indispensable, while the theorem itself needs no majority assumption on the *weights*. An earlier version of this paper summarized part (2) as a weight-ratio bound $\rho < n_b/n_a$; that summary was wrong, and the correction is not cosmetic, because the majority condition belongs on δ' instead. **Measured at Dirichlet** α=0.5: the displacement bound holds in all 17 rounds with $w_a > 0$, and is conservative there by a factor of at least 10.5 (mean 215), because the realized displacement is set by a force balance among the residuals rather than by the worst case a bound must cover. Condition (1) is satisfied in 16 of those 17 rounds: 15/15 at n_a=1, where $C_A = 2.67$ and the requirement at the measured mean $\bar{R}'_B/R'_B = 0.429$ is $\Delta'_{\text{sep}} > 4.29\, R'_B$ against a measured $\Delta'_{\text{sep}}/R'_B$ never below 5.88; and 1/2 at n_a=2, where $C_A = 6.00$ tightens the requirement past what one of the two rounds supplies. **The heterogeneity limit recorded for Corollary 4 applies here too**: at α=0.1 condition (1) holds in only 5 of 12 positive-weight rounds. The checkable form certifies less exactly where the condition it replaces also stops binding, and it does not rescue the guarantee there.

A further reduction to ρ alone is vacuous on our runs, and we report it as vacuous. The natural next ask is the analogue of part (1)'s $\rho < r^\star$: a threshold on ρ computable from the *pre-composition* round geometry, so that no pilot is needed at all. Such a threshold exists formally. Bounding $\Delta'_{\text{sep}} \ge w_{\min}[\Delta_{\text{sep}} - (\rho - 1)N_A]$ and $R'_B \le 2w_{\min}[\rho R_B + (\rho - 1)\|\mu_B\|]$, where Δ_{sep}, R_B and μ_B are the untransformed analogues and $N_A := \max_a \|u_a\|$, makes $w_{\min}$ cancel and turns (1) into a linear inequality in ρ with a critical value $\rho^\star$ depending only on $(\Delta_{\text{sep}}, R_B, \|\mu_B\|, N_A, n_a, n_b)$. **It is empty on our data.** It requires $\Delta_{\text{sep}} > 4(1{+}C_A)R_B$, and the measured $\Delta_{\text{sep}}/[4(1{+}C_A)R_B]$ clears 1 in exactly 1 of the 17 positive-weight rounds. Its maximum is 1.015 at α=0.5, attained in that n_a=1 round where $C_A = 2.67$ makes the requirement loosest, with mean 0.754 over those 15 rounds and a maximum of only 0.389 over the two n_a=2 rounds; at α=0.1 the maximum is 0.853. Consequently $\rho^\star \le 1$ in 16 of 17 rounds at α=0.5 and in all 12 at α=0.1, against a measured ρ of at least 1.034 and 1.015 in the two settings respectively; in the one round where $\rho^\star$ does exceed 1 it reaches only 1.006, against a measured ρ of 1.196 in that round. The reduced condition is therefore met in *no* round, which is a stronger negative than a merely conservative threshold. The looseness is in the composition of the two bounds rather than in either one separately, and closing it would need a curvature assumption on the benign configuration that we are not prepared to assume. **We therefore claim no pilot-free screen for the geometric class**, and the calibration protocol below prices it accordingly.

What part (2) does not assume. Recorded explicitly, because the step invites readings it does not support: there is no honest-majority assumption on the *weights*, no requirement that adversarial updates lie outside the benign convex hull, and the count majority $n_b > n_a$ is used in part (2) only through Corollary 3, nowhere in the theorem's own proof.

Theorem 9(2) orders residuals *per client*: each adversarial point receives less inverse-residual weight than each benign point. That alone does not bound the adversary's *total* share of the aggregate: several adversaries could each be down-weighted yet collectively hold substantial mass. The separation margin controls this, but only quantitatively:

Corollary 4 (Bounded aggregate adversarial influence)**.** *Let $m := \Delta'_{\text{sep}}/[2(R'_B + \delta')]$ be the realized separation margin, so Theorem 9(2) reads $m > 1$. Every benign residual is at most $D_b := R'_B + \delta'$ and every adversarial residual at least $D_a := \Delta'_{\text{sep}} - R'_B - \delta' = (2m-1)D_b$. Bounding $\sum_j 1/d_j$ below by the benign terms alone and each adversarial term above by $1/D_a$,*

$$\Lambda_a \;=\; \sum_a \lambda_a \;\le\; \frac{n_a/D_a}{n_b/D_b + n_a/D_a} \;=\; \frac{n_a}{n_b(2m-1) + n_a},$$

so μ' is a convex combination whose adversarial share is at most Λ_a: writing $\mu' = (1-\Lambda_a)\bar{u}_B + \Lambda_a \bar{u}_A$ for its benign and adversarial parts, the displacement from the benign-only combination $\bar{u}_B$ is at most $\Lambda_a \|\bar{u}_A - \bar{u}_B\|$.

The bound is only as good as the margin, and **the separation condition alone is not enough**: as $m \to 1$ it degrades to $n_a/(n_b + n_a)$, permitting the adversary 40% of the mass at n_a=2, n_b=3. A margin comfortably above 1 does the work. We measure $m \ge 2.82$ in every round with $w_a > 0$ (mean 4.71 over those rounds), capping the adversarial share at 5.1% for n_a=1 and 12.6% for n_a=2. The realized figure is far tighter than the bound: measuring Λ_a directly, it never exceeds 1.8% in any round, and the bound is satisfied in all 17 measured rounds with $w_a > 0$. With Lemma 2 this completes the picture, and the mechanism survives for *opposite* reasons: at w_a=0 the adversaries hold most of the Weiszfeld mass ($\Lambda_a = 0.71$ measured) but carry no payload; at $w_a > 0$ they carry payload but hold at most 12.6%.

This is the one place in the paper where the theory predicts a number and the measurement lands inside it, so we state the pair plainly: **bound** 12.6%**, realized** $\leq 1.8\%$, in every measured round with $w_a > 0$. **And it is also the one place where heterogeneity makes the guarantee measurably empty.** The bound is a function of the realized margin m alone, and at Dirichlet α=0.1 RFA's **all-round mean** margin **collapses from** m=3.24 **to** m=1.01, exactly the limit at which Corollary 4 degrades to $n_a/(n_b{+}n_a)$ and certifies nothing, with two of three certified compositions crossing the ASR threshold there on the raw reading, one of three once our own accuracy floor is applied (§S24.4(E)). So the theorem's explanatory value is genuine and measured in *both* directions: it predicts a number that holds where heterogeneity is moderate, and its own quantity records where it stops holding. **That is not a robustness check**: the theorem's own quantity is what records where it stops holding, which is why it is carried as a finding rather than as a caveat (§D.2, §H).

Λ_a is level (P4) in measurable form, but (P4) is a *level*, not a single metric. Before the instantiation, one clarification the rest of the paper depends on: **(P4) is a conceptual level of the hierarchy and is instantiated defense-specifically**, because different aggregator families have different notions of what it means for adversarial input to enter the aggregate at all. A *selector* admits or does not, so its (P4) is an indicator in $\{0, 1\}$; a *weighted averager* has an adversarial coefficient share in $[0, 1]$; a *coordinate-wise order statistic* has a fraction of selected coordinates that are adversarial. There is no single scalar all three compute, and this is why the influence column of Table A8 is **necessarily aggregator-specific** rather than a common measurement across rows: a fact about the level, not a shortcoming of the table. With that said: Definition 2's levels are not equally meaningful. (P1)–(P3) are properties of a statistic and of a decision rule reading it; **(P4) is the first level whose content is the adversary's influence on the aggregate**, and for a geometric aggregator Λ_a is exactly that level made measurable: Corollary 4 bounds it, Lemma 2 drives it to a benign-only convex combination, and §D.2's instruments intervene on it directly. So the negative results do not leave the theory empty-handed: **the level the evidence points to is one the theory can bound, just not from a** (d_1, d_2) **definition pair alone**, which is what a screen would need. That gap is the open problem we state in §H. **Three quantities are distinguished throughout and never used interchangeably**: *admitted adversarial mass*, whether an adversarial update enters d_2's aggregate at all, binary for a selector and a coordinate fraction for an order statistic; the *adversarial coefficient share*, the fraction of the upstream coefficient vector's total mass held by adversaries, which is what the instruments of §D.2 pin; and *aggregate adversarial influence* Λ_a, the weight the adversary's direction actually carries in the aggregate. Unqualified, "admission" means the first. (Symbols are collected in Appendix J.)

Lemma 2 (Zero-weight annihilation)**.** *Zero-weight clients are* not *removed from the composition: d_2 receives them as zero vectors and still sees all K points. Our implementation runs the ε-smoothed Weiszfeld iteration ($\varepsilon = 10^{-8}$), whose fixed point is $\mu' = \sum_i \lambda_i u'_i$ with $\lambda_i \propto 1/\max(\|u'_i - \mu'\|, \varepsilon)$: a convex combination for* every *configuration, including the singular case $\mu' \in U'$ that a zero-weight client at the origin can create, so no non-degeneracy assumption is required. If $w_a = 0$ for every adversarial client, then $u'_a = 0$ and, writing $\Lambda_a := \sum_a \lambda_a$, the adversarial terms vanish identically: $\mu' = \sum_{b\ benign} \lambda_b u'_b = (1 - \Lambda_a)\,\bar{u}$ with $\bar{u}$ a convex combination of benign transformed updates. The aggregate is a* contraction by $1 - \Lambda_a \in (0, 1)$ of a benign-only weighted average*: the attack direction never enters μ'. The origin points' only effect is that contraction: a reduction in effective step size, i.e. a utility cost, not a security failure. The same holds for CoordMedian, where a zeroed coordinate carries no payload but shrinks the aggregate when selected.*

The iteration is truncated, and the lemma's conclusion does not depend on that. The shipped code stops at `max_iter` $= 50$ or when successive iterates differ by less than `tol` $= 10^{-6}$, in `fl_core/federated.py` and, with the same two constants hardcoded, in the upstream-stage branch of `apply_d1_transform`. It returns the iterate *preceding* the one that passes that test, which on a first-iteration exit is the plain mean. So $\lambda_i \propto 1/\max(\|u'_i - \mu'\|, \varepsilon)$ holds at the fixed point the lemma reasons about and not at the vector actually returned, whose weights are reciprocal distances to the *previous* iterate, or uniform. **What the lemma needs survives truncation exactly**: every iterate has the form $\sum_i \lambda_i u'_i$ with λ convex, and a zero-weight client contributes $\lambda_a \cdot 0 = 0$ whatever λ_a is, so the adversarial terms vanish identically at every iterate and the contraction reading is unchanged. We record this because the lemma is stated about a fixed point and the implementation does not reach one.

Why the positive-weight restriction matters. Defining ρ over all weights would make it *infinite in every round*: FoolsGold's normalization subtracts the minimum weight, so 2 of 5 clients receive exactly zero (measured, N=10, K=5). A bound of the form $\rho < 4$ would then never be satisfiable, and Theorem 9(2) would be vacuous on the very composition it is stated over. Restricting ρ to strictly positive weights and handling w_a=0 separately via Lemma 2 is what makes the analysis apply to the algorithm we actually run. **Both channels are drawn in Figure A2**, which is the composed pipeline as a DAG rather than the evaluation design of Figure 2: the statistic path C2 is stated over, and the zero-weight arc that reaches the aggregate without passing through σ_2 at all.

Operational calibration protocol. To evaluate the criterion for $(d_1 \to d_2)$: **C1** is available from standard single-defense baselines (zero cost). **C2/C3 for the theorem-backed class** (FG or reputation as d_1; RFA, CM, TM as d_2): for the *rank* aggregators, Theorem 9(1)'s bound $\rho < r^\star$ needs only ρ and the coordinate-wise separation,

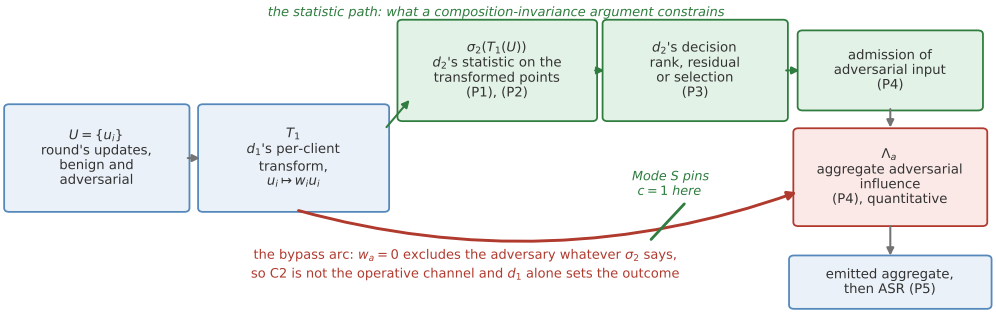

Figure A2: **The composed pipeline as a DAG, which is a different graph from Fig. 2**: that one is the *evaluation design*, this one is what a client update passes through. **The statistic path** (green) is what a composition-invariance argument constrains: d_1's per-client transform T_1 moves the points, d_2 computes σ_2 on the transformed points, σ_2 drives d_2's decision, and the decision determines admission, i.e. levels (P1)–(P4) of Def. 2 in order. **The bypass arc** (red) is the one the paper's argument turns on: d_1's weights reach Λ_a (Cor. 4) without passing through σ_2 at all, so when d_1 zeroes an adversary's weight the adversary is excluded whatever σ_2 says and d_1 alone sets the outcome (Lem. 2). Mode S cuts that arc by pinning every adversary's coefficient at c=1, which is why it identifies the intervention's effect at a fixed adversarial coefficient share, **not the statistic path in isolation**, since T_1 also moves the magnitude of what d_2 emits (Table 2's Δ agg. column), and also why it needs adversary identity and is an instrument rather than a defense. **A statistic-level argument constrains the top path and leaves the bottom one free**, which is the whole of why (P1)–(P3) do not deliver (P5).

both read off the C1 baseline runs, so this case really is free. **For the *geometric* case an earlier version of this paper claimed the same, and that was false**: Corollary 3 removes the one quantity that would have required running the composition, but the benign spread that replaces it is measured on the transformed points, so RFA carries the same short-pilot cost as the other families and we no longer exempt it. **C2/C3 for other families**: one short pilot (5–10 rounds, $\sim$10% of full cost) measuring $\sigma_2(T_1(u_a))$ vs. $\sigma_2(u_a)$ suffices. The full 50-round experiment runs only for criterion-PASS candidates. $\delta_2 = 0.5$ separates all PASS from C2-FAIL cases on the development set; insensitive to choice: any $\delta_2 \in [0.3, 0.7]$ yields identical classifications on all 42 pairs.

Proposition 10 (Rank-invariance of coordinate-wise aggregators)**.** *Let d_2 be a coordinate-wise rank aggregator (TrimmedMean or CoordMedian) with admission statistic $\sigma_k(u) = \mathrm{rank}_k(u)$ among the received updates $\{u_1, \ldots, u_m\}$. Let $T_1(u_i) = \alpha \cdot u_i$ be a* uniform *positive rescaling ($\alpha > 0$, same for all clients). Then:*

1. *$\sigma_k(T_1(u_i)) = \sigma_k(u_i)$ for all i, k, i.e., C2 and C3 hold trivially.*
2. *Under* heterogeneous *rescaling $T_1(u_i) = \alpha_i \cdot u_i$ with $\alpha_i \neq \alpha_j$, rank preservation fails when rescaling reorders coordinates across clients: $\exists\, k$ s.t. $\mathrm{rank}_k(\alpha_i u_{i,k}) \neq \mathrm{rank}_k(u_{i,k})$.*

Proof. (1) Rank is invariant to positive scalar multiplication: $u_{i,k} > u_{j,k} \iff \alpha u_{i,k} > \alpha u_{j,k}$ for $\alpha > 0$. (2) Counterexample: let $u_{1,k} = 3$, $u_{2,k} = 2$, with $\alpha_1 = 0.1$, $\alpha_2 = 1$. Then $\alpha_1 u_{1,k} = 0.3 < \alpha_2 u_{2,k} = 2$, reversing the rank. □

Corollary 5. *CoordMedian (pure rank selection) is robust to C3 boundary compression under any uniform rescaling but vulnerable under heterogeneous client-specific rescaling (e.g., FoolsGold or reputation weights). This predicts that compositions X→CoordMedian will show* intermediate *ASR when d_1 applies heterogeneous weights: consistent with observed NC→cm (0.433) and rfa→cm (0.485), both near the 0.5 threshold.*

Proposition 10 concerns rank under *uniform* rescaling. The sharper question is which admission signals survive *heterogeneous* per-client rescaling, since that is what every per-client defense in our menu applies: NormClip ($c_i = \min(1, \tau/\|u_i\|)$), reputation ($c_i \propto e^{-d_i/s}$) and RFA-reweighting ($c_i \propto 1/\max(\|u_i - \mu'\|, \varepsilon)$) are all of the form $T(u_i) = c_i u_i$ with $c_i > 0$ strictly, each positive by construction. FoolsGold is of the same form with $c_i \geq 0$ but attains $c_i = 0$ in *every* round (its shift-and-clamp gives the least-trusted client weight exactly zero, 2 of 5 clients here), so it falls outside the strict hypothesis below and is governed by Lemma 2 instead: the same distinction we draw for FLTrust's ReLU zeros in §S24.1. The strictly-positive case has a sharp answer, and it partitions the menu:

Proposition 11 (Invariance classes under positive per-client rescaling)**.** *Let $T(u_i) = c_i u_i$ with $c_i > 0$,* ***strictly****: the form taken by NormClip ($c_i = \min(1, \tau/\|u_i\|)$), Reputation ($c_i \propto e^{-d_i/s}$) and RFA-reweighting ($c_i \propto 1/\max(\|u_i - \mu'\|, \varepsilon)$), each positive by construction.* ***FoolsGold is deliberately not in that list****: its shift-and-clamp gives the least-trusted client weight exactly* 0 *in* every *round (2 of 5 clients here), so a zero coefficient is*

not a positive *rescaling and FoolsGold-as-d_1 falls outside the hypothesis, governed by Lemma 2 rather than by (a)–(c). Then:*

- *(a)* ***Cosine similarity is exactly invariant:*** *$\cos(c_i u_i, c_j u_j) = \cos(u_i, u_j)$. FoolsGold's weight vector is a function of the pairwise-cosine matrix alone, so $w(T(U)) = w(U)$ identically, and* ***C2 holds for every $\cdot\to$FoolsGold composition*** *(the $\cdot$ standing for any upstream d_1 throughout).*
- *(b)* ***Coordinate ordering is not invariant*** *in general; it survives under the bounded-ratio condition of Theorem 9(1).*
- *(c)* ***Consensus distance and residual magnitude are not invariant.*** *Heterogeneous c_i can invert their ordering, so $\cdot\to$reputation compositions fail C2, and $\cdot\to$RFA needs the separation condition of Theorem 9(2).*

Proof. (a) $c_i u_i / \|c_i u_i\| = u_i / \|u_i\|$ for $c_i > 0$, so every pairwise cosine, each client's normalized maximum similarity, and hence every FoolsGold weight is unchanged. (c) Counterexample for consensus distance: with $u_1{=}(1,0)$, $u_2{=}(0,1)$, $u_3{=}(3,3)$ the coordinate-wise median is $(1,1)$ and u_3 is farthest; under $c = (1,1,0.1)$ the median becomes $(0.3, 0.3)$ and u_3 becomes the *closest* point, reversing the ordering reputation depends on. □

This corrects a natural but wrong intuition. Norm clipping does *not* blind FoolsGold: it cannot, by (a), though it does blind reputation, whose signal is consensus distance rather than similarity. We verified (a) in the shipped implementation: FoolsGold's weight vector changes by at most 3×10^{-8} (float32 rounding) under NormClip, across 1000 synthetic trials and live FL rounds in which 1–2 of 5 clients were genuinely clipped (raw norms up to 24.5 against $\tau{=}5$). Note this also means Theorem 9(3)'s norm-equalizing degenerate case requires τ below *all* client norms; at our $\tau{=}5$ only 1–2 of 5 clients clip per round, so norms are not equalized and that case does not describe our runs.

C.4 The framework in general form

Positive rescaling is the instantiation that matters for current defenses, but the structure is more general, and stating it separates what is essential to the framework from what is an artifact of today's defense menu. Write d_2's admission mechanism as a statistic $S(U)$ of the round's updates together with a decision rule that consumes only the *ordering* S induces over clients, which are extreme, which are far from consensus, which are down-weighted. Let T be d_1's per-client transform. Then $(d_1 \to d_2)$ is **mechanism-preserving** exactly when

$$S(T(U)) \;=\; \varphi\big(S(U)\big) \quad \text{for some } \varphi \text{ preserving the ordering } d_2 \text{ acts on,} \tag{2}$$

which is C2; C3 additionally asks that φ not compress the quantitative gap below d_2's operating threshold. Proposition 11 is the case $T(u_i) = c_i u_i$: for S = cosine similarity, $\varphi = \mathrm{id}$, so (2) holds unconditionally; for S = coordinate value, φ is a per-client positive scaling, order-preserving only under the bounded-ratio condition $\rho < r^\star$ of Theorem 9(1); for S = consensus distance or residual magnitude, no φ exists, because the reference point (median, geometric median) moves with T, so $S(T(U))$ is not a function of $S(U)$ at all. Three consequences worth stating. First, this is why the conditions are checkable without running an experiment: one reads S off d_2's definition and T off d_1's, and the question becomes algebraic. Second, it identifies what a genuinely new instantiation would require: an S and a non-rescaling T, e.g. a d_1 that projects updates onto a subspace or applies a coordinate-wise mask, for which (2) would need to be re-derived rather than inherited. Third, it marks the framework's real boundary: a d_1 whose transform is not per-client (one that mixes updates across clients, or carries cross-round state) admits no T of this form, and nothing here applies to it.

D The intervention evidence: the confounded ladder, and the instrument it motivated

The controlled upstream transform. We introduce a synthetic d_1 whose only parameter is the *dispersion* of a positive per-client rescaling. For the K participants of a round, at position j in a $(\text{seed}, \text{round})$-keyed pseudorandom permutation,

$$c_j \;\propto\; \exp\Big(\kappa \cdot \Big(\tfrac{2j}{K-1} - 1\Big)\Big), \qquad \text{normalized so } \tfrac{1}{K}\textstyle\sum_j c_j = 1. \tag{3}$$

The abscissa is measured, not nominal. What matters to C2 is not ρ but what the rescaling does to the statistic d_2 actually reads, so before freezing any prediction we pushed live FL updates through the shipped transform and recorded each arm's decision at every κ on the *same* raw updates, computing no ASR. That establishes three things before any outcome exists (Table A5): *there is a dose*: disturbance is exactly 0 at the identity rung and rises monotonically with κ in every non-invariant arm; *the class-(a) invariance claim holds in the*

Table A5: The four arms of the dose–response ladder, and the frozen predictions. The statistic each d_2 reads is, in order, summed pairwise Euclidean distance, distance to the client median, pairwise cosine, and coordinate ordering. *Standalone* is d_2's own ASR (clean accuracy) on that attack, the C1 input, held fixed within every arm. *Measured disturbance* is the fraction of live rounds on which the rung changes that arm's decision (for `coord_median`, the mean fraction of *coordinates* whose median-attaining client changes) measured before the freeze on the same raw updates, with no ASR computed (`results/dose_disturbance.json`). Shapes and decision rules are frozen at commit `e711a95`.

	arm (d_2 / attack)	**class** (Prop. 11)	**standalone** ASR (acc)	**measured disturbance** κ=0	0.5	1.0	2.0	**predicted shape in** κ
1	`krum` / scaling	(c) not inv.	0.061 (0.577)	0.00	0.60	0.73	**0.87**	monotone rise
2	`reputation` / scaling	(c) not inv.	0.017 (0.777)	0.00	0.87	0.87	**0.93**	monotone rise
3	`cos_krum` / pixel	**(a) exact**	0.300 (0.598)	**0.00**	**0.00**	**0.00**	**0.00**	**FLAT** (control)
4	`coord_median` / pixel	(b) cond. inv.	0.443 (0.767)	0.00	0.25	0.38	**0.49**	rise, shallower

shipped code: `cos_krum`'s selection changes in **0** of 120 cells at ρ up to 54.6; and *class (b) is graded rather than binary*, at roughly half the class-(c) disturbance (0.487 vs. 0.867/0.933 at κ=2.0), which is what arm 4's prediction turns on.

Frozen predictions. The *primary* test is pooled: across all 16 (arm, κ) cells, the rise in ASR over that arm's own κ=0 rung is predicted to be positively rank-correlated with the measured disturbance rate (Spearman, one-sided, $\alpha = 0.05$). Its predictor is a *measured property of the statistic*, not the identity of the defense and not its standalone effectiveness, both held fixed within every arm. The *secondary* predictions are the per-arm shapes of Table A5, tested by a one-sided Jonckheere–Terpstra trend test across the four ordered rungs at n=5, with arm 3's "flat" scored as an *equivalence* claim at a 0.15 margin fixed and justified in advance (under a quarter of the $+0.629$ within-defense gap for Krum (the n=3 value on record at the freeze; the later top-up puts it at $+0.716$, so the margin is if anything conservative) and $0.300 + 0.15 = 0.450 < 0.5$, so "within the margin" and "still suppressing" coincide for this arm) and arm 4's class-(b) signature being the *shallower slope*.

One prediction we deliberately do not make. Theorem 9(1)'s ordering-survival condition $\rho < r^\star$ would give arm 4 a crossing rung fixed by a measured $r^\star$, but **no frozen prediction here uses** $r^\star$: we measured it and it is **not a stable per-attack constant**, ranging 0.00–12.59 on model-scaling (median 0.57) and 0.61–1.63 on the pixel backdoor (median 0.92) across the 12 rounds that sampled an adversary, so an illustrative value such as $r^\star$=10 is a stipulation rather than a measurement; §C.3 states the condition symbolically and stipulates no value for it. This is a limitation of our own theorem that the dose-response measurement surfaced, reported as one (L5) rather than quietly left unchecked.

D.1 Outcome of the confounded design: one arm confirmed, two refuted, one indeterminate, and the pooled test fails

What did the outcome-gated ladder actually return, arm by arm? All 80 runs completed (16.5 h wall clock). Table A6 is the whole ladder; Figure A3 plots it on the nominal dose axis against the frozen shapes and Figure A4 on the independently measured disturbance axis with the real-transform anchors, scored by `analyze_dose_response.py` against the rules frozen at `e711a95`, every number recomputed per seed from `results/dose_response/`. The **primary** pooled test does not confirm: across the 16 cells, the rise in ASR over each arm's own κ=0 rung correlates with the measured disturbance at Spearman $\rho_s = +0.351$, one-sided $p = 0.091$, so the frozen criterion ($\rho_s > 0$ at $p < 0.05$) is not met and **we report the primary prediction as refuted**. Of the four frozen shapes one is confirmed, two are refuted, and one is indeterminate; **in three of the four arms ASR *falls* as the dose rises**. This is the ladder's headline result and it is a negative one. It is also the outcome the pre-registration anticipated and named in advance: "pooled correlation absent while individual arms move: the mechanism is defense-specific and measured disturbance is not a transferable predictor."

Arm 2 confirms: the ladder yields one prospective confirmation, while the pooled prediction is rejected. Reputation under model-scaling rises monotonically from 0.017 at the identity rung to 0.487 at $\rho = 7.39$ and 0.779 at $\rho = 54.6$ (Jonckheere–Terpstra $z = +4.64$, one-sided $p = 1.7 \times 10^{-6}$; permutation $p < 5 \times 10^{-5}$), while clean accuracy moves only $0.777 \to 0.701$, so what is lost is suppression, not the model. Same d_2, same attack, same standalone baseline at every rung: nothing about effectiveness varies, and 55× dispersion in a *positive per-client rescaling* (a transform the criterion treats as benign for invariant statistics) carries a defense from 0.017 across the 0.5 threshold. Two qualifications keep this in proportion. It validates the *destruction* half of C2's mechanism (a non-invariant statistic, actually disturbed, loses suppression); the *preservation* half (invariance ⇒ suppression survives) is arm 3, which came out indeterminate. And even within this arm the measured disturbance is not what orders the curve: it saturates at 0.867 by κ=0.5, where ASR is still 0.022,

Table A6: The dose–response ladder as measured, n=5 per cell, per-attack mean ASR with mean clean accuracy in parentheses. *Predicted* is the shape frozen at commit `e711a95` before any run; $p_{\uparrow}$ is the one-sided Jonckheere–Terpstra probability for an *increasing* trend across the four ordered rungs, which is the only direction any frozen rule tests. Verdicts apply the frozen criteria unchanged (Table A5 and the supplement §S6); an arm satisfying neither its confirmation nor its refutation criterion is reported as *indeterminate* rather than scored in our favour. † below the 0.35 clean-accuracy gate: a collapsed model, not suppression. Student-t 95% intervals and per-seed values are in the supplement (§S8).

	arm (d_2 / attack)	**predicted shape**	**mean ASR (clean accuracy)** κ=0	0.5	1.0	2.0	**JT** $p_{\uparrow}$	**verdict**
1	`krum` / scaling	monotone rise	0.062 (0.57)	0.011 (0.65)	0.014 (0.68)	0.054 (0.62)	0.63	**refuted**
2	`reputation` / scaling	monotone rise	0.017 (0.78)	0.022 (0.78)	0.487 (0.77)	**0.779** (0.70)	1.7×10^{-6}	**confirmed**
3	`cos_krum` / pixel	**FLAT** (control)	0.490 (0.61)	0.450 (0.57)	0.153 (0.47)	0.086 (0.22)†	0.999	indeterminate
4	`coord_median` / pixel	rise, shallower	0.443 (0.77)	0.428 (0.76)	0.327 (0.75)	0.170 (0.69)	0.998	**refuted**

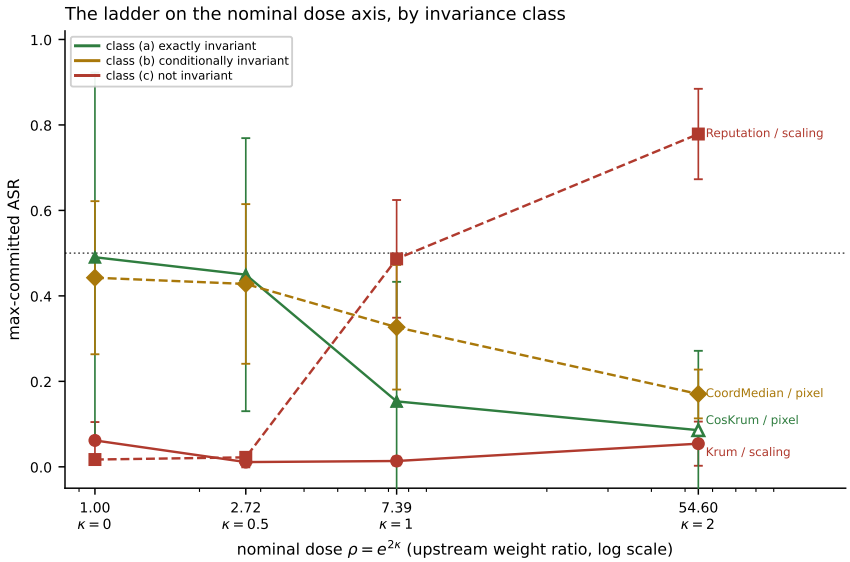

Figure A3: **The ladder and its frozen predictions.** The 16 cells on the nominal dose axis $\rho = e^{2\kappa}$ (n=5 per point, Student-t 95% intervals), coloured by Proposition 11 class. Of the two class-(c) arms whose statistics the framework's fragility claim is stated over, only reputation rises; ASR *falls* with the dose in three of the four arms, against the predicted rises. Hollow markers fail the 0.35 clean-accuracy gate and are uninterpretable. Drawn by `paper/figures/dose_response_figure.py`, which recomputes every value from `results/` and imports the arm construction from the scoring script so the two cannot diverge.

and the jump to 0.487 happens at the next rung at the same measured disturbance, so the *dose* moves ASR here while the flip rate does not.

Ecological validity: the real transforms do not lie on the synthetic curve. Pre-registered caveat (iv) required this check, and it fails in an informative way. RFA→Krum sits at ASR 0.779 with measured disturbance 0.667, where the synthetic ladder's nearest rung (0.600) gives 0.011: a residual of +0.767. NormClip→Reputation sits at 0.806 with disturbance 0.333, against 0.017 at the identity rung: +0.789. Two anchors do land: RFA→Reputation, 0.895 at disturbance 1.000 against 0.779 at 0.933 (+0.116), and both `cos_krum` anchors, which coincide with its curve (+0.000 and −0.044) because that statistic is invariant under all three transforms. This is not a range mismatch: read off the same live rounds, NormClip realizes ρ of mean 3.61 (max 4.90) and RFA mean 11.93 (max 17.38), both inside the ladder's $\{1.00, 2.72, 7.39, 54.60\}$ span. The conclusion is about our abscissa: **the rate at which a transform flips a defense's decision is not a sufficient statistic for what that transform does to the defense**: two transforms at matched flip rate differ by up to 0.77 in ASR. **And the axis they *are* unmatched on is now measured, not conjectured.** Adversary-correlated dispersion $\Delta_c := \mathbb{E}[c_i \mid i \in A] - \mathbb{E}[c_i \mid i \notin A]$ (§C) separates the transforms carrying those anchors from the ladder: RFA reaches $\Delta_c = -1.21$ and NormClip -0.72, against -0.34 at the ladder's most dispersed rung: 2–3.6× more extremely targeted at ρ that is 3.7 and 13× *smaller*, so **matched on the statistic, badly unmatched on the targeting, and ρ does not order Δ_c. This rescues no predictor**: Δ_c's magnitude does not order suppression

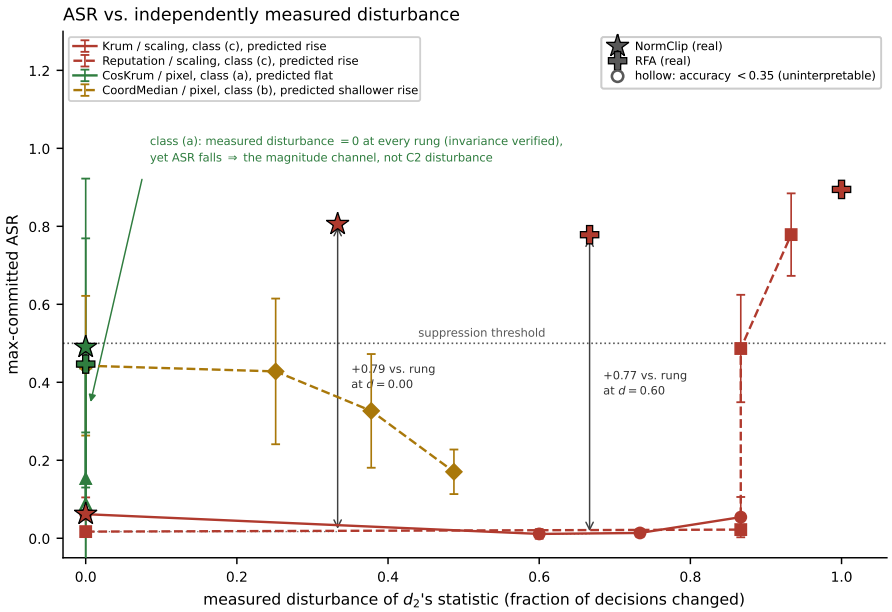

Figure A4: **Why a transform's flip rate is not a sufficient statistic for what it does.** ASR against the *independently measured* disturbance of each arm's own statistic (n=5, Student-t 95% intervals), the abscissa fixed before the freeze. Stars and crosses are the two *real* upstream transforms measured on these same (d_2, attack) cells (NormClip, RFA), placed at their own measured disturbance rate: at matched disturbance they sit $+0.77$ and $+0.79$ ASR *above* the synthetic curve, while their realized weight ratios (mean 3.61 and 11.93, maximum 4.90 and 17.38) fall *inside* the ladder's span. The class-(a) arm's four rungs coincide at disturbance exactly 0 ($0/120$ selections changed, maximum relative score drift 1.1×10^{-4}) and its ASR falls anyway: Lemma 2's magnitude channel, not C2 disturbance.

either; rfa→krum still fails at 0.779 at $\Delta_c = -1.21$, while `reputation`, targeted harder still on both attacks, has rep→coord_median suppressing, and FoolsGold's Δ_c is *positive* ($+0.663$). These figures are **post-hoc and exploratory**, computed with no ASR and rescoring nothing, and are tabulated in the supplement (§S11). Arm 4 has no real-transform anchor at all, since `coord_median` was not on the metric-swap menu, so its ecological validity is untested.

What we now claim less of. C2's causal content is established prospectively for one statistic (distance to the client median) and refuted for another (summed pairwise distance) at ρ up to 54.6; the pooled claim that measured disturbance predicts loss of suppression *across* defenses is refuted, and we no longer make it. Within-defense disturbance remains the only design the testability boundary admits, and §C.1's general statement stands untouched by this outcome; it is a statement about admissible designs, and this is the admissible design, run and reported as it landed. What changes is the empirical status of C2: one confirmed arm, not a validated screening mechanism.

D.2 The targeted intervention: isolating C2 from the attenuation channel

With the adversary's share pinned, how much of each arm's rise survives? **Reading `cos_krum`'s entry in Table 2.** Its -0.425 is the full-ladder endpoint, which lies below our accuracy floor; **the witness we rest on is instead its first, gate-passing rung** (0.173, §6). The rank discordance between $\Delta\Lambda_a$ and ΔASR across the two moving arms is §6's refutation of our own predicted ordering. This section gives the emitter and the per-artifact provenance for every cell of that table.

Round counts and per-row provenance, relocated here from the table's caption. *Admission* here is Table 2's Δ adm., the *support* of the admitted adversarial mass, which that caption keeps distinct from the mass itself ($\Delta\Lambda_a$); it never moves in any measured round of any arm: 60 rounds per CIFAR-10 `cifar_cnn` arm, 56 on the EMNIST-byclass replication and 72 on the ResNet18 replication, 368 in all, including the 48 per arm where reputation and `coord_median` carry adversarial mass at baseline, so the constancy is not an artifact of arms that never admit an adversary at all. The decision/admission dissociation is accordingly not one aggregator's property: it holds for a selector, a cosine-invariant selector, a coordinate-wise order statistic and a weighted averager, which is what the table's *Kind* column records. Row 5 replicates row 1 on a second dataset and a

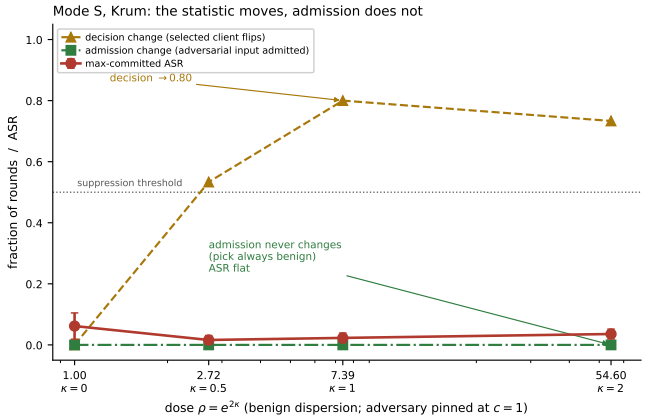

Figure A5: **Mode S, the Δ_c-pinned dose: Krum's statistic is disturbed, its admission of adversarial input is not, and suppression does not follow either.** The instrument of §5 at the four rungs of the benign-dispersion dose, with every adversary's own weight pinned (Δ_c exactly 0 at every rung, so the attenuation channel is cut by construction rather than controlled for): the selected client flips in up to 0.80 of rounds, the admitted adversarial support changes in none of them, and max-committed ASR does not rise past the pre-registered ± 0.15 practical-equivalence margin, whose lower side ASR's own floor already satisfies on this cell, so only the upper side is a test here (the supplement §S18). It is the Δ_c-pinned companion to the Δ_c-swept Mode A (Figure A6), and the same evidence Figure 1(a)'s middle annotation pools over four arms rather than Krum alone. Bars are Student-t 95% intervals, n=5; numbers recomputed per seed by `analyze_targeted_dose.py`.

differently configured CNN, with its channels measured *before* its ASR existed; row 6 is the pre-registered score-only control, row 1 with the magnitude channel closed, so its decision, admission and influence columns are row 1's own, and **it is the one row printed at two seed counts**, its Δ ASR at the n=20 top-up of §D.2 and its channel columns at the n=5 measurement they were made at, which is why its Δ ASR is never differenced against row 1's; row 7 replicates row 1 with *architecture* varying alone, its channels also measured before its ASR existed (§D.2). **Row 7 is printed last rather than beside row 5, which is where it belongs by subject**: four sites in this paper index the score-only control as row 6, and the emitter orders the rows so that adding a seventh cannot silently falsify them. The table is *not* an out-of-sample test. **Row 5 is the frozen n=3 in every column, its Δ ASR included**, because its three channel columns were measured at that seed count and were not re-measured: the same cell's Δ ASR at n=5 is -0.013, reported with its provenance in disclosure *(i)* below, and substituting it here would put one n=5 cell in a row of n=3 ones. **Row 6 is treated the other way and the asymmetry is deliberate rather than an inconsistency**: a pre-registration written before that top-up ran requires the row to report n=20, so the mixed count is printed in the n column and named in both captions instead of being avoided, whereas no comparable freeze exists for row 5 and nothing is gained by mixing it.

What Mode S does to the ladder's confirmation. Reputation was the ladder's one confirmed arm (§D.1); Mode S shows most of that confirmation was attenuation. With the adversary's share held fixed, reputation's rise at ρ=54.6 shrinks from the ladder's $+0.762$ to $+0.178$ (one-sided Welch t=-5.22, p=0.0017 that it is smaller), and what remains tracks a small *admission* change (0.000 $\rightarrow$ 0.020) rather than the large decision change (0.000 $\rightarrow$ 1.000), so H-admission is confirmed here too (Table A7). The confound is thus *explained*, not explained away: a genuine, much smaller, admission-driven effect survives it. The class-(a) control cuts the other way and refutes us: `cos_krum`'s selection is bit-identical at every rung (decision and admission both 0.000), yet ASR is 0.584 at the identity (it never suppresses) and *falls* 0.425 across the ladder ($p_{\downarrow}$=0.0008; the κ=2 rung is below the accuracy gate). Exact invariance of the statistic neither implies suppression nor holds it fixed; the magnitude channel moves it regardless. We report this as a refutation of Proposition 11's practical content, not a confirmation of its algebra.

Mode A: the two regimes, and where they hold. The payload dose gives Theorem 9 a single falsifiable shape: ASR single-peaked at r=1, since an over-attenuated adversary carries no payload (Regime B) and an over-amplified one is discarded by a rank aggregator (Regime A). It is **confirmed for `coord_median`**, the rank aggregator the regimes are stated for: ASR rises 0.124 $\rightarrow$ 0.315 $\rightarrow$ 0.443 into r=1 and falls 0.443 $\rightarrow$ 0.296 $\rightarrow$ 0.098 past it, both sides significant ($p_{\downarrow}$=3.0×10^{-4} amplified, 6.4×10^{-4} attenuated) at accuracy $\geq$ 0.74 throughout (Figure A6). It is **refuted for reputation**, which is monotone: ASR is maximal at the most attenuated rung (0.894 at r=0.135) and near zero once $r \geq 1$, because an attenuated adversary reads as *consistent* to a reputation score and stops being down-weighted. Reputation is not a rank aggregator, so the two-regime shape was never licensed for it; the split is the theorem behaving exactly where it applies and not where it does not.

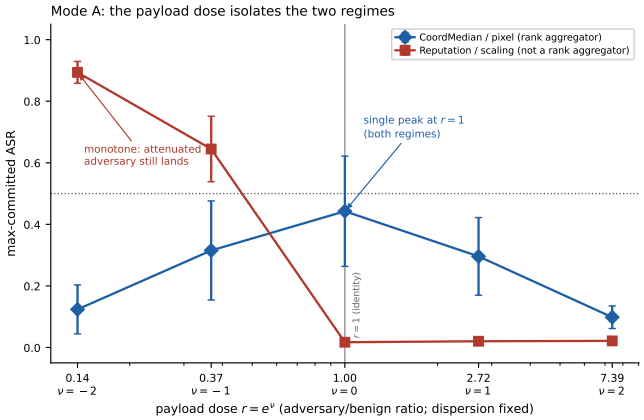

Figure A6: **Mode A, the payload dose: the two mechanism-preserving regimes, and where they hold.** With dispersion fixed and only the adversary/benign weight ratio r swept, `coord_median` (a rank aggregator, the case Theorem 9's regimes are stated over) is **single-peaked at** $r{=}1$ as the theorem predicts (over-attenuated adversaries carry no payload, over-amplified ones are discarded) while reputation is monotone, off-spec and refuted for that shape. Mode A is the Δ_c-*swept* companion to Mode S's Δ_c-pinned design (Proposition 13); Mode S itself is Figure A5. Bars are Student-t 95% intervals; numbers recomputed per seed by `analyze_targeted_dose.py`.

Table A7: Mode S (statistic-only), $n{=}5$ per arm ($n{=}8$ for `cos_krum`), per-attack mean ASR with the two measured channels. The decision channel moves far more than ASR throughout. **The *admission* row is the admitted adversarial *mass*** Λ_a, i.e. Table 2's influence column $\Delta\Lambda_a$ and *not* its Δ adm. column, which records the support of that mass and is 0.000 in every measured round of every arm; this section's hypothesis names (H-admission) are stated at the mass. † below the 0.35 accuracy gate. ‡ the pre-registered replication arm, *rank-discordant* by design (smallest decision change, largest admission change of any interpretable arm) and scored against thresholds frozen separately at `06d9bce` before `results/dose_replication/` existed: $\Delta{=}{+}0.098$ falls below reputation's $+0.178$, so the cross-arm admission *ordering* is refuted while the negative replicates on a different invariance class (below). Verdicts otherwise scored against the readings frozen at `5130cec`; per-seed values in the supplement (§S8).

arm	channel	$\rho{=}1$	2.72	7.39	54.6	test	verdict
`krum` / scaling	ASR	0.062	0.016	0.023	0.036	$p_\uparrow{=}0.70$	H-admission
	decision	0.000	0.533	0.800	0.733		confirmed,
	admission	0.000	0.000	0.000	0.000		H-stat. refuted
`reputation` / scaling	ASR	0.017	0.019	0.025	0.195	rise $+0.178$	H-admission
	decision	0.000	0.933	1.000	1.000	vs. ladder $+0.762$	confirmed
	admission	0.000	0.002	0.008	0.020	($p{=}0.0017$)	(mostly atten.)
`cos_krum` / pixel	ASR	0.584	0.411	0.295	$0.159^\dagger$	$p_\downarrow{=}8{\times}10^{-4}$	**refuted**
	decision	0.000	0.000	0.000	0.000	$\lvert\Delta\rvert{=}0.425$	(Prop. 11
	admission	0.000	0.000	0.000	0.000		practical)
`coord_median` / pixel‡	ASR	0.443	0.442	0.490	0.541	rise $+0.098$	**ordering**
	decision	0.000	0.232	0.358	0.482	$< +0.150$	**refuted**;
	admission	0.000	0.005	0.013	0.033	($p_\uparrow{=}0.069$)	negative repl.

A pre-registered replication on a second downstream defense, and what it refutes. One arm cannot establish which channel matters, and the two Mode-S arms above move together on both candidate predictors: reputation is higher on *both* decision and admission change and also had the larger rise, so nothing in that pair separates them. `coord_median` under the pixel backdoor is *rank-discordant* and therefore does: of the interpretable Mode-S arms it has the **smallest decision change** (0.482) and the **largest admission change** (0.033), so the decision reading predicts its rise is the *smallest* of the three and the admission reading predicts the *largest*. It is also a structurally different defense, a coordinate-wise order statistic (Proposition 11(b)) against Krum's selector (class (c)), and it suppresses the pixel backdoor at usable accuracy, so there is suppression to lose. Thresholds were frozen against the two published rises (confirm if $\Delta > +0.178$, refute if $\Delta < +0.150$, indeterminate between) at commit `06d9bce` before `results/dose_replication/` existed, with the identity rung imported bit-identical and 15 new runs spent (`experiments/pre_registration_dose_replication.md`).

The outcome is $\Delta = +0.098$ at n=5: the admission *ordering* is refuted. The arm with the largest admission change rose *less* than an arm with a smaller one, so **admission magnitude does not order outcomes across arms**. Neither does the decision reading, which predicted the *smallest* rise of the three when $+0.098$ is the middle one, so this arm refutes a rank prediction from *each* candidate channel. The pooled model agrees: re-fit with these cells it is still not confirmed (R^2=0.239; the admission coefficient's one-sided p moves $0.055 \rightarrow 0.041$, but the adversarial-share coefficient is negative, so the pre-specified conjunction fails). **What does replicate is the negative**: a decision change of 0.482 costs no suppression by the frozen margin, at mean accuracy 0.698, so the accuracy gate does not bind and the flagship result now holds on two downstream defenses in *different* invariance classes. Two qualifications, the first pre-specified in the freeze and the second post-hoc: this cell's identity ASR is 0.443, so equivalence is to the *identity rung* and not to a low absolute ASR (an "every rung <0.5" criterion has no headroom here and was deliberately not used); and the rise is positive in $5/5$ seeds (paired t=3.39, one-sided p=0.014, *not* pre-registered), so on this class the cost of disturbing the decision is **small and consistent rather than zero**. All numbers are recomputed per seed by `analyze_dose_replication.py`, which imports its rung builder and trend test from the frozen suite's scorer.

Two admissible estimands, opposite conclusions on the same cell. This is the paper's central empirical result. On this identical cell (matched aggregate displacement, decision change and clean accuracy, Table A3), **the confounded design does not lose power, it points the other way**: at n=20 the outcome-gated ladder moves -0.273 ($0.478 \rightarrow 0.205$) where the instrument moves $+0.125$ ($0.478 \rightarrow 0.603$), the two paired 95% intervals $[-0.334, -0.213]$ and $[+0.095, +0.155]$ disjoint and both excluding zero. Closing the attenuation channel does not sharpen a weak positive effect, it **inverts it**. **The two designs are not the same intervention**, and Table A3 states the one non-match: Λ_a moves in both (0.081 against 0.033, both n=5), so neither column is a clean statistic-only intervention and no p-value attaches to the difference between them. What the contrast establishes is that the choice between two admissible estimands sets the inferred direction, so a criterion calibrated against the confounded contrast would have been calibrated against the confound, and a cross-defense contrast would have read the fall as evidence about the statistic. That is why the identification boundary of §C.1 is not a technicality about power.

What the two arms jointly establish, stated formally. The arms above and the class-(a) control (`cos_krum`) are between them a *constructive* answer to the question C2 was written to settle. Both halves are witnessed by runs already reported, so the following formalizes measurements rather than adding a claim.

Proposition 12 (Empirical non-identifiability of preservation levels; Proposition 3 of the body, restated with proof)**.** *Within the positive per-client rescaling class there exist transform families T_1, T_2 and downstream defenses d_2 such that (a) T_1 and T_2 induce arbitrarily different statistics and* decisions *at d_2 while inducing the* same *admitted adversarial mass and the same suppression; and (b) T_1 and T_2 induce the* same *admitted adversarial mass (indeed the same statistic and the same decision) while inducing* different *suppression. Consequently no predicate of the downstream statistic alone determines the admitted adversarial mass, and no predicate of the admitted mass alone determines suppression.*

Proof. Both halves are exhibited by runs reported in this section. (a) Mode S into Krum under model-scaling, κ=0 against κ=2: ρ moves from 1 to 54.6 and Krum's decision changes in up to 0.80 of rounds, while the support of the admitted adversarial mass changes in none of the 60 measured rounds (a floor on this arm: no adversary is admitted at any rung), the adversarial coefficient share is constant at 0.266667, and suppression shows **no evidence of a practically meaningful change under the pre-registered ± 0.15 practical-equivalence margin** ($|\Delta|$=0.026). (b) `cos_krum` under the pixel backdoor: by Proposition 11(a) its statistic is invariant, and its decision and admission are measured bit-identical (0.000) at every rung, yet ASR falls 0.173 by the first rung alone, at ρ=2.72, where the accuracy gate still passes ($0.599 \rightarrow 0.567$; paired t=-1.976, one-sided p=0.044, falling in $5/8$ seeds). □

What the proposition does not say. A preserved statistic is not thereby uninformative; it is not *sufficient*, remaining evidence about what d_2 reads and none about what it suppresses.

The corollary is why no statistic-only metric can stand in for suppression: no metric defined *solely* on preservation of the downstream statistic can identify attack-suppression equivalence, because (a) exhibits statistic disturbance without suppression change and (b) exhibits suppression change without statistic disturbance. **Scope, stated immediately**: this is a non-existence claim *within the positive-rescaling class on these defenses and attacks*, established by exhibited counterexample: **not** a claim that no metric of any kind can order security, and not a claim about transforms outside the class.

Proposition 13 (ρ and Δ_c are independently controllable)**.** *Within the same class, the statistic-level dispersion $\rho = \max_i c_i / \min_i c_i$ and the adversary-correlated dispersion $\Delta_c = \mathbb{E}[c_i \mid i \in A] - \mathbb{E}[c_i \mid i \notin A]$ are not functionally related: there is a family with $\Delta_c \equiv 0$ exactly while ρ ranges over $[1, 54.6]$, and a family with the statistic's dispersion held fixed while Δ_c ranges over $[-1.138, +2.469]$. Hence neither quantity determines the other.*

Proof. By construction, with the realized values tabulated in the supplement (§S11) and computed with no ASR involved. Mode S pins every adversarial coefficient at $c{=}1$ and spreads only the benign clients, giving $\Delta_c \equiv 0$ at every ρ up to 54.6; Mode A holds the benign weights uniform and sweeps only the adversary-to-benign ratio r, moving Δ_c over $[-1.138, +2.469]$ at fixed statistic dispersion. □

This is why the two instruments were the right ones, and it makes the supplement's §S11 warning precise: **a transform reported by ρ alone leaves Δ_c entirely free**, so ρ cannot be a sufficient statistic for what the transform does to a defense. It does *not* follow that Δ_c is a predictor: §D.1 reports that its magnitude fails to order suppression either.

The dissociation is not Krum-specific: all four aggregators, all four channels. Table A8 collects the endpoint change ($\kappa{=}0$ against $\kappa{=}2$) in each measured channel under Mode S for every downstream defense in the paper. **Recomputed, not transcribed**: every cell is emitted by `build_channel_table.py`, which only reads frozen artifacts under `results/`, `admission_measurement.json`, `femnist_admission.json` and `resnet18_admission.json` for the channel columns, the `targeted_dose/`, `dose_replication/`, `dose_femnist/` and `dose_resnet18/` summaries for the ASR column; it also asserts that the identity rung displaces nothing (0.00e+00 for all seven rows). The sixth row's aggregate column comes from `displacement_decomposition.json` and its ASR from `score_only/summary.json`, both read-only; the emitter refuses to merge that arm into the ASR sources it shares cell keys with, so it cannot shadow the first row's published number. The influence column is **aggregator-specific by necessity** (a selector has no weights) and its instantiation is: adversarial share of the Weiszfeld/reputation weight mass for the averager, adversarial share of selected coordinates for the order statistic, and the indicator that an adversarial update is selected at all for the two selectors. The fifth row replicates the first row's arm on a second dataset and a differently configured CNN (EMNIST-byclass with `simple_cnn`) with **its channels measured prospectively, before its ASR existed**; the result is reported below. **The seventh row replicates it with architecture varying alone** (CIFAR-10 with `resnet18`), its channels likewise measured before its ASR existed, and it is the one row that shares *both* the first row's dataset and its cell key: the ASR lookup is keyed on the triple (dataset, model, cell) precisely so that the two cannot shadow each other, and the emitter raises rather than resolving a collision by file order (§D.2). **What the Δ adm. column pools**: the four `cifar_cnn` CIFAR-10 rows contribute 60 measured rounds each, 240 in total, and adversarial mass is present at baseline in 100 of those 240 rounds while its support moves in none of them, which is the count Figure 1's annotation reports and why that annotation carries no single arm's name. **That** 100 **is not spread evenly, and the split is the answer to the floor objection**: it is 0 rounds for Krum, 4 for `cos_krum`, and 48 each for reputation and `coord_median`, so the admission reading is carried by the two arms that have room and not by the two that are at or near the floor. The emitter prints each arm's pre-transform admitted fraction, averaged over the rounds an adversary actually joins, and asserts that it is identical at the identity rung and at $\kappa{=}2$, since a level that moved with the dose would not be a baseline. **It also holds the aggregator fixed and varies the attack instead**, which is the only way to tell an aggregator's floor from an attack's, and the four CIFAR-10 aggregators read as follows under `committed_scaling` against `committed_pixel`, as level (nonzero rounds of 12): Krum 0.000 (0) against 0.333 (4), `cos_krum` 0.083 (1) against 0.083 (1), reputation 0.0014 (12) against 0.251 (12), `coord_median` 0.156 (12) against 0.286 (12). So Krum's floor belongs to the attack: a model-scaling adversary is a norm outlier and Krum is built to reject one, while under the pixel backdoor the same Krum admits adversarial mass at baseline. That weakens the reading on which the floor is a property of the aggregator, and it does not remove the limitation, because the adjudicating arm does commit to scaling.

A second dataset and a differently configured CNN, pre-registered. The fifth row's arm is the first row's arm re-run with EMNIST-byclass (62 classes, $1{\times}28{\times}28$) and `simple_cnn` in place of CIFAR-10 and `cifar_cnn`, everything else held: Mode S, Krum as d_2, the committed model-scaling attack, the same four frozen rungs $\kappa \in \{0, 0.5, 1, 2\}$. The decision rule lives in `experiments/pre_registration_dose_femnist.md`, git-committed at `478b555` before any EMNIST-byclass ASR existed, and the runner refuses to start unless its `PREREG_COMMIT` matches that hash.

One naming correction, made here rather than left to be inferred. Earlier drafts labelled this arm "FEMNIST" in the tables while the prose said EMNIST-byclass, and the tables were the wrong ones: the `femnist` key loads *torchvision* `EMNIST(split="byclass")` and partitions it with this paper's own label-Dirichlet split, **not** LEAF FEMNIST's partition by writer. So the arm is a second dataset and a differently configured CNN, and it is **not** a more federated partition than the CIFAR-10 arm; it is the same partition scheme on different data. Every table now reads EMNIST-byclass. The frozen artifact paths, the pre-registration filename and the JSON dataset key retain the `femnist` spelling because renaming a committed artifact would break the provenance the rest of this appendix rests on.

The premise was measured first, with no ASR anywhere. Krum's selection changes on 0.643–0.714 of rounds across the three disturbed rungs while the support of the adversarial mass it admits is unchanged in all 56 measured rounds (again a floor: no adversary is admitted at any rung), so the arm has a real disturbance to

Table A8: **The four channels, all four aggregators, under Mode S.** Endpoint change $\kappa{=}0 \to \kappa{=}2$. **The last three channels are the levels of Def. 1**, decision P_3, admission P_4 and ASR P_5, so every row is a filled-in ΔP_3–ΔP_5 report of the kind §4's clause (4) asks for, and the arm is classified by the highest level preservation survives at; P_1 and P_2 need no measurement, being settled algebraically by Prop. 11 before any run, and their columns are absent for that reason rather than unrun. **The first column, Δ agg., is the relative displacement of the aggregate the defense *emits***, $\|\mathrm{agg}(T(U)) - \mathrm{agg}(U)\|/\|\mathrm{agg}(U)\|$: what enters training, and **not** the value of a scoring statistic; the measured channels are aggregate displacement, decision, admission and influence. The dissociation between the *decision* channel and the *admission* channel is not a property of one aggregator: it holds for a selector, a cosine-invariant selector, a coordinate-wise order statistic and a weighted averager. **The two channels do not have equal room, and the Δ adm. column is a support flip rather than a magnitude**: averaged over the rounds an adversary joins, the pre-transform admitted fraction is 0.000 on every Krum row and 0.083 on Cosine-Krum, which is one round in twelve, against 0.286 on coord. median and 0.0014 on reputation, each nonzero in all twelve. So the decision reading rests on Krum, whose decision channel is unconstrained and moves on 0.733 of rounds, and the admission reading rests on coord. median and reputation, the two arms with room; a zero on a Krum row is a floor and is not evidence either way. The table does **not** order the arms ($\Delta\Lambda_a$ and ΔASR disagree in rank, which is this section's refutation stated in one row-pair) and it is *not* an out-of-sample test: all rows are Mode-S ladders reported elsewhere in this paper. **The sixth row is the score-only control**: the same arm as the first with the magnitude channel closed, so its Δ agg. is the re-selection component alone and its decision, admission and influence columns are the first row's own numbers, score-only leaving the selection untouched. **That is why its 0.986 *exceeds* the first row's 0.892 with the magnitude channel closed**: 0.892 is the vector sum of that same 0.986 re-selection component and a 0.973 rescaling component, which partly cancel, so a larger number in this column is a *narrower* channel and not a wider one (§S9). **The seventh row varies architecture alone**: the first row's cell on `resnet18` instead of `cifar_cnn`, with dataset, N, K, f, α, attack and defense all at their frozen CIFAR-10 values, so it is the only row that differs from the first in a single factor. Its channel columns were measured *prospectively* (`results/resnet18_admission.json` records "No ASR. Freezes nothing." and predates every ResNet18 ASR in this paper), and its two unrun interior rungs are why §D.2 reports no trend statistic for it. n **is the seed count behind the Δ ASR column and nothing else**: the channel columns are per-round measurements, whose counts are given above. It is emitted per row as $\texttt{n_seeds} = \min(n_0, n_2)$ over the row's two rungs and it is **not uniform**, reading 3 for the second-dataset replication, 8 for `cos_krum` and 20 for the score-only row. **That last pair of cells is the one thing in this table the emitter does not produce, and we say so rather than let the two disagree silently**: `build_channel_table.py` reads `results/score_only/` and returns the published $n{=}5$ contrast, -0.023, whereas what is printed here is the $n{=}20$ top-up of App. D.2, assembled from three artifacts by that arm's own analyzer with both legs' seed sets asserted first. So row 6's Δ ASR and n are read from an analyzer rather than from the emitter, and *both* values stand at their own seed counts, -0.023 at $n{=}5$ in App. D.2 and -0.009 at $n{=}20$ here. The emitter also records `seeds_match` on the six rows whose two rungs come from one paired suite, and it is true on all six; the score-only row's ASR is read from a separate directory, so it carries no such field.

		magnitude	statistic	attenuation		total	
Aggregator	Kind	Δ agg.	Δ dec.	Δ adm.	$\Delta\Lambda_a$	Δ ASR	n
Krum	selector (pairwise distance)	0.892	0.733	0.000	0.000	-0.026	5
Reputation	weighted averager (cross-round state)	0.659	1.000	0.000	0.020	$+0.178$	5
Cosine-Krum	selector (cosine distance)	1.297	0.000	0.000	0.000	-0.425	8
Coord. median	coordinate-wise order statistic	0.670	0.482	0.000	0.033	$+0.098$	5
Krum (EMNIST-byclass)	selector (pairwise distance)	0.842	0.643	0.000	0.000	-0.017	3
Krum, score-only	selector, magnitude closed	0.986	0.733	0.000	0.000	-0.009	20
Krum (ResNet18)	selector (pairwise distance)	0.824	0.556	0.000	0.000	-0.041	3

test. Had the decision change come out near zero the arm would have been reported ineligible rather than run: that is the `cos_krum` failure mode, and this check is what distinguishes the two.

Suppression shows no evidence of a practically meaningful change. $\Delta = -0.017$ (mean ASR $0.027 \to 0.010$) against the frozen ± 0.15 practical-equivalence margin, whose lower side the identity mean 0.027 satisfies by arithmetic before any run, so only the upper side is a test here and what it returns is a bound of -0.001 on any rise (the supplement §S18), beside CIFAR-10's -0.026 on the same cell, with mean clean accuracy

rising $0.767 \to 0.835$ so the accuracy floor binds on no rung. Under the frozen rule this is a **replication on a second dataset and a differently configured CNN**. **That is narrower than the wording the pre-registration froze**, which reads "a second dataset and architecture": `simple_cnn` and `cifar_cnn` are two configurations of one shallow-convnet family, so we report the narrower phrase and leave the frozen file unedited, claiming less than a rule licenses being always permitted where claiming more is not. The prohibition we keep verbatim is the one that matters: never "generalizes", which one further dataset does not license.

Three disclosures, each of them from the freeze rather than from hindsight. (i) $n{=}3$ seeds per rung, not 5, fixed in advance to match the EMNIST-byclass payoff matrix's own trial budget. **The frozen verdict is that $n{=}3$ one, and the same ladder also reads at $n{=}5$ with no new run needed**: the comparability amendment of §D.3 ran seeds 45 and 46 of this exact cell at all four rungs, so pooling them with the frozen three, frozen rows winning and the frozen directory never written, gives $\Delta{=}-0.013$ beside -0.017, mean ASR $0.024 \to 0.011$ at accuracy $0.758 \to 0.841$: the same label under the same unmoved margin. Two more seeds *widen* the interval across zero, $[-0.034, +0.009]$ against $n{=}3$'s $[-0.033, -0.001]$, so they remove an apparent small fall rather than confirm one, and **one of the five carries the opposite sign** (seed 45, $+0.015$, its identity rung landing at ASR 0.000018), which a 3-seed mean cannot show. Both counts are reported and neither replaces the other (`analyze_dose_femnist.py --pooled`). Those two seeds were pre-registered in the comparability document before they ran and are post-hoc in this arm's own. (ii) The secondary Jonckheere–Terpstra test was specified to detect a monotone *rise* (the outcome that would have refuted the flat prediction with attenuation already closed by construction) and the observed trend is a *fall* ($J{=}16.0$, $z{=}-1.563$, $p_{\uparrow}{=}0.941$; at $n{=}5$, $J{=}57.0$, $z{=}-1.211$, $p_{\uparrow}{=}0.887$, the same direction). We report it and we do **not** score it as support for flatness; the pre-registration states in advance that at $n{=}3$ a null from this test is not evidence of flatness, and two more seeds do not convert one into evidence. (iii) The two branches had unequal headroom. With the identity rung at ASR 0.027, refuting us required mean ASR($\kappa{=}2$) > 0.177 against a ceiling of 1 and was fully reachable, whereas the indeterminate branch ($\Delta < -0.15$) was arithmetically unreachable. **Equivalence here is to the identity rung, not to a low absolute ASR.**

Harness sanity, which is an identity check and not corroboration. The $\kappa{=}0$ rung is identity$\circ$Krum, i.e. Krum alone, and it lands *bit-identically* on the EMNIST-byclass Krum figure from our earlier payoff-matrix sweep (0.0269784552 at accuracy 0.7670795973). Both pipelines are deterministic given a seed and both run three trials from base seed 42, so the agreement shows they execute the same protocol rather than a similar one: the strongest form the pre-registered check can take, and for the same reason **not** independent corroboration of that figure: the same computation run twice cannot corroborate itself. **The pooled seeds of disclosure *(i)* now carry the same check across two runners.** Having already read those two seeds from the comparability suite, we ran this arm's own runner over the same eight (rung, seed) cells, pre-registered as an amendment at `ad5479b` before those runs: it reproduces all eight to ten decimal places, largest gap 0.00e+00, so the pooled reading is not an artifact of which suite measured it (`analyze_dose_femnist.py --pooled`). The same reading applies: identical protocols, not independent confirmation.

A second architecture family, pre-registered, with the gate scored first. The seventh row's arm is the first row's arm with **architecture varying alone**: `resnet18` (11.2M parameters, BatchNorm replaced by GroupNorm-8, `conv1` adapted to $32{\times}32$, `maxpool` $\to$ Identity) in place of `cifar_cnn`, and CIFAR-10, $N{=}10$, $K{=}5$, $f{=}0.2$, α, the committed model-scaling attack and Krum as d_2 all at their frozen values. The decision rule is `experiments/pre_registration_dose_resnet18.md`, git-committed at `89046f6` before any ResNet18 ASR existed, and the runner refuses to start unless its `PREREG_COMMIT` matches that hash. **Two endpoint rungs only**, $\kappa \in \{0, 2\}$ of the frozen four-rung ladder ($\rho = 1.00 \to 54.60$), at seeds 42–44: the interior rungs are not run, **so this arm carries no trend statistic at all** and the runner prints that restriction rather than leaving it to be inferred.

The gate was scored before the ladder, and it is what makes the arm interpretable. No ResNet18/Krum figure existed on disk, so the pre-registration named the void branch first: unless Krum alone suppresses the committed scaling attack on this architecture (mean ASR < 0.5 at mean clean accuracy ≥ 0.35) *there is no suppression for the ladder to preserve and the arm is void, not negative*. It passed: at $\kappa{=}0$, mean ASR 0.071 at mean clean accuracy 0.536, $n{=}3$. Had it failed, the runner refuses the $\kappa{=}2$ leg outright rather than spending it, and we would be reporting a void.

Suppression does not rise, and under the rule this arm was frozen against the negative replicates on a second architecture family. $\Delta = -0.041$: mean ASR 0.071 at $\kappa{=}0$ against 0.030 at $\kappa{=}2$, its magnitude inside the frozen ± 0.15 practical-equivalence margin, beside `cifar_cnn`'s -0.026 and EMNIST-byclass's -0.017 on the same cell. **We do not write "shows no evidence of a practically meaningful change" about this arm, and disclosure *(v)* is why**: §S13 defines that phrase by the paired 95% interval lying inside the margin, and at $n{=}3$ this arm's does not. **Both endpoints are $n{=}3$ on the same three seeds**, so the difference of rung means and the paired mean difference are the same number to four places (-0.0410), and `seeds_match` is true on the row; no leg is a mixed-n comparison. Mean clean accuracy *rises* $0.536 \to 0.557$, so the 0.35 floor binds on neither rung. Under the frozen three-way rule ($|\Delta| < 0.15$ replicates, $\Delta > +0.15$ architecture-specific, $\Delta < -0.15$ a fall) this is a **replication across architecture families**, and it is the arm that earns

the phrase: `simple_cnn` and `cifar_cnn` are two configurations of one shallow-convnet family, whereas an 11.2M-parameter residual network is not.

Five disclosures about this arm, four of them limits. (i) **The channels were measured prospectively and the admission column is vacuous here**, exactly as on rows 1 and 5. `results/resnet18_admission.json` records "No ASR. Freezes nothing." and predates the ladder; it reads the decision disturbed and the admitted adversarial mass not (Δ dec. 0.556 against Δ adm. 0.000 at the top rung), which is the premise the flagship arm rests on and the reason the cell was eligible at all. **But adversarial mass is never present at baseline in any of this arm's 72 measured rounds**, so its Δ adm. $= 0.000$ is vacuous and we do not report it as evidence; `count_admission_rounds.py` prints that verdict per row rather than letting a zero speak for itself. (ii) **Clean accuracy is far below this architecture's own reach**: 0.536–0.557 against the 0.66–0.76 the ResNet18 mini-sweep of the supplement §S20.2 attains in the same regime, ResNet18 at 50 rounds being far from converged. The gate is a floor and not a claim of a converged model, and no statement here is about converged training. (iii) **One cell, one attack, and it widens nothing else.** It does not add a dataset, an N, a K, an f, a round count or an attack, it carries no composition suite, and it does not touch the sign-reversal cell of §D.3. (iv) **Why this arm fits an overnight budget when §D.3's disclosure *(iv)* prices ResNet18 at 65 h**: that figure is a *comparability* cell, two designs $\times$ four rungs $\times$ five seeds $= 40$ runs at $\approx$5900 s. This arm is one design $\times$ two endpoint rungs $\times$ three seeds $= 6$ runs, which is 13.4 h as measured. The two numbers price different objects and neither supersedes the other. (v) **The interval is wider than the margin, so the frozen ± 0.15 is load-bearing for this arm and, among the arms of the margin ladder, for no other.** The paired 95% interval is $[-0.183, +0.101]$ (per-seed ASR $0.114/0.050/0.048$ at κ=0 against $0.008/0.034/0.047$ at κ=2 on seeds 42–44, hence per-seed differences -0.106, -0.017, -0.000 at sd 0.057; the Reproducibility statement's per-seed promise covers this arm too, and these are those values), so this arm's $m^* = \max(|\text{lo}|, |\text{hi}|)$ is 0.183 against the comparability ladder's binding 0.034 (§S13), and the body's ladder claim is scoped to the supplement's Table S10 for that reason rather than stated of the paper as a whole. TOST against the frozen margin *does* reject, at the 90% interval α=0.05 reads: p=0.040, post hoc, the closest call in Table S9, and the smallest margin at which it still rejects is 0.137. **So this arm replicates under the rule it was frozen against and establishes equivalence at no tighter margin**; three seeds on two rungs is what that buys, and the remedy is seeds rather than wording. **And that m^* is set by the side of the margin that was never at risk**: 0.183 is $|\text{lo}|$, while the identity-rung mean ASR on this cell is 0.071, so a fall of 0.15 is arithmetically unavailable and the lower leg is satisfied before any seed runs (the supplement §S18). Read on the one leg that is a test here, the binding quantity is the 95% upper bound $+0.101$, which is inside the frozen margin. We keep the two-sided verdict above because it is the conservative one and it is what the freeze scores, and we state beside it that ± 0.15 being load-bearing for this arm is a property of the two-sided form rather than of the evidence against a rise.

The same cell at a $4\times$ longer horizon, pre-registered, which answers (ii) with a measurement. Disclosure (ii) concedes that ResNet18 at 50 rounds is far from converged, and the objection that follows is that a security negative read at 0.536–0.557 clean accuracy may be an artifact of an undertrained model rather than a property of the composition. We therefore re-ran this exact cell at 200 rounds: same N=10, K=5, f=0.2, the same committed model-scaling attack, Krum as d_2, the same endpoint rungs $\kappa \in \{0, 2\}$, the same seeds 42–44 and the same imported `run_one`, with only the round count changed. The decision rule is `experiments/pre_registration_dose_resnet18_converged.md`, git-committed at `ca89f96` before the runner was written, and the runner refuses to start unless its `PREREG_COMMIT` matches that hash; its harness check first reproduced a stored 50-round per-seed row through the same imported code path to $< 10^{-9}$, a value comparison against stored rows rather than a hash of a results file. Output: `results/dose_resnet18_converged/summary.json`.

Both gates passed and the negative survives the longer horizon, but the margin does not bind, because the identity rung has less headroom than the margin. Mean clean accuracy reaches 0.652 at κ=0 and 0.688 at κ=2, clearing the pre-registered 0.65 convergence-label floor and reaching the 0.66–0.76 band (ii) names for this architecture in this regime (the κ=0 rung sitting just under its lower end), and the parent arm's own validity gate passes as well. $\Delta = \mathbf{-0.022}$, mean ASR 0.045 at κ=0 against 0.023 at κ=2, n=3 on one seed set, so $|\Delta| < 0.15$ and the frozen three-way rule returns the verdict it returned at 50 rounds. **We do not report this as a strengthened replication, and the freeze forbade doing so in advance**: an identity rung at 0.045 cannot fall by more than 0.045, so $|\Delta| < 0.15$ was attainable by arithmetic alone, and the pre-registered informativeness condition (identity ASR at or above 0.15, so that the margin binds) *failed*, exactly as that document predicted it would. The identity rung reads 0.045 at 200 rounds and 0.071 at 50 rounds, each stated at its own horizon; **no difference between the two is taken**, here or anywhere, because a longer horizon lengthens both the training and the adversaries' poisoning exposure, so a cross-horizon Δ would confound them.

What the longer horizon does not buy, from the freeze rather than from hindsight. (a) n=3 again: the paired 95% interval is $[-0.182, +0.138]$ at sd 0.064, with per-seed differences -0.096, $+0.015$ and $+0.016$ (per-seed ASR $0.111/0.012/0.011$ at κ=0 against $0.015/0.027/0.027$ at κ=2, which the Reproducibility statement's per-seed promise covers), so the mean is carried by seed 42, the one seed whose identity rung was appreciably backdoorable at all. This arm's m^* is 0.182, so like its 50-round parent it fits inside no margin

below the frozen 0.15, and *no* equivalence is claimed at any margin tighter than that one. That m^* is |lo| and not |hi|, so the arithmetic of the paragraph above carries over from the point rule to the interval: the lower leg is satisfied before any seed runs, and the one leg that tests anything returns a bound of $+0.138$ on a rise (the supplement §S18). The remedy for $n{=}3$ is more seeds, which this arm deliberately does not spend. (b) It is *not* a row of the comparability ladder, whose emitter reads only the frozen artifacts (§S13), so the ladder's binding m^* and the body's scoping of which single arm needs a margin above 0.15 are claims about that table and are unchanged by this arm. (c) Two endpoint rungs again, never described as a four-rung ladder, so no trend statistic is computed here either. (d) The six runs cost 51.4 h of wall time, 6.6–14.3 h each, the spread being contention with a concurrent arm on the same GPU rather than a property of the cell; the freeze budgeted 40–54 h from two measured rates that disagreed and named the slower one as the rate it planned against.

What the instrument does *not* hold fixed, named in full, and what it pins exactly. Mode S pins the *adversarial* coefficients; it does not pin the benign ones and was never claimed to. So the dose also moves (a) Krum's pairwise distances, hence the statistic itself; (b) *which* benign client Krum selects, in $0.533/0.800/0.733$ of rounds at $\rho{=}2.72/7.39/54.6$; (c) the **magnitude and geometry of the update actually aggregated**, since the selected client enters scaled by its own coefficient; and (d) the model trajectory, which diverges across rungs in consequence. Channel (c) is not a conjecture but a **number we already measure and print**: the Δ agg. column of Table A8 is $\|\mathrm{agg}(T(U)) - \mathrm{agg}(U)\|/\|\mathrm{agg}(U)\|$, the relative displacement of the aggregate the defense *emits*, and it reads $\mathbf{0.892}$ on this cell. The supplement §S9 splits that 0.892, as vectors, into 0.986 of *re-selection* (a different benign client is chosen, and in $0/11$ of the rounds where that happens it is an adversary) and 0.973 of *rescaling* (the chosen client arrives scaled by its own coefficient); for `cos_krum` the displacement is *entirely* rescaling, since $0/15$ selections change at any rung. The claim this arm licenses is therefore sharper than a claim about the statistic alone: **a** 0.892 **displacement of the emitted aggregate, together with decision changes in up to** 80% **of rounds, leaves suppression inside the pre-registered practical-equivalence margin, provided the adversarial coefficient share does not move.** That last reading is one-sided on this cell: the identity-rung mean ASR is 0.062, so the margin's lower leg is satisfied by arithmetic before the dose is applied, and what the interval bounds is a *rise*, at $+0.026$ against the frozen 0.15 (the supplement §S18). **What is pinned is pinned exactly**: $c_{\mathrm{adv}}{=}1.0$ at every rung with maximum deviation 0.00e+00 and the share constant to 3.5e−07 on CIFAR-10, 3.9e−07 on EMNIST-byclass: float32 norm read-back noise, not slack in the instrument (`measure_admission.py`). Naming (a)–(d) is what keeps the claim to the one channel the instrument closes by construction.

Closing channel (c) as well: the pre-registered score-only control. Channel (c) (the magnitude and geometry of what is actually aggregated) is the one confound Mode S leaves open, so we closed it too rather than argue about it. **Score-only Krum** scores on the transformed stack exactly as before and then aggregates the *untransformed* selected update: the dose can change *which* client contributes and nothing whatever about *what* that client contributes. Statistic and decision are disturbed identically to the uncontrolled arm (the selection is computed on the same transformed stack, so Δ dec. is the same 0.733) while Δ agg. is no longer 0.892 but the re-selection component alone, 0.986 relative to $\|u_{s_0}\|$ (the supplement §S9), the rescaling component being closed by construction. This is what the control demands, in its own terms: hold the selected update's magnitude and direction fixed while perturbing the selection statistic independently.

Frozen before the arm existed. The decision rule is `experiments/pre_registration_score_only.md`, git-committed at `35788d9` before any score-only ASR existed, and the runner refuses to start unless its own `PREREG_COMMIT` matches that hash. *Primary:* Δ = mean ASR($\kappa{=}2$) − mean ASR($\kappa{=}0$) against the suite's existing ± 0.15 practical-equivalence margin, on `krum` / `committed_scaling` / CIFAR-10 / `cifar_cnn`, the same four rungs and seeds 42–46. $\Delta > +0.15$ would have said that **the published flat result was partly an artifact of the magnitude channel** and that the paper's central claim narrows to "with benign magnitudes free to move": that branch is why the arm was worth three hours of compute. The $\kappa{=}0$ rung is imported rather than re-run, because at the identity the transform returns the stack unwrapped and score-only Krum *is* Krum alone; that is an exact claim, not an approximation, and it was verified bit-identically at one seed before the arm started.

The negative survives with the magnitude channel closed. $\Delta = \mathbf{-0.023}$ (mean ASR $0.062 \to 0.039$) against -0.026 for the uncontrolled arm on the identical cell: **closing the magnitude channel moves the effect by** 0.003. The refuting threshold 0.212 is nowhere approached, and mean clean accuracy clears the 0.35 floor at every rung ($0.565/0.629/0.624/0.630$), so no rung is a collapsed model. **The central claim is therefore instrumented against two independent confounds** (attenuation, closed by Mode S, and magnitude, closed here) and the sharpened statement is that a 0.986 displacement of the emitted aggregate together with decision changes in 0.733 of rounds leaves suppression inside the pre-registered practical-equivalence margin. That containment is one-sided here as well: the identity-rung mean ASR is 0.062, below the margin, so its lower leg is satisfied by arithmetic and this arm's 95% bound on a rise is $+0.061$ (the supplement §S18).

Three disclosures, from the freeze rather than from hindsight. (i) The secondary Jonckheere–Terpstra test was specified to detect a monotone *rise* (the outcome that would have refuted flatness with attenuation already closed) and the observed trend is a *fall* ($J{=}58.5$, $z{=}-1.110$, $p_{\uparrow}{=}0.867$, permutation $p_{\uparrow}{=}0.874$). We report

it and do **not** score it as support for flatness; the pre-registration states in advance that at $n{=}5$ a null from this test is not evidence of flatness. (ii) The two branches had unequal headroom. With the identity rung at 0.062, refuting us required mean ASR($\kappa{=}2$) > 0.212 against a ceiling of 1 and was fully reachable, whereas the indeterminate branch ($\Delta < -0.15$) was *arithmetically unreachable*: ASR cannot fall below 0, so the largest possible fall on this cell is 0.062. (iii) **Selectors only.** A weighted averager and a coordinate-wise order statistic emit no single selected client, so "the selected update" is undefined for them and *this* control does not cover them, which means *this* control covers no `coord_median` arm and none of reputation's $+0.178$ ($n{=}5$). **What a coordinate-wise instrument can hold fixed instead is each coordinate's own contributing value**, and §D.10 builds one under its own freeze and runs it on the `coord_median` cell carrying the headline reversal, which is the $+0.098$ arm's own cell, where it returns the opposite finding to this arm's; `coord_median` under committed *scaling* ($+0.196$, $n{=}5$) is a different attack and stays uncontrolled, and the two instruments are one idea at two aggregator forms whose results transfer in neither direction. Score-only Krum is also an instrument and not a defense: it aggregates an update its own scoring stage never saw.

The corollary the pre-registration staked on `cos_krum`, and exactly what it settles. Section 5 of the same document asserted, before the run, that score-only `cos_krum` must be *bit-identical* across rungs: its statistic is exactly invariant under positive rescaling, so the selection cannot move, and with the emitted update untransformed the whole 50-round trajectory must repeat. It does (accuracy 0.528400 and ASR 0.003111 at all four rungs, $|\Delta| = 0.00\text{e}{+00}$ throughout) and the $\kappa{=}0$ value reproduces the published cell at that seed exactly, so the two pipelines agree at 0.000000 rather than approximately. **What the check does not do is confirm the arm's published fall.** It runs one seed, as frozen, and that seed is unrepresentative of the arm: under the uncontrolled dose its own ladder *rises* $+0.224$ where the eight-seed mean *falls* -0.425. Closing the magnitude channel removes all 0.224 of that seed's movement, which corroborates the *channel* (for `cos_krum` the emitted displacement is entirely rescaling, so no other route is open) but attributing the arm's mean fall to magnitude rests on that decomposition rather than on this check, and settling it empirically would need the arm's other seven seeds. Expanding a pre-registered seed count after seeing the data is precisely what pre-registration exists to prevent, so we report the invariance as confirmed and the arm-level attribution as **corroborated mechanically, not established at the arm mean**.

The same control at $n{=}20$: the negative survives, and the effect moves further inside the margin. The control above is $n{=}5$, the seed count its parent freeze fixed for *comparability* with the uncontrolled arm; that arm has since reached $n{=}20$ on this cell, so the control was extended to match rather than left as the lower-powered leg of the comparison. Decision rule: `experiments/pre_registration_score_only_kappa2_n20.md`, git-committed alone at `15d02e4` before the runner was written and naming `35788d9` as its parent, with the runner refusing to start unless that hash matches the log and the file is clean in `git status --porcelain`. Fifteen new runs at $\kappa{=}2$ on seeds 47–61; the $\kappa{=}0$ leg is imported at those seeds from `results/dose_seed_topup/` under the parent's own exact identity, that at the identity rung the transform returns the stack unwrapped and score-only Krum *is* Krum alone. Artifact `results/score_only_kappa2_topup/summary.json`, which holds the fifteen new runs alone; the $n{=}20$ contrast is assembled from it and the two imported legs by `analyze_score_only_kappa2_topup.py`, which asserts both legs are on exactly seeds 42–61 before it prints anything.

$\Delta = \mathbf{-0.0088}$ **at** $n{=}20$**, paired** 95% **CI** $[-0.0335, +0.0160]$, sd 0.0529, $t_{19}{=}2.093$, mean ASR 0.0439 at $\kappa{=}0$ against 0.0352 at $\kappa{=}2$, both legs on exactly seeds 42–61, per seed in the supplement's Table S3. Mean clean accuracy is 0.5618 and 0.6265, clearing the 0.35 floor by at least $+0.2118$, so no rung is void. The frozen label is *flat* and the frozen literal is **the negative survives with the magnitude channel closed at** $n{=}20$: substantial statistic and decision disturbance does not produce a corresponding change in suppression even when the aggregated update's magnitude and direction are held exactly fixed, at the same seed count as the headline reversal. **The two relations the freeze requires be stated together differ, and neither is reported as the other**: the interval lies entirely inside ± 0.15, and it does *not* exclude zero. This is a margin arm, so containment against ± 0.15 is what it scores and no nonzero effect is claimed from it. The realized half-width 0.0248 is narrower than the 0.0319 projected at freeze time, so the contrast is powered for its margin in the sense the freeze's `can_certify` clause asserts. **The published** $n{=}5$ **verdict is reported unchanged beside it**, -0.0232 at sd 0.0682, CI $[-0.1078, +0.0614]$, recomputed from the artifacts bit-equal to the record; the fifteen new seeds alone read -0.0039, CI $[-0.0309, +0.0231]$.

Four clauses travel with this number, and the first is a departure from the parent freeze that we disclose rather than absorb. *(i)* `35788d9` fixed *one cell, one dataset, one architecture, one attack, one defense,* $n{=}5$, and this suite departs from that literal text in the seed count. The $n{=}5$ was a comparability commitment, the control having to share seeds with the uncontrolled arm, and the uncontrolled arm is now at $n{=}20$ on these seeds, so extending the control *restores* the seed match rather than breaking it; the departure from the literal text is nonetheless real, the $n{=}20$ result is a **disclosed post hoc extension and never prospective**, and the $n{=}5$ verdict above is reported unchanged rather than replaced. *(ii)* **Only the frozen primary contrast is extended.** The score-only ladder across κ and its Jonckheere–Terpstra trend test stay at $n{=}5$; after this top-up the honest shape of this arm is $\kappa{=}0$ and $\kappa{=}2$ at $n{=}20$ with the interior rungs at $n{=}5$, which is why Table 2's row 6 prints a per-leg n and why no display prints that ladder without one. *(iii)* **Closing the magnitude channel does not**

remove the oracle. Mode S still reads adversary identity to pin $c_{\text{adv}}=1$, at $n=20$ exactly as at $n=5$, so this arm licenses no statement about oracle-free behaviour in either direction. *(iv)* **Magnitude is closed; the trajectory is not**, and **this arm covers selectors only**: which client is selected still changes across rungs so the trajectory still diverges, and `generic_compose` raises for non-selectors, so no `coord_median` arm is magnitude-controlled by *this* suite. The cell carrying the paper's headline reversal is controlled by a separate coordinate-wise instrument with its own freeze (§D.10), which returns the opposite finding; nothing here transfers to it and nothing there transfers back.

D.2.1 The 2×2 channel factorial, and the decomposition it refuses to license

Mode S opens two channels into a selector: the *statistic* channel (the dose moves Krum's pairwise distances, so *which* client is selected changes) and the *magnitude* channel (the selected client enters scaled by its own coefficient). Score-only closes the second. The fourth cell closes the first instead: **emit-only Krum** scores the *untransformed* stack, pinning the decision to the identity rung's, and aggregates the *transformed* selected update. The four cells are then a factorial, and if the two channels acted independently the three measured effects would satisfy $\Delta_{\text{full}} = \Delta_{\text{SO}} + \Delta_{\text{EO}}$. **They do not.** Rules were frozen in `experiments/pre_registration_emit_only.md` at `b995f1b` before any emit-only ASR existed, with the runner refusing to start unless its own `PREREG_COMMIT` matches that hash; the $\kappa=0$ rung is imported rather than re-run, because at the identity the transform returns the stack unwrapped and emit-only *is* Krum alone, an exact claim verified bit-identically before the arm started. All three harness checks the freeze required pass at $0.00\text{e}{+00}$ across five published arms and the score-only control, so the refactor that added this cell perturbed no published computation.

Cell	statistic channel	magnitude channel	mean ASR	ΔASR
$\kappa=0$ identity (imported)	closed	closed	0.062	—
Score-only	**open**	closed	0.032	−0.030
Emit-only	closed	**open**	0.107	+0.046
Full Mode S	**open**	**open**	0.023	−0.039
additive prediction $\Delta_{\text{SO}} + \Delta_{\text{EO}}$				+0.016
residual against measured Δ_{full} (frozen margin 0.05)				**−0.054**

The table above is the published $n=5$ design, retained as the pre-registration requires; the verdict it produced has since been withdrawn by our own top-up. Seeds 47–61 were added under rules frozen at `684b31e`, and at $n=20$ $\Delta_{\text{EO}} = +0.010$ (95% CI $[-0.028, +0.047]$), so the magnitude channel is *established* inert rather than merely point-inside, while the residual is -0.019 ($[-0.061, +0.023]$), **inside** the additivity margin. **So the $n=5$ reading that the channels interact does not survive**, and what remains is *consistent with* separability and not an establishment of it, the interval reaching -0.061 and so not contained in ± 0.05. Its half-width, 0.042, in fact came in *narrower* than the 0.057 that top-up's own pre-registration projected, so what blocks containment here is the off-centre point estimate rather than the width the projection worried about. Read the $n=5$ paragraphs below as the record of what was published and why it was fragile, not as the current verdict.

The primary rung is forced, and the additive model gets the sign wrong. At $\kappa=2$ mean clean accuracy is 0.333, below the 0.35 floor, so the primary contrast is reported at $\kappa=1$ and **that substitution is forced by the floor rather than chosen after seeing ASR**: Section 5 of the pre-registration named $\kappa=2$ as the rung at risk, and named it for this cell specifically, because emit-only aggregates the *rescaled* update and so degrades training in a way the other three cells do not. Seed-matched at $\kappa=1$, $\Delta_{\text{EO}} = +0.046$ sits inside the frozen inert margin of 0.05, so **the magnitude channel is inert on its own**, which is what we predicted. But $\Delta_{\text{SO}} + \Delta_{\text{EO}} = +0.016$ against a measured $\Delta_{\text{full}} = -0.039$: an additivity residual of -0.054, outside the frozen 0.05 margin, and **the additive model does not merely mis-size the joint effect, it reverses its sign**. So **no per-channel decomposition of the uncontrolled arm is licensed**, and we withdraw any reading of the re-selection/rescaling split as a decomposition of suppression. That split remains sound as the thing it says it is: the supplement's algebraic decomposition (§S9) decomposes a *displacement*, and says in its own terms that it makes no claim about suppression.

Three reasons this verdict is fragile, all of them ours to report. (i) It is knife-edge on both margins at once, $|\Delta_{\text{EO}}|$ falling *inside* the inert margin by 0.004 and the residual *outside* the additivity margin by 0.004. (ii) Both pre-registered secondaries are null: the emit-only per-seed sign count is 3 up against 2 down, and Jonckheere–Terpstra across the four rungs gives $z = -0.437$, $p_{\uparrow} = 0.669$. (iii) One seed carries the result. Seed 42's emit-only run collapses to clean accuracy 0.2613, below our floor *at that seed* though the rung mean of 0.487 clears it, and its Δ is $+0.272$ against -0.007, $+0.007$, $+0.013$, -0.058; excluding it would give $\Delta_{\text{EO}} = -0.011$ and residual -0.003, reading as separable. **We do not exclude it**: Section 5 fixes $n=5$ on seeds 42–46 and applies the floor to a rung's *mean* rather than per seed, so no per-seed exclusion is licensed, and the pre-registered top-up settled it without one. **At $n=20$ the knife edge is gone and the verdict it produced**

is withdrawn: $\Delta_{\mathrm{EO}} = +0.010$, 95% CI $[-0.028, +0.047]$, so the magnitude channel is *established* inert, and the residual is -0.019, $[-0.061, +0.023]$, inside the additivity margin, so the interaction reading does not survive. Seed 42 remains the only draw of 20 past either outlier threshold and is still excluded from nothing; realized sd falls $0.130 \rightarrow 0.080$. Separability is now *consistent with* and **still not established**, its interval reaching -0.061 and so not contained in ± 0.05 even though its half-width, 0.042, came in *narrower* than the 0.057 that top-up's pre-registration projected at this variance, and dropping the one seed that produces the unwelcome answer is precisely what pre-registration exists to prevent. Section 6.1 of the same document declared, before the run, that at n=5 only a large $|\Delta_{\mathrm{EO}}|$ would be decisive, and $|\Delta_{\mathrm{EO}}|$ is not large. **We therefore report this arm as not decisive, and act on it only in the direction that withdraws a claim**: low power is a reason to stop asserting a decomposition, never a reason to assert one. Every number above is emitted by `experiments/analyze_emit_only_factorial.py`, which reads only frozen artifacts and asserts all four primary quantities reproduce from the per-seed cells to $< 10^{-9}$ before printing. **Selectors only**: the factorial is defined for `krum` and `cos_krum`, the harness raising for any other d_2, so it is one selector family, one attack, one dataset, one architecture.

What this settles, and what it does not. The intervention delivers the within-defense contrast at a fixed adversarial coefficient share that the two cross-defense suites could not: within a single defense, with attenuation excluded by construction *and*, in the score-only arm, the magnitude of the aggregated update held exactly fixed as well, **substantial disturbance of the statistic d_2 reads need not produce a corresponding change in suppression once the adversarial coefficient share is held fixed** (on the adjudicating Krum arm under Mode S and, by pre-registered replication, on a downstream defense in a different invariance class, on a second dataset and a differently configured CNN, and on a second architecture family with architecture varying alone) while *admission* of adversarial input is the channel the evidence points to. The statistic-level reading the framework was first written in is refuted; the admission-level reading is confirmed as a *direction* in these interventions rather than established in general. It does *not* rescue the framework's positive direction wholesale: the pooled two-channel model over all 26 cells is **not** confirmed (R^2=0.239), and the replication refutes its cross-arm ordering outright, so admission is a better single predictor than the statistic without being a quantitative one. It addresses neither recall nor discovery (FG→RFA remains a false negative), and nothing in §D.1 is rescored: its pooled refutation, its two refuted arms and its indeterminate arm stand exactly as reported.

Mode M: coordinate masking, a second transformation class. Every other arm in this paper transforms client updates by a positive per-client scalar, the family Prop. 11 classifies and Thm. 9 bounds. **Mode M** is outside it. Benign client i has a random m fraction of the coordinates of each parameter tensor zeroed, drawn from an independent Bernoulli$(1-m)$ seeded by `default_rng([seed, round, slot])` so it is reproducible and uncorrelated with adversary status, and the surviving coordinates are then **renormalized to the update's original L_2 norm**. Renormalization is load-bearing and was required in the pre-registration *before any ASR existed*: an unnormalized mask shrinks benign norms and re-opens the magnitude channel that §D.2's control exists to close. The unnormalized variant was considered and rejected for that reason, and the rejection is recorded rather than omitted. **Adversarial updates are returned unmodified, exactly**: the maximum deviation of every adversarial coefficient from 1 is 0.00 with object identity preserved, a strictly stronger pinning than Mode S's 3.5×10^{-7} float32 read-back. Rungs are $m \in \{0, 0.2, 0.5, 0.8\}$ and m=0 returns the update list unwrapped, so the identity rung is bit-identical to Krum standalone by construction; we check rather than assert this, the identity rung agreeing to 10^{-9} across all three directories that carry those seeds.

The channels were measured before any ASR existed, as the EMNIST-byclass arm's were. Krum's decision changes in $0.000/0.133/0.400/0.467$ of rounds across the ladder while the adversarial mass it admits is 0.000 at every rung, at an aggregate displacement of 1.095 against Mode S's 0.892 at a *larger* decision change. So the arm was established prospectively as a generalization test, and the pre-registration (`684b31e`) then split four ways, including two that would have refuted us: a monotone rise in m would confirm the statistic reading, and masking that made Krum admit the adversary would void the equivalence claim at that rung and confirm the *admission* reading. **Neither fired.** At n=10 (seeds 42–51) the paired contrast m=0 → m=0.8 is $\Delta = -0.016$, 95% CI $[-0.051, +0.019]$, contained in the frozen ± 0.15 margin at a half-width 23.1% of it (TOST p=5.3×10^{-6}), every rung mean clearing the 0.5 ceiling and the 0.35 accuracy floor, and Jonckheere–Terpstra finding no increasing trend ($p_{\uparrow}$=0.895). The mask analogue of the score-only control gives -0.018, $[-0.047, +0.010]$, agreeing with the full ladder to 0.002: since adversarial updates are unmasked in *both* conditions the two ladders differ only when a benign client is selected, so their agreement localizes the absent effect to *which client Krum picks* rather than to the content it emits. Containment is one-sided on this arm too: the identity-rung mean ASR is 0.053, so the margin's lower leg is satisfied by arithmetic and the leg that tests anything bounds a rise at $+0.019$ (the supplement §S18).

What this arm does and does not license. Only the flat-with-admission-unchanged outcome generalizes the negative, and that is the outcome observed; had admission moved, the arm would have been reported as a confirmation of the admission reading instead. **One mask is one transformation, not the class of non-rescalings**: rotations, projections, quantization and sign compression are untested, so the supportable claim is that the negative replicates on a second, structurally different transformation class, never that it holds for arbitrary upstream processing. Mode M has nothing to impose in $24/120$ measured rung-rounds, identical to

Table A9: **The same aggregators under two transformation classes, every prediction frozen before its measurement.** Prop. 11 classifies aggregators under *positive per-client rescaling* only; coordinate masking is outside that family, so Thm. 9 and Prop. 11 make **no** prediction there, which is what makes the second class a test rather than a second demonstration. Rescaling predictions were frozen at `5130cec`, masking predictions at `684b31e`, each before the corresponding run existed; `experiments/build_prediction_matrix.py` refuses to emit if any observed cell disagrees with its frozen document. **`cos_krum` is the adjudicating row**: exactly invariant under rescaling, its selection bit-identical at every rung, it loses the guarantee under masking on the same attack and the same seeds, which is what it means for the classification to be a statement about a transformation *class* and not about an aggregator. **Every cell in this table is a decision-channel measurement, not an ASR**: under masking, ASR is measured for `krum` alone (§D.2), so the other three masking rows are channel-only and this table carries no ASR column to report them in.

		Δ dec., rescaling		Δ dec., masking	
aggregator	Prop. 11 class	predicted	obs.	predicted	obs.
`cos_krum`	(a) exactly invariant	unchanged	0.000	lapses	**0.467**
`krum`	(c) not invariant	changes	0.733	changes	0.467
`coord_median`	(b) conditionally inv.	partial	0.482	partial	0.700
`reputation`	(c) not invariant	changes	1.000	changes	0.533

Mode S's figure, and like Mode S it is an *instrument* and not a defense, since it reads adversary identity. The `cos_krum` mask ladder was considered and dropped for compute before any ASR existed, which is why the masking column carries one outcome and three channel-only entries.

D.3 Seven cells, and a pre-registered mechanism that did not survive them

Is the sign reversal a property of the design, or of this aggregator and attack? Table A3 contrasts the two designs on the one arm that motivated the instrument. That contrast is one cell, and a single cell cannot say whether the sign reversal is a property of the evaluation design or of `coord_median` under a pixel backdoor. We therefore ran the contrast on seven cells: four that already existed in frozen artifacts and are **re-scored, never re-run**, and three that did not exist and were run for this purpose (60 runs for the first two of them, plus 8 under Amendment 2 and 40 for the third under Amendment 4 below, which is the run budget that amendment froze).

Both designs are the same computation. The two ladders differ in exactly two tokens, the family name and whether `generic_compose` receives an `adv_mask`; data partition, participant draw, attack, local training and seeds are identical. `run_comparability_cells.py` refuses to start unless it first reproduces a published `results/dose_response/` cell bit-identically ($|\text{diff}| = 0.000\text{e}{+00}$), so the new cells are the same experiment as the four they are compared against, and `analyze_comparability.py` refuses to print any pooled number unless all four published contrasts reproduce.

What was frozen, and when. Before any new run, we pre-registered a mechanism (`experiments/pre_registration_comparability.md`, frozen at `c986ef4`; two amendments at `1026a96` and `f16083b`; the three hashes are kept because every other pre-registration in this paper carries its own, and dropping them for brevity would make this the one prereg citation a reader cannot check): the two designs disagree exactly where the downstream defense admits adversarial mass, gated on the admission-level change $\Delta\Lambda_a$ at the top rung. The four existing cells separate perfectly under that rule, which is precisely why **they are training data and cannot also be evidence for it**. The rule mechanically generated a prediction for each new cell, committed in advance: DISAGREE for `coord_median`/scaling ($\Delta\Lambda_a$=0.0086, disclosed at the time as the weakest margin of any cell), AGREE for `krum`/EMNIST-byclass ($\Delta\Lambda_a$=0.0000). The document names in advance what would refute it and commits to reporting refutation rather than fitting a threshold.

The out-of-sample result: one confirms, one refutes, one was never predicted (Table A10). `coord_median` under *model scaling* disagrees, as predicted, on the weakest margin in the set: at n=20 on both legs the outcome-gated ladder reads $+0.062$ $[+0.023, +0.102]$ where the instrument reads $+0.210$ $[+0.173, +0.247]$. This is the cell that matters for the headline, because it is the paper's sign-reversal aggregator under the *other* committed attack, and it says **the disagreement is not specific to the pixel backdoor**. It is a disagreement, not a second sign reversal, and we do not call it one. **The top-up that brought this row to n=20 changed the disagreement's character without changing its label**, and the disclosures below report that change together with the one outcome our own freeze failed to enumerate. The EMNIST-byclass cell then **refutes the rule**: $\Delta\Lambda_a$ is exactly 0.0000, the rule predicts the designs coincide, and they do not ($+0.157$ against -0.013, opposite signs). **Those opposite signs are not a second sign reversal, and we do not count them as one:** both intervals contain zero ($[-0.175, +0.489]$ and $[-0.034, +0.009]$), so neither estimate is

Table A10: **The confounded ladder against the Mode-S instrument, on seven cells.** Δ ASR from κ=0 to κ=2, paired 95% t intervals, each design at its own full n, **which the n column now prints as confounded/controlled** so the strongest row and the weakest are distinguishable on the page. **Rows 4 and 5 are the two rows at n=20 on both legs**, each following its own pre-registered endpoint top-up at seeds 47–61 (`experiments/pre_registration_reversal_seed_topup.md` and `experiments/pre_registration_coordmedian_scaling_n20.md`); at the frozen n=5 row 4 read -0.272 $[-0.422, -0.122]$ against $+0.098$ $[+0.018, +0.178]$ and row 5 read $+0.002$ $[-0.078, +0.081]$ against $+0.196$ $[+0.112, +0.280]$, **the same verdict for both by the same definition**, though row 5's outcome-gated interval no longer contains zero, and this section's top-up disclosures report what that changes. **AGREE** := intervals overlap and point estimates share a sign, or both lie within ± 0.05 of zero; **DISAGREE** := intervals disjoint or signs opposite; **SIGN REVERSAL** := signs differ *and* both intervals exclude zero. The four unmarked cells are the training data the rule was read off. **The rule predicted AGREE for the EMNIST cell and the designs disagree, so it is refuted**; per the frozen clause it is withdrawn, the four-cell pattern is demoted to a description, and the table is reported in full. **The CIFAR-100 cell carries no prediction and its $\Delta\Lambda_a$ entry is absent rather than zero**, because the rule that would have generated one was already dead when that cell was pre-registered (Amendment 4); it is scored on the two endpoint rungs the primary contrast reads, and because no trend secondary was pre-registered for this table, the trend across all four rungs is reported as a *post hoc* description below rather than as a test here. Every cell emitted by `analyze_comparability.py`, **except row 5's two n=20 readings, which `experiments/run_coordmedian_scaling_topup.py` emits from its own artifact**: that top-up leaves `results/comparability_six_cells.json` byte-untouched, so the frozen artifact still reads row 5 at n=5 and the two emitters are named separately rather than one standing for both.

cell	n	outcome-gated	Mode S	$\Delta\Lambda_a$	pred.	observed
krum/scaling	5/20	-0.008 $[-0.089, +0.074]$	-0.010 $[-0.032, +0.012]$	0.0000	AGREE	AGREE
cos_krum/pixel	5/8	-0.405 $[-0.796, -0.014]$	-0.425 $[-0.732, -0.118]$	0.0000	AGREE	AGREE
reputation/scaling	5/5	$+0.762$ $[+0.658, +0.866]$	$+0.178$ $[-0.112, +0.469]$	0.0196	DISAGREE	DISAGREE
coord_median/pixel	**20/20**	-0.273 $[-0.334, -0.213]$	$+0.125$ $[+0.095, +0.155]$	0.0332	DISAGREE	SIGN REVERSAL
coord_median/scaling*	20/20	$+0.062$ $[+0.023, +0.102]$	$+0.210$ $[+0.173, +0.247]$	0.0086	DISAGREE	DISAGREE
krum/scaling, EMNIST*	5/5	$+0.157$ $[-0.175, +0.489]$	-0.013 $[-0.034, +0.009]$	0.0000	AGREE	**DISAGREE**
coord_median/pixel, CIFAR-100*	5/5	-0.213 $[-0.284, -0.141]$	$+0.071$ $[+0.031, +0.110]$	—	—	**SIGN REVERSAL**

$*$ out-of-sample; the other four are the cells the rule was read off.

separated from no effect and what this cell establishes is a failed prediction of the rule, not a reversal of the measured one. Whether the *reversal* travels is a different question from whether the *rule* does, and the third out-of-sample cell is the one that addresses it.

The third out-of-sample cell moves the sign reversal to a third dataset, and it replicates (Amendment 4, committed at `9b8a395` before the runner was pointed at CIFAR-100). The objection this cell answers is one we sharpen rather than defend: before this row the table already spanned two attacks, two datasets and two CNN configurations, but **the sign reversal itself did not**. It existed on CIFAR-10 with `cifar_cnn` and nowhere else, because the cell that added a second dataset was `krum`/scaling, a *different* cell, and that cell is explicitly not counted as a reversal. Cell 7 therefore holds the aggregator, the attack, the rung grid and the seed set fixed and changes only the dataset, to CIFAR-100 under the same architecture: a 100-way task, which moves the backdoor's base rate and the clean-accuracy regime together. The outcome-gated ladder falls -0.213 $[-0.284, -0.141]$ where the instrument rises $+0.071$ $[+0.031, +0.110]$, opposite signs with both intervals excluding zero, so under the same frozen definition applied to every other row this is a **second sign reversal**. **It is a replication of the phenomenon, not an enlargement of the claim**: three datasets and two CNN configurations is the scope after it, the reversal is *dataset-conditional* on exactly that evidence, and nothing here licenses a statement about a dataset or an architecture outside the three measured.

Five disclosures about that cell, every one of them from the amendment rather than from hindsight. (i) **The admissibility gate was two-sided and frozen before the run**: the identity rung had to reach mean clean accuracy ≥ 0.35 *and* mean ASR inside $[0.15, 0.85]$, the upper bound because a saturated rung kills the *rise* leg exactly as EMNIST's 0.027 identity rung killed the *fall* leg in the cell above. It passed on both sides (0.393 accuracy, 0.703 ASR, means over seeds 42–46), and the amendment named in advance the finding it would have reported instead: *no third dataset within this compute budget admits this cell*. (ii) **No prediction was made, and the $\Delta\Lambda_a$ entry is absent rather than** 0.0000. The rule was refuted by the EMNIST cell and withdrawn before this cell existed, so it generates nothing here, and **a dead mechanism is not revived by a cell that happens to agree with what it would have said**. (iii) **The cell is scored on the two endpoint rungs**, which is exactly what the frozen primary contrast reads, so the row is complete as printed. All four rungs of the

frozen grid *are* run, on both designs, at seeds 42–46; **an earlier version of this disclosure said the interior rungs were not run and that the cell therefore carried no trend statistic, which was true when written and became false when the runner finished**, and we correct it here rather than silently. What the completed interior buys is a description and not a test: **the comparability pre-registration registers no trend secondary**, so the Jonckheere–Terpstra statistics that follow are *post hoc*, they are emitted with `pre_registered: false`, and no verdict in Table A10 is a function of them. Both ladders start from the same identity rung and then move apart across every rung, not only at the top: the outcome-gated design falls monotonically (0.7034, 0.6804, 0.6508, 0.4905; J=32.0, z=−2.89, $p_{\downarrow}$=0.0019) while the instrument rises monotonically (0.7034, 0.7035, 0.7220, 0.7740; J=96.0, z=+1.41), so the reversal is a divergence of the whole ladder rather than an artifact of its endpoints. **The rising leg does not clear** 0.05 **as a rank trend** ($p_{\uparrow}$=0.079, permutation $p_{\uparrow}$=0.084) even though its endpoint interval $[+0.031, +0.110]$ excludes zero, and **we report that weaker reading beside the stronger one** rather than the half that helps: at n=5 per rung a rise of this size is not separable from chance by ranks alone. Both $p_{\downarrow}$ values here are normal-approximation figures, following the supplement's §S8, because the permutation leg is computed for the increasing direction only. The CIFAR-10 cell of the same aggregator and attack has the same shape on both legs (z=−2.96, $p_{\downarrow}$=1.5×10^{-3} falling against z=+1.48, $p_{\uparrow}$=0.069 rising), under the same caveat on the rising one; its falling leg is the arm-4 trend already reported there, and it agrees because `analyze_comparability.py` *imports* that suite's trend test rather than reimplementing one. **Those four figures are at** n=5**, and stay there, even though that cell's endpoint row is now** n=20: the top-up ran the two *endpoint* rungs only, so its four rungs hold $[20, 5, 5, 20]$ seeds, and Jonckheere–Terpstra's null distribution is a function of those group sizes. A mixed-n trend is therefore not the statistic this grid was described with. The scorer computes it anyway and prints it beside the quoted one so that its exclusion is visible rather than assumed (z=−5.54 falling, z=+3.07 rising); **we do not quote it**, and no verdict reads it. The quoted values are the trend over the five seeds all four rungs share, which is the pre-top-up block, so they are the same numbers the earlier draft printed and not a re-test on new data. (iv) **CIFAR-100 was chosen on wall time and we record that as the limitation it is**: at 660 s per run it was the third dataset that fit an overnight budget, where EMNIST-byclass (29.5 h per cell, and an identity rung that fails the gate a priori) and ResNet18 (65 h) did not. **Those are *comparability-cell* costs, two designs × four rungs × five seeds, and they do not price the single-design endpoint arm of §D.2**, which is six runs and did fit one night; the figures are not in conflict because they cost different objects. This is the third dataset that fit, not the third that was most informative. (v) **The two designs coincide run for run at the identity rung of this cell**, bit-identically, as they also do for `coord_median`/scaling and do *not* for the EMNIST cell, so it is a property of the cell and not of the design. Both Δ estimates therefore share their κ=0 leg and the difference *between* the two designs is carried entirely by the κ=2 rung, which is where the attenuation channel is open. We say so because two identical columns in an artifact look like duplicated data and are not.

Row 5's top-up, and the branch of our own freeze that fired. `coord_median`/scaling was the last row still at the frozen n=5 on both legs, and it is the cell where both margin legs are genuine tests rather than one being satisfied by arithmetic, so it is where an interval is worth narrowing. `experiments/pre_registration_coordmedian_scaling_n20.md` was committed alone at `74044f1` with the output path verified not to exist, so `git log` orders the freeze before the data rather than our word doing it; 45 runs followed at seeds 47–61, the two endpoint rungs only, into `results/coordmedian_scaling_topup/summary.json`. It is 45 and not 60 because at κ=0 the two designs are the same computation, and **that is established on *this* attack rather than inherited from the pixel cell**: four harness runs at seed 42 reproduced both published κ=0 rows and both published κ=2 rows at $|\Delta| = 0.000\text{e}{+00}$ in ASR *and* clean accuracy, and their verdict dict is persisted into the artifact instead of left in a terminal. The freeze named two refuting branches. **The one that fired is the second**: the outcome-gated interval, which contained zero at n=5, excludes it at n=20. The first did not fire. Had the paired difference between the designs covered zero, this cell would have gone DISAGREE to AGREE and the paper's count of five disagreeing cells would have fallen to four, in §1 and in Figure 1. **So the label survives and the reason for it does not**: at n=5 one design detected nothing where the other detected an effect, and at n=20 both detect one and the disagreement is a disagreement about *magnitude*, +0.062 against +0.210 with disjoint intervals.

One outcome fell outside both branches we named, and we report it as a defect in the freeze rather than assign it to the nearer branch. The paired difference between the designs is +0.147 $[+0.103, +0.191]$ at n=20, against +0.194 $[+0.108, +0.281]$ at n=5. Its *sign* verdict is the one the freeze anticipated, an interval excluding zero, but its *margin* verdict moved from outside the frozen 0.15 to inside it, and our branches were written over the sign alone. **The two designs are therefore separated from each other and separated by less than the margin this paper calls practically meaningful, at the same time**, and neither half may be reported as the other. The same tension sits on the outcome-gated leg, which is inside the margin at $|+0.062|$ while excluding zero: **that arm stays inside the frozen margin and can no longer be phrased as no effect**, which is a stronger constraint on our own language than the n=5 row imposed. Neither half is an *equivalence reading* in the sense §S13 reserves that term for, which is a statement about the *controlled* estimate, and this arm's controlled interval lies outside the margin. **The property the cell was chosen for survived its own top-up**: the identity rung's mean ASR is 0.557 at n=20, above the frozen 0.15, so the largest available fall still exceeds the margin and both legs remain tests. That figure is an n=20 level and is *not* the 0.519 this paper

prints for `pure_coord_median` and for CoordMedian's standalone scaling baseline, which are n=5 readings of the same quantity on seeds 42–46 and which this top-up neither supersedes nor enters. **Endpoint-only has a declared price**: the four rungs of this ladder now hold $[20, 5, 5, 20]$ seeds, so no trend statistic is computed on the new seeds and none is quoted, exactly as the pixel cell's is not. The 0.35 accuracy floor was live and no rung is void (lowest rung mean 0.706, lowest single seed 0.633). $\Delta\Lambda_a$=0.0086 is imported from the frozen channel measurement and was not re-measured here, this arm emits no admission or channel quantity, and its controlled leg is masked, so **it bears on the oracle objection of §G not at all**.

Per-seed values for this arm, which the Reproducibility statement promises for every arm. Identity ASR, then the outcome-gated κ=2 and then the Mode S κ=2, on the 20 paired seeds; the first five are imported from `results/comparability_cells/summary.json` and the other fifteen were computed here, and each row of the artifact's `per_seed` block records which of the two it is. seed 42, 0.549 → 0.486 and 0.722; 43, 0.341 → 0.432 and 0.534; 44, 0.460 → 0.491 and 0.772; 45, 0.768 → 0.711 and 0.908; 46, 0.476 → 0.483 and 0.637; 47, 0.445 → 0.666 and 0.617; 48, 0.773 → 0.846 and 0.866; 49, 0.541 → 0.693 and 0.819; 50, 0.530 → 0.688 and 0.788; 51, 0.634 → 0.710 and 0.835; 52, 0.354 → 0.515 and 0.733; 53, 0.785 → 0.750 and 0.825; 54, 0.615 → 0.600 and 0.905; 55, 0.397 → 0.550 and 0.620; 56, 0.522 → 0.643 and 0.710; 57, 0.493 → 0.620 and 0.678; 58, 0.800 → 0.752 and 0.938; 59, 0.483 → 0.539 and 0.745; 60, 0.474 → 0.478 and 0.759; 61, 0.709 → 0.743 and 0.930. **Per seed the two designs separate more cleanly than their means alone show**: the controlled difference is positive on all 20 seeds, its smallest being +0.040 on seed 53, while the outcome-gated difference is negative on 5 of the 20, and the paired difference between the designs is positive on 19 with seed 47 the one exception at −0.049. Those three counts are *descriptive and post hoc*: the freeze scored means and paired intervals, so no sign test and no per-seed count statistic is computed on this arm and none is quoted as a verdict.

We report the refutation as a refutation. $\Delta\Lambda_a$ does not explain when the two designs disagree. The four-cell separation that suggested it stands as a description of four cells and nothing more, and no threshold was fitted after the fact to rescue it. **The observation survives the mechanism's death**: across seven cells the two designs disagree on five, on two different attacks, three datasets and two CNN configurations, and two of those disagreements are sign reversals. What we do not have is an account of *when* to expect it, which is the honest state of this result and a direct limit on how far the protocol of §4 can be turned into a rule of thumb.

The short form first, because it is all the argument needs: we pre-registered a mechanism before the runs that tested it, one out-of-sample cell refuted it, and it is therefore withdrawn. **One process failure is disclosed below that, because it nearly reversed this section** and because a reader entitled to the first sentence is entitled to how nearly it went the other way. The EMNIST cell was first scored on non-identical seed sets (n=5 against a frozen n=3), the disagreement carried entirely by two seeds with no counterpart in the other design. We amended the pre-registration to complete the controlled ladder to the frozen seed set rather than resolve the cell on the three shared seeds, which give AGREE: choosing them would have chosen the answer. The completed comparison is the fully paired n=5 in the table and it still disagrees. That same amendment asserted that cross-design pairing was the convention the freeze had used. **It was not**, and what caught the error was not our judgement but the analyzer's own gate: under cross-design pairing the four *published* cells stop reproducing (`krum`/scaling's controlled ladder moves −0.010 at n=20 to −0.026 at n=5), and the script refused to print. A third amendment withdraws the claim, and the full sequence is left in the pre-registration.

D.4 Two further regimes: an adversary that adapts to the downstream defense, and N=100 at 20% participation

Is the design disagreement an artifact of a non-adaptive attacker, or of ten clients at full participation? Neither question is answered anywhere else on disk, and each is answered by rerunning *both* ladders on the cell the instrument was built for, `coord_median` under the committed pixel backdoor. The adaptive machinery this paper already has (§E) was pointed at FoolsGold→RFA and scored whether *the screen's* verdict survives, so no artifact held a criterion-aware adversary and `coord_median` together; and the existing N=100 material stores game equilibria, which is why the supplement §S20.1 labels it residual. **The pre-registration was frozen before its output directory existed** (`experiments/pre_registration_regime_dissociation.md`, at `b317af0`, Amendment 1 at `f4c3fec`), and `run_regime_dissociation.py` refuses to start unless `PREREG_COMMIT` matches the amended hash. **Each regime changes one thing**: A changes the adversary, B changes N and K. The cell, the attack, the κ grid, f=0.2, α=0.5, the round count and `cifar_cnn` are held at the frozen values in both.

The pre-registered estimand is the disagreement itself and nothing else. $D = \Delta(\text{confounded}) - \Delta(\text{instrument})$, paired by seed, at a two-sided 95% t-interval. Its κ=0 leg cancels exactly, because at κ=0 both ladders return the update list unwrapped and both rungs are `coord_median` alone, so $D_s = \mathrm{ASR}_{\mathrm{conf}}(\kappa{=}2, s) - \mathrm{ASR}_{\mathrm{inst}}(\kappa{=}2, s)$. **Only D is powered at n=3, and the freeze says so with arithmetic rather than with a hope**: at the n=3 half-width $4.303\,\mathrm{sd}/\sqrt{3} = 2.484\,\mathrm{sd}$, the frozen variances give 0.299 for Δ(confounded) against a mean of −0.272 and 0.161 for Δ(instrument) against +0.098, so a rule requiring both intervals to exclude zero is one this arm was arithmetically guaranteed to fail; D's frozen mean is

-0.408 at sd 0.109, hence a half-width of 0.272. Pre-registering the two Δs would have been pre-registering a null. Each Δ is therefore **reported descriptively, with its half-width beside it and no significance attached**, and **no equivalence claim is made anywhere in this arm**: the phrase *no evidence of a practically meaningful change* is defined in the supplement against the paired 95% interval lying inside the margin, $n{=}3$ cannot earn it, and `EQUIV_MARGIN` is deliberately not imported by this runner. Table A11 reports both regimes seed by seed.

Table A11: **The design disagreement reproduces in both new regimes; the *sign* reversal reproduces in one.** Per-seed ΔASR on the two endpoint rungs, paired within seed, $n{=}3$. D is the pre-registered primary and its interval is the only inferential quantity here; **both Δ columns are descriptive**, printed with the frozen $n{=}3$ half-widths 0.299 and 0.161 in mind, and no verdict reads them. Under an adversary that solves for `coord_median` the instrument still moves *up* while the outcome-gated ladder moves down, on all three seeds. At $N{=}100$ with $K{=}20$ both designs move down and only the magnitude disagrees, so **regime B reproduces the disagreement and not the reversal**, which scopes the reversal to the participation regime rather than refuting it. The frozen row is Table A3's published cell, at the n each of its quantities was frozen at, and it is a reference point rather than a third arm. Emitted by `experiments/run_regime_dissociation.py --analyze-only` over `results/regime_dissociation/summary.json`.

regime	seed	Δ confounded	Δ instrument	D
A: adaptive adversary	42	-0.265	$+0.122$	-0.387
$N{=}10$, $K{=}5$	43	-0.214	$+0.009$	-0.223
	44	-0.249	$+0.113$	-0.362
	mean	-0.243	$+0.081$	$\mathbf{-0.324}$ $[-0.543, -0.105]$
B: $N{=}100$, $K{=}20$	42	-0.276	-0.183	-0.093
	43	-0.333	-0.212	-0.120
	44	-0.367	-0.201	-0.165
	mean	-0.325	-0.199	$\mathbf{-0.126}$ $[-0.217, -0.035]$
frozen cell, as published	mean	-0.273	$+0.125$	-0.408

Three gates, all of them run before the arm was scored, and two of them returned exact zeros rather than small ones. (1) *The adaptive adversary must beat the plain backdoor it replaces*, or regime A tests nothing: a grid of $\varepsilon \in \{1, 2, 4\}$ crossed with `decorrelate` $\in \{\text{True}, \text{False}\}$ at $\kappa{=}0$, seed 42 only, argmax mean ASR, then frozen for every rung and seed. (2) *Every rung must clear the* 0.35 *clean-accuracy floor*, which both regimes do ($0.769/0.691/0.690$ and $0.474/0.426/0.441$). (3) *The shared* $\kappa{=}0$ *rung must be proved shared, not assumed*: the instrument's own $\kappa{=}0$ rung is run at one seed in each regime and differenced against the confounded one, giving $(0.000000, 0.000000)$ in accuracy and ASR at a tolerance of 10^{-6} in *both* regimes, the adaptive one included. **The two frozen runners were also proved unchanged by value before either regime ran**: the three optional keyword arguments this arm adds to `run_dose_response.py` and `run_targeted_dose.py` are exercised at their defaults against stored cells of `results/dose_response/` and `results/reversal_seed_topup/`, worst absolute deviation $0.00\text{e}{+00}$ over three runs plus a cross-ladder identity comparison. **An md5 of an artifact would not have been that proof**, which is why the check is a value comparison against per-seed rows.

The calibration grid is the reason regime A exists, and its outcome vindicates a decision taken before the data. `decorrelate` confines each adversary to a disjoint coordinate block, which is an anti-FoolsGold device; the freeze argued that against a coordinate-wise median a lone adversary inside its own block is plausibly *weaker*, and refused to import the existing suite's argmax on that ground. The grid says so unambiguously: with `decorrelate` on, ASR is 0.150, 0.192 and 0.199 across the three radii, **every one of them below the plain backdoor's** 0.395, while with it off they are 0.398, 0.430 and 0.406. Importing $\varepsilon{=}1$ with decorrelation because it won against FoolsGold→RFA would have installed an adversary weaker than no adaptation at all, failed gate (1), and licensed no sentence about adaptivity. **A hyperparameter's argmax does not travel across downstream defenses**, and the general lesson is the one we would want applied to us: an adaptive-attack baseline tuned against one defense is not an adaptive-attack baseline against another.

Six disclosures, five of them limits. (i) **The winning adversary beats the plain backdoor by** $+0.035$ **at one seed** (0.4299 against 0.3946), which is a margin, not a rout. What this arm establishes is that the disagreement *survives* an adversary adapting to d_2; it does **not** establish that a strong adaptive attack against `coord_median` was found, and no sentence anywhere in the paper says so. (ii) **The adversary does not anticipate** d_1, and that is a property of the object rather than a gap in the attack: Mode S reads adversary identity in order to pin the adversarial coefficient, so it is an instrument and not a deployable defense, and no real attacker can adapt to a stage that does not exist in deployment. The arm therefore bounds adaptation to the pipeline a deployed attacker can see. (iii) **Regime B's sign reversal does not reproduce, and we report that as the finding it is.** Both designs read the dose as protective at $N{=}100$, the instrument's Δ landing at -0.199 and negative on all three

seeds ($-0.183/-0.212/-0.201$), so what travels to 20% participation is the disagreement and not its sign. That interval happens to exclude zero, and **it is still not reported as inferential**, because the freeze forbids it at n=3 and a bound may not become a claim because it came out convenient. (iv) **Regime B confounds N with K**, deliberately, because the practitioner-facing regime is many clients with partial participation and separating the two would have been a third arm. Nothing here attributes the attenuation of the reversal to either factor alone. (v) **Two rungs, three seeds, one cell, one attack, one dataset, one architecture.** Neither regime is a four-rung ladder, no trend statistic is computed here and none may be quoted from it, and the arms add no dataset, attack, f or round count. (vi) **The base rung does not suppress the attack in either regime** (0.387 and 0.514), which the frozen data already showed, so no sentence from this arm may say that `coord_median` alone suppresses this backdoor.

One unplanned corroboration, reported because it is stronger than what the freeze expected. The pre-registration cautioned that its N=100 rung and the residual N=100 material "are not bit-comparable", and on that basis regime B ran its own κ=0 rung and never substituted the older one. It ran anyway, and the two agree exactly: freshly measured ASR $0.4261/0.5532/0.5631$ at accuracies $0.4773/0.4855/0.4598$ against the identical figures per seed in `results/cifar10_100clients_rich/seed_4{2,3,4}/payoff_results.json`. **So the caution was false and it was false in the safe direction**: it caused an extra measurement rather than a substituted one, the two runners are protocol-identical on this cell, and by the standard we apply to ourselves elsewhere, a computation reproduced across two runners establishes protocol identity and not corroboration of the effect. We leave the committed pre-registration as written rather than amend it a second time after seeing results, and record the correction here.

Amendment 1, and why it was needed before any run. The freeze required the runner to assert that the adversarial tensor *changed* at the first round, as a guard against a hook that is silently never called. That assertion is false by design for the committed pixel backdoor, whose payload lives in the poisoned dataset and whose `manipulate_update` is the inherited identity, so the runner would have aborted on round zero of every non-adaptive leg. Amendment 1 replaces it with an **attack-conditional** assertion: the update must change exactly when the attack overrides `manipulate_update` or the criterion-aware layer is active, and must be *unchanged* otherwise. **This is strictly stronger than what it replaced**, because the defect the original clause existed to catch, a hook that never fires, produces "nothing changed", which is also this attack's correct behaviour, and that exact collision once left two composition arms in this repository computing the same thing. All four failure modes were deliberately induced and each was caught. The amendment was committed **before** `results/regime_dissociation/` existed and before any run of this arm, and the pre-amendment text remains at `b317af0`.

D.5 One Mode-S intervention off the frozen Dirichlet concentration

Every Mode-S row elsewhere in this paper is at the frozen α=0.5. This arm moves α alone on the sign-reversal cell (`coord_median` under the committed pixel backdoor, endpoint rungs $\kappa \in \{0, 2\}$), holding the cell, the attack, the rung endpoints, N, K, f, the round count, `cifar_cnn` and the seed block 42–46 at their published values. It is pre-registered in `experiments/pre_registration_heterogeneity_modeS.md`, committed as its own commit at `e7b7c83` before either output path existed; both scripts refuse to start unless that document is committed at the recorded hash with a clean working tree. The α grid and the treatment of the frozen column are the paper's own: §S24.4(E) already sweeps compositions at exactly $\alpha \in \{0.1, 1.0, 10.0\}$ and carries α=0.5 as each pair's headline value rather than a rerun, so each row here is comparable to an existing row at the same concentration under the same convention. This arm uses n=5 where that sweep used 3, because small n is the objection it answers.

α	Δ_S	paired 95% CI	n	mean acc. κ=0/2	premise
0.1	$+0.218$	$[+0.064, +0.372]$	5	$0.49/0.47$	holds
1.0	$+0.137$	$[+0.046, +0.228]$	5	$0.79/0.75$	holds
10.0	$+0.165$	$[+0.127, +0.204]$	5	$0.81/0.77$	holds
0.5, frozen, not re-run	$+0.098$	$[+0.018, +0.178]$	5	—	holds

Two verdicts, frozen separately before any number existed, because a previous arm of ours cleared its pre-registered threshold in two regimes while reproducing the sign in only one. The *sign* verdict is that the within-defense rise reproduces at all three concentrations, each interval excluding zero at n=5. The *magnitude* verdict is that each interval overlaps the frozen α=0.5 interval, which is an interval-overlap statement and not the equivalence phrase the supplement defines: `EQUIV_MARGIN` is not imported by this arm.

Both Mode-S premises were measured at every concentration, not assumed. At each α the identity rung is the exact identity, with zero coefficient and zero channel-displacement violations over 18 rows tested with equality rather than a tolerance, and the adversarial coefficient is exactly 1.000000. Baseline adversarial admission into coordinate median is 0.261, 0.270 and 0.271, nonzero in 30 of 30 rounds at each concentration,

against the frozen α=0.5 level of 0.315: the adversary is admitted at every concentration, so no row here rests on an aggregator that excludes it outright.

A gate we corrected after seeing it fail, with the threshold unchanged. Our first implementation compared the adversarial coefficient share within each (seed, round) cell and held the worst spread over cells to the frozen 10^{-6}. All three concentrations failed that form, at 2.56e−6, 3.81e−6 and 2.88e−6, and were labelled confounded. So does the published α=0.5 leg, at 3.97e−6, which is larger than all three, so that form cannot discriminate concentration and the verdict it emitted asserted a mechanism the measurement does not support. The tolerance is calibrated in its own defining comment for the spread of per-rung *mean* shares, recorded there as 3.48e−7 on CIFAR-10, and that is the statistic both existing users of the constant compare; the gate now tests it with the same strict inequality. The tolerance is still 10^{-6} and was not widened. On the corrected statistic the published α=0.5 reads 9.22e−8 and the three concentrations read 1.09e−7, 1.12e−7 and 6.75e−7. The per-cell figures are retained per concentration as diagnostics, and the gate now measures the published configuration before judging any concentration and returns indeterminate rather than condemning a new one with an instrument that condemns the published one.

What this arm does not establish. Clean accuracy at α=0.1 is 0.49 at either rung against the 0.75–0.81 the other two span, so that leg clears the frozen 0.35 floor and is interpretable, but its absolute ASR is not comparable to the others. Three concentrations on one cell, one attack, one dataset and two rungs license no statement about heterogeneity in general, and no mean is taken across concentrations: each Δ_S carries its own n and its own interval, and no concentration's rung is differenced against another's. The frozen α=0.5 verdict is not amended; it is printed as frozen beside these rows. The attack is the committed pixel backdoor, non-adaptive by construction, so nothing here concerns adaptivity.

D.6 An oracle-free upstream defense on CIFAR-100: attempted, and not established

Can the Mode S contrast be run with a real deployed defense as the upstream transform, on a second dataset, at nonzero baseline adversarial admission? That one experiment would answer three separate objections at once: Mode S's rungs pin the adversary with an `adv_mask` that a deployed server does not have, every composed-pair experiment in this paper was on CIFAR-10 when this arm was pre-registered, and the flagship arm's admission baseline is a floor. `norm_clip` is the candidate, and the fit looks exact: its coefficient is $c_i = \min(1, \tau/\|u_i\|) > 0$, so it lies inside Theorem 9's positive per-client rescaling class, it needs no mask, its own τ is the dose, and T=id is recovered at τ above every client norm. We pre-registered it in two stages with stage 2 gated on stage 1's verdict (`experiments/pre_registration_normclip_cifar100.md`, committed at `9cbd009` before either runner existed, with one append-only amendment at `2eeff40`), ran stage 1, and **the premise failed**. Stage 2 never started, so **this arm reports no composed-pair ASR result on a second dataset**, and none may be inferred from what follows; the arm that answers that second objection is §S24.3, pre-registered separately and run later. The CIFAR-100 cell of §5 is a different object and is not withdrawn by this: its upstream transform is the Mode S instrument rather than a deployed defense, so it replicates the design's sign reversal on a second dataset without composing two defenses.

The premise was three conditions frozen before any ASR existed, and two of them passed. (i) *`coord_median`'s baseline adversarial argmedian fraction on CIFAR-100 must clear a pre-stated floor of* 0.05 *in at least half its adversary rounds*: measured 0.2885, nonzero in 15/15, against CIFAR-10's 0.2855. The admission *level* therefore replicates across the change of dataset and label space, which is a real if small positive result and the only one this arm yields. (ii) *At the identity rung every c_i is exactly* 1 *and every channel displacement is exactly* 0: exact at both epoch counts, tested with == rather than against a tolerance. (iii) *The adversarial share of coefficient mass must not move beyond the frozen* 10^{-6}: it moves 0.2533 → 0.2603 at τ^*, a spread of 6.9×10^{-3}, with 45 of 90 client-rounds clipped and the smallest coefficient 0.838. So $\Delta_c \neq 0$, the arm measures the attenuation channel rather than the statistic channel, and by the freeze's own demotion clause it is **not established on CIFAR-100**.

The freeze's guess at the mechanism was wrong in sign, and its verdict string still stands as the verdict. That literal reads *arm confounded on this dataset: the clip attenuates adversaries*. The clip attenuates *benign* clients. Over the 15 adversary rounds the mean coefficient on adversarial clients is 1.000000 (not one adversary was clipped in any round) against 0.962780 on benign clients, so $\Delta_c = +0.037$ and the adversarial share rises rather than falls. Under a pixel backdoor adversarial updates are not norm outliers; they sit just below the benign median, so a clip placed *at* that median is a benign-only attenuator. We print the pre-committed literal as the verdict, because that is what the artifact records and what was frozen, and the measured direction beside it, rather than letting the literal's wording stand as a finding.

The failure is structural rather than a badly chosen τ, and that is this arm's actual contribution. A clip's coefficient admits exactly two regimes. If τ sits above every client norm it does not bind, every c_i is exactly 1, and $\Delta_c = 0$ *by construction*, which is precisely our own $\Delta_c = 0.000$ row for this transform under this attack (Table S4, at $\rho = 1.00$ because no client exceeds τ=5), a zero that measures nothing, as that table's

discussion already says. If τ binds, Δ_c is a difference of two group means of $\min(1, \tau/\|u\|)$ and is zero only by coincidence, its sign set by which group's norms sit higher: -0.721 under model scaling, $+0.037$ here. **For this family neutrality and informativeness are mutually exclusive**, at any τ and on any dataset, because it is a fact about the functional form and not about our configuration. So the mask Mode S carries is not a convenience, and **no oracle-free member of Theorem 9's class serves as a clean probe of the statistic channel**. What we had stated about one table row generalizes to an impossibility for the whole family, and we had not drawn that consequence.

τ was fixed by a rule before the number existed, and the amendment is why the confound is visible at all. The rule is the median client update norm, so roughly half of clients clip; stage 1 measured $\tau^*{=}4.0825$. It was not re-chosen after the failure and no second τ is reported: searching for a τ that passes condition (iii) after seeing this one fail is what pre-registering a rule instead of a value was for. The amendment, appended before either runner existed, moved the rule's instantiation from the one-epoch pass that every existing admission measurement in this paper uses to the two-epoch pass the ASR ladders train. **At one epoch the same τ^* clips 0 of 90 client-rounds and condition (iii) passes with a spread of exactly 0**: the confound exists only at the epoch count the ladder would have run at, so an unamended freeze would have certified a stage 2 whose upstream defense was inert. Artifact `results/normclip_cifar100_admission.json`, 3 seeds $\times$ 6 rounds $\times$ 2 epoch counts $\times$ 2 rungs, no ASR anywhere in it.

D.7 Both channels without adversary identity, at $n{=}20$ off the admission floor: the pre-registered prediction is refuted, and the arm is not a clean statistic probe

Can the statistic-channel or the magnitude-channel measurement be reproduced without using adversary identity to build the intervention? That is the first of the three objections §D.6 set out to answer, and the one its failed premise left open. Nothing else on disk answers it either: every published measurement of those two channels is a Mode S, Mode A or Mode M rung, and `apply_d1_transform` *raises* for all three without an `adv_mask`, so every published measurement reads adversary identity by construction rather than by choice. This arm answers it at $n{=}20$ and off the admission floor, and **two things went against us**. The pre-registered prediction for the statistic channel is **refuted**, which we report as a contradiction of the reading this paper leads with and not as a scope condition on it, because that is what the freeze requires. And a measurement we made only *after* the verdict shows the arm does not hold the adversarial coefficient share fixed, so it is not the clean statistic-channel probe the freeze took it for. We report the frozen verdict as frozen and that measurement beside it, and we do not use the second to retract the first.

The design uses no new transformation class and no new methodology. It does not need Δ_c-neutrality, which §D.6 shows no *oracle-free* member of Theorem 9's positive per-client rescaling class can combine with informativeness, because it closes a channel *downstream* of the coefficient, for every client at once, using the two off-diagonal cells of the factorial this paper already pre-registered (§D.2.1): `score_only` scores the transformed stack and emits the selected client's *original* update, closing the magnitude channel (leg 1), and `emit_only` is the mirror, pinning the decision and leaving magnitude open (leg 2). Neither control is told who the adversary is. The upstream transform is `rfa`, a deployed defense this paper composes elsewhere, whose per-client coefficients are the final Weiszfeld weights of its geometric-median iteration and are computed from the update stack alone; its realized spread is $\rho \in [1.306, 2.194]$, so it is comparable in size to Mode S's $\rho{=}2.72$ rung rather than a weaker intervention. Three arms, 20 seeds each, 42–61, fixed before the first run with no top-up clause; 60 runs, 17.3 h. Pre-registered in `experiments/pre_registration_oracle_free_channels.md`, committed as its own commit, `4e7090b`, at a point when both output paths were verified not to exist, with both scripts refusing to start unless that document is committed at the recorded hash with a clean working tree. Of §D.6's three objections this arm answers two, adversary identity and the admission floor, and not the remaining one: it is on CIFAR-10, so it adds nothing on the dataset axis and does not disturb §D.6's statement that no composed-pair ASR on a second dataset follows from *that* arm. The dataset axis is taken up instead by §S24.3.

The host cell is off the admission floor, by measurement rather than by assertion. Krum under the committed pixel backdoor admits baseline adversarial mass 0.333, nonzero in 4 of 12 adversary rounds, against 0.000 in 0 of 12 for the same Krum under committed scaling. Admission has somewhere to fall here, so an unchanged admission in this cell is a measurement and not a ceiling artifact, and the floor the flagship arm is charged with is the attack's rather than the aggregator's. This arm recomputes that baseline itself rather than carrying it across from the flagship arm's admission artifact, and reproduces it exactly. The identity rung's mean ASR is 0.3232, against 0.062 on the cell the published score-only control ran (§D.2), where we disclose that a fall of more than 0.062 was *arithmetically unreachable*. **That difference in headroom is the whole point of moving cells**, and it is why this arm could return a verdict that one could not. It is also why §D.9 runs the masked Mode S dose on this same cell.

Oracle-freeness is enforced by the callee, not intended by the caller. The runner passes `adv_mask=None` whenever the upstream stage is named directly, so a dose family reached by this path raises instead of running.

That the transform actually ignores identity is asserted on live update stacks rather than read off the branch structure: over 5 rounds, `apply_d1_transform` with `adv_mask=None` is bit-identical to the same call with the true mask, tested with `torch.equal` on every tensor of every client, and it is not the identity, with first-round per-client norm ratios 1.301, 0.794, 0.943, 1.297 and 0.665. The same check confirms that the published Mode S transform raises without the mask, so both halves of the contrast are witnessed by one measurement rather than by two conventions.

The premise was measured before any ASR existed, on a three-way rule frozen before its numbers were read, and it passed. Across 5 seeds $\times$ 3 live rounds, with both rungs on the same raw updates, `rfa` moves Krum's *decision* in 8 of 15 rounds (0.533) and its *admitted adversarial mass* in 0 of 15, at an aggregate displacement of 0.868. The identity rung runs through the same code path with d_1=`fedavg`: over 15 rows neither quantity moves in a single one, every realized weight ratio reads 1.000000000000, and every adversarial coefficient is exactly 1 on every round where an adversary is sampled, so the measurement reproduces its own baseline exactly rather than to a tolerance. The rule's two constants, a decision floor of 0.10 and an admission-flatness bound of 0.05, are *imported* from the Mode M measurement rather than restated, and the rule was applied to the published Mode S/Krum arm before it was applied here: at decision change 0.733 and admission change 0.000 it types that arm the same way, so this is not an instrument that condemns only what is new. One diagnostic explains the choice of host: under the same `rfa`, `cos_krum`'s decision does not move at all (0.000), which is why that aggregator cannot carry this arm and Krum can. **This measurement is 15 rows at rounds 0–2 and not the 50-round ladder**, so the admission channel is asserted flat where it was measured and not throughout the runs behind the ASR figures below.

Both pre-registered predictions fail, in opposite directions. Leg 1 was frozen as $|\Delta\mathrm{ASR}|$ *inside* the 0.15 equivalence margin; it reads $+\mathbf{0.2969}$ (sd 0.4243, paired 95% CI $[+0.0983, +0.4955]$, n=20), outside the margin *and* excluding zero, which is the exact conjunction the pre-registration named in advance as contradicting this paper's headline reading. Mean ASR rises $0.3232 \to 0.6201$ at mean accuracy $0.5158 \to 0.5671$, so no accuracy collapse explains it, and no rung on either arm of this leg crosses the 0.35 floor. **An earlier version of this sentence added that no rung comes *within* 0.10 of the floor, and that was false of the identity arm**: three of its twenty rungs do, reading 0.3962, 0.4261 and 0.4474, while the statistic arm has none and its minimum is 0.4571. We correct it here rather than drop it, because §D.8 draws a contrast against it that is now a quantitative one. Leg 2 was frozen as nonzero with its interval excluding zero; it reads -0.0167 (sd 0.0874, CI $[-0.0588, +0.0254]$, n=19), inside the margin with the interval containing zero. The two verdicts are read separately per leg, as the freeze requires, and neither is reported as the other.

What the $+0.297$ is, stated as the distribution rather than as its mean. This cell is bimodal: per-seed ASR sits either above 0.5 or below 0.10, which is why a real effect carries an sd of 0.4243. The identity arm has 6 of 20 seeds in the high mode and 12 below 0.10; the statistic arm has 12 of 20 high and *none* below 0.10. Eight seeds move from the low mode to the high one and none move the other way, so the effect is directional seed by seed and not an artifact of averaging: score-only `rfa` doubles the number of seeds in which the backdoor takes hold and empties the near-zero mode. Four per-seed differences are negative, and the two large ones (-0.575, -0.505) are identity-high seeds landing mid-range at 0.350 and 0.474 rather than seeds the intervention protected.

Leg 2 is unresolved, which is not a zero, and one rung of it is void. The paired half-width is ± 0.042, so the arm is consistent with no effect and equally with an effect of a few points either way; it is not evidence of absence, and the equivalence phrase the supplement defines against the stricter of our two interval conventions is not applied to it. Seed 61's magnitude rung reads clean accuracy 0.3334, below the frozen 0.35 floor, so that rung is *void rather than negative* and the pair is dropped, which is why leg 2 reads n=19 against leg 1's n=20; both verdicts are unchanged if it is retained (-0.0159, CI $[-0.0557, +0.0240]$, n=20), so nothing turns on the exclusion. The runner's own accuracy gate compares each arm's *mean* against the floor and prints the per-seed minimum beside its verdict, which is why its emitted line reads `PASS` next to a minimum of 0.3334; the per-seed exclusion is applied on top of it and the per-run marker in the log is what flags the rung. This leg leaves the mode structure nearly intact (6 high, as in the identity arm, and 13 below 0.10 against its 12) while differing from the identity arm on all 20 seeds, so the channel is open and simply does not resolve. **Its failure was foreseeable from an artifact we already had** and we should have said so in the freeze: §D.2.1's own top-up reads $\Delta_{\mathrm{EO}} = +0.010$, 95% CI $[-0.028, +0.047]$ at n=20, and calls the magnitude channel *established inert*. Our leg-2 prediction was taken from Proposition 3's second clause rather than from that number, and the two intervals in fact overlap, so this leg is better read as consistent with a published inertness on a different attack and transform than as a surprise.

The measurement we made after the verdict: this arm does not hold the adversarial coefficient share fixed. `rfa` pins nothing, and pinning the adversarial coefficient is exactly what Mode S needs the mask for, so we checked post hoc what the freeze did not gate on. Over the same 12 adversary rounds, the adversarial share of coefficient mass moves in **12** of **12**, by up to 0.0859, against the frozen `SHARE_TOL` of 10^{-6} that Mode S holds to 3.5×10^{-7}; the implied Δ_c has mean -0.089 over a range $[-0.537, +0.147]$, changing sign across rounds (5 positive, 7 negative). So the attenuation channel that Figure 1's design closes by construction is **wide open here**, and `score_only` does not close it: it fixes what the selected client contributes, not what every

client's coefficient does to the scoring that selects it. The mean direction, adversaries attenuated relative to benign clients, is the direction that moves an adversarial update toward the benign cluster in Krum's distances, so it is a named candidate route to the $+0.297$ rather than a channel we can exclude. **Three things follow, and we state all three.** (i) *The frozen verdict is not amended.* Leg 1 is refuted, and the paragraph above reports it as a contradiction, because the freeze forbids converting a refutation into a scope condition after the fact and this measurement does not license doing so. (ii) *This arm is not a clean probe of the statistic channel alone*, so its refutation does not transfer to the published Mode S result, whose own settled statement is conditional on *once the adversarial coefficient share is held fixed*; the two interventions differ in two factors at once, the attack and the upstream transform, and this arm decomposes neither. (iii) *The defect is in our pre-registration and it is ours to own.* §D.6 demotes an arm on exactly this condition, our freeze recorded `share_tol` in its own rules block, the runner recorded `adv_coeff_share` on every round, and the gate was still not written. By the paper's own existing precedent this leg would have been typed confounded in the attenuation channel had the gate been included, and we did not include it.

What the arm therefore establishes, which is narrower than either prediction and is not good news for us. Removing adversary identity re-opened the channel the flagship design exists to close, because closing it is what the oracle is for. That is the empirical companion to §D.6's impossibility for the whole positive-rescaling family: there we showed no oracle-free member of that class is simultaneously Δ_c-neutral and informative, and here an oracle-free transform outside the neutrality requirement produces a large suppression change with Δ_c open. **The honest summary is that the oracle is load-bearing for our negative rather than a laboratory convenience**, and that we have not produced an oracle-free measurement of the statistic channel on any cell. Proposition 3 is an existence claim over exhibited counterexamples and is not disproved by a cell in which a statistic disturbance does cost suppression; what fails is the *generalization*, and a reader who wanted the negative to travel without an oracle should read this arm as evidence that it does not.

Both controls are asserted at the bit level against independently constructed aggregates. The standing failure mode in this codebase is a manipulation hook that is never called while the run passes silently, so neither control is taken on its docstring. `emit_only`'s returned aggregate equals FedAvg over the *raw*-stack selection's transformed update, and `score_only`'s equals FedAvg over the *transformed*-stack selection's raw update, both tested with `torch.equal` rather than against a tolerance, each against an aggregate the checker builds for itself from the selection that control is defined to use. So `emit_only`'s decision is pinned to the identity arm's by construction and not by approximation, and no seed's `emit_only` ASR equals the identity arm's, so the pinned decision did not make the arm inert. The two controls select different clients in 3 of the 5 rounds checked, so they are not the same computation.

Three harness checks ran before the ladder, and their verdicts are in the artifact rather than in a terminal. The one keyword this arm adds to shared code is proved bit-neutral by recomputing a frozen Mode S cell through the patched runner, at $\Delta = (0, 0)$ on both accuracy and ASR; the identity arm is proved to be Krum standalone *across suites* by reproducing the pilot's own `krum` under pixel row at a shared seed, also at $\Delta = (0, 0)$, which is the cross-suite assertion an existing dose-response runner already makes against that same pilot file; and the `rfa` branch is proved entered and non-inert. All three verdicts are persisted, because a check whose verdict is discarded leaves the claim it supports witnessed only by a terminal and costs hours to repeat.

What this arm does not establish. One aggregator, one attack, one dataset and one upstream defense license no statement of the form *the negative fails without an oracle* in general, any more than they would have licensed its converse; what they license is *in this cell, without adversary identity*, and that is the only form used above. The refutation is of a pre-registered directional prediction, in the sense this paper fixes for that word throughout, and not of any mathematical claim. The two controls modify the aggregator's internals, so this remains an instrument for causal identification and not a deployable defense: it answers the adversary-identity objection and leaves the deployability one open exactly where the rest of the paper leaves it, on the same ground already given for the published control, that score-only Krum aggregates an update its own scoring stage never saw. The attack is the committed pixel backdoor, non-adaptive by construction, so nothing here concerns adaptive robustness. And no row is pooled with any published row: the $n{=}3$ pilot value for this cell is a harness-check target and is never used as data, so all 60 runs behind the numbers above are new. §D.8 takes the gap this arm leaves as its subject: it screens every deployed oracle-free family for an instrument that would close it, reports what that screen returned, constructs an instrument when the screen returns nothing eligible, and runs it, with the result going against this paper a second time.

D.8 Separating the two channels: a screen that returns nothing eligible, a pre-registration of ours that failed its own gate before any result existed, and a constructed boundary instrument whose ladder contradicts the reading we lead with

Can the decision channel be moved while the adversarial coefficient share is held to the tolerance Mode S itself meets, and without reading adversary identity? §D.7 moved both channels at once and disclosed it post hoc, and the paragraph above records that the gate was recorded in that freeze's own rules block and then

not written. This subsection writes it. Four things are reported: a screen of the deployed oracle-free class that returns nothing eligible, a constructed boundary instrument that meets the gate on a supremum taken over every possible adversary set, so that no mask is needed to compute it, a pre-registration of our own that failed that gate in a premise check run before any result existed, which cost us a superseding document, and the ladder that document froze. **The branch that fired is the one the freeze named in advance as contradicting this paper's headline reading.** With the adversarial share of coefficient mass held an order inside `SHARE_TOL`, the tolerance Mode S itself meets, on every round of both arms, and with no adversary identity read anywhere in the construction, moving Krum's selection across a single boundary crossing moves ASR by $+\mathbf{0.322}$ at n=20, outside the frozen 0.15 margin and with the paired interval excluding zero. We report that as a contradiction rather than convert it into a scope condition, because that is what the freeze requires, and the paragraphs below state precisely what it does and does not contradict. The first three things reported read the update stack and no outcome; the fourth is an outcome measurement and is the only one of the four that bears on whether suppression survives the intervention.

The screen returns `NO_ELIGIBLE_FAMILY`, and an empty result is the outcome it was pre-registered to be allowed to return. Over 5 seeds and 3 live rounds on the same host cell as §D.7, Krum under the committed pixel backdoor, each of the four oracle-free d_1 families `apply_d1_transform` implements is measured on one raw stack per round: decision movement on Krum, admission flatness, and the adversarial share of coefficient mass. Every threshold is *imported*, the decision floor 0.10 and the admission-flatness bound 0.05 from the Mode M measurement and `SHARE_TOL` $= 10^{-6}$ from the admission measurement, so none was chosen after these numbers were read. The identity rung is exactly flat on all 15 rows, with $\max|\rho - 1| = 0$. `norm_clip` moves nothing at all, and that is a fact about τ=5.0 against these update norms rather than a property of the defense: the largest client norm seen over 5 rounds is 3.044, so $c_i \equiv 1$ and the transform *is* the identity. The other three all move the decision and all move the share past tolerance in 12 of 12 adversary rounds: `reputation` at decision rate 0.200 with share excursion up to 0.0998, `foolsgold` at 0.533 with 0.400 and an admission change of 0.067 that also breaches the flatness bound, and `rfa` at 0.533 with 0.0859, which is §D.7's own confound recovered by an instrument that was written to look for it. **The dichotomy is the finding**: on this cell every family that moves the decision moves the share, and the only share-neutral family is the one that does nothing. That is the quantitative shadow of §D.6's impossibility, measured across the deployed class rather than argued for one member, and it is *not* a claim that no such transform exists in general.

The constructed instrument, which we call the *boundary blend*, and why it does not contradict the impossibility. Blend each client's coefficient toward `rfa`'s, $c_i(t) = (1-t) + t\,c_i^{\mathrm{rfa}}$ for $t \in [0,1]$, reading c^{rfa} back from the transformed norms exactly as the admission measurement already does. Krum's argmin changes at isolated values of t while the share moves continuously, so a pair of points straddling one crossing has a decision change by construction and a coefficient difference that shrinks with the bracket. The difference vector does not depend on which clients are adversarial, so the neutrality is *uniform* rather than neutrality for one mask, and we measure it as such: the reported quantity is the supremum of the share gap over all 30 nonempty proper subsets of the K=5 participants, which upper-bounds the realized gap for every adversary set and needs no mask to compute. The impossibility in §D.6 is about *exact* neutrality, which oracle-freeness forces to hold for every adversary set at once and which therefore forces c constant; the blend is not exactly neutral, so it is consistent with that theorem, and what this instrument adds is the tolerance-parameterized statement that *approximate* uniform neutrality is reachable at any tolerance. The price is stated rather than hidden: informativeness is confined to rounds in which a crossing lies on the blend path, which is 13 of the 25 rounds measured here.

Our first freeze of that instrument failed its own share gate, and we report the failure against ourselves. It bisected the crossing of the float64 Gram statistic, while the shipped `krum_selection` runs in float32 and crosses at a slightly different t, so the converged bracket is tight around the wrong predicate: measured on live stacks before any arm was run, its supremum share gap exceeds `SHARE_TOL` in 13 of 13 flip rounds, and it fails its own *flip* gate as well. One sentence in that document needs correcting rather than defending, the number it offered as an existence proof of a within-tolerance pair being a minimum where a supremum is what the gate reads. The supplement (§S15) gives those measurements, that correction, and the outcome of the repair the document's own linearity argument implied, which is perfectly share-neutral and informative on nothing. **The general lesson is one sentence: bisect the predicate you intend to satisfy, and stop on the gated quantity rather than on a proxy for it.**

The corrected instrument, and an append-only supersession. The bisection now runs on the shipped selection itself; a float64 scan may only *propose* a bracket, which the shipped predicate must confirm at both endpoints, and an unconfirmed bracket yields no dose for that round rather than a dose the gate cannot vouch for. It halts when the mask-free supremum is at or below `SHARE_TOL`/10, which is the gated quantity and not a proxy. Measured on the same rounds: 13 of 13 shipped selection differences at a supremum of at most 9.60×10^{-8}, in 20 to 21 bisection steps, inside a step cap that is a *cost* bound and never a relaxed tolerance. The first freeze is left byte-untouched and superseded by `experiments/pre_registration_oracle_free_decomposition_refined.md` at `947c000`, which cites it, records the failure with the

measurements the supplement gives, pins the premise artifact by md5, and corrects that sentence, with the supersession enforced by the runner rather than promised; the supplement (§S15) carries the superseded document's own hash, that enforcement, the realized cost, and the four harness checks that ran before the ladder and whose verdict dict is persisted into the artifact. This is the same course §D.6 took with a frozen arm whose premise failed, and it is why the count of pre-registration documents rises rather than holding.

Both verdicts on the leg, read separately. The primary estimand is the paired difference $B - A$ at the same 20 seeds, 42–61, both arms new, no mixed-n leg and no top-up clause. It reads $+\mathbf{0.3218}$ (sd 0.3288, paired 95% CI $[+0.1680, +0.4757]$, half-width ± 0.1539, n=20). The *sign* verdict is that the interval excludes zero, positive. The *margin* verdict is that $|\Delta|$ lies outside the frozen 0.15 margin. Neither is reported as the other, because an interval can sit inside the margin and still exclude zero, and the freeze fixed the two labels separately for that reason. Mean ASR rises $0.2485 \rightarrow 0.5703$ while mean accuracy *also* rises, $0.4944 \rightarrow 0.5271$, so no accuracy collapse explains the move. The frozen 0.35 floor is live and no rung is void: the per-seed minimum is 0.3604 on arm A at seed 61 and 0.3713 on arm B at seed 55. The floor nonetheless sits closer here than on §D.7's leg 1, even though nothing crosses it on either: five arm-A rungs and four arm-B rungs come within 0.10 of it, nine of forty, against three of forty there, which is the count that paragraph now states after correcting an earlier claim of none.

What the $+0.322$ is, stated as the distribution rather than as its mean. This cell is bimodal in the same way §D.7's is, which is why a real effect carries an sd of 0.3288. Arm A has 4 of 20 seeds above ASR 0.5 and 13 below 0.10; arm B has 11 above 0.5 and 1 below 0.10. Six seeds move from the near-zero mode into the high one and none move the other way, so the effect is directional seed by seed rather than an artifact of averaging: one grid step across the crossing nearly empties the mode in which the backdoor fails to take hold. Eighteen of the 20 differences are positive and the median is $+0.2445$, below the mean, so the distribution is right-skewed and the mean is the larger of the two summaries. The mean is *attenuated* by arithmetic rather than inflated: four seeds already sit above ASR 0.9 on arm A with almost no headroom left, and their differences are $+0.033$, $+0.001$, $+0.010$ and -0.003 for that reason alone. Of the two negative differences only one is material, seed 42 at -0.108, from 0.434 down to 0.326; the other is -0.003 on a saturated seed.

The share gate is asserted on every round of both arms rather than sampled. Across all $2{,}000$ rounds, 2 arms $\times$ 20 seeds $\times$ 50 rounds, the mask-free supremum over all 30 nonempty proper adversary subsets never exceeds `SHARE_TOL`, and never exceeds the stricter stopping target of `SHARE_TOL`$/10$ either: its maximum anywhere is 9.999×10^{-8}, against the 3.5×10^{-7} Mode S holds to *with* a mask, and the largest per-client coefficient gap is 4.98×10^{-7}. Of those rounds $1{,}128$ carry a crossing and the shipped selections differ in every one of them, so the dose is delivered and not merely sought; the remaining 872 are bit-identical across the arms at a supremum of exactly 0. No round exhausted the 40-step cost bound, the most any round needed being 21 steps and 26 shipped Krum evaluations.

What the arm contradicts, and the one channel it still cannot separate. It contradicts the second half of the dissociation this paper leads with, that Krum's decision can flip while suppression stays inside the pre-registered ± 0.15 margin: on this cell, without adversary identity, and with the attenuation channel held to 10^{-7}, an order inside `SHARE_TOL` and $3.5\times$ tighter than the 3.5×10^{-7} Mode S reaches with a mask, a decision flip carries $+0.322$ and leaves that margin. We report that as a contradiction and do not convert it into a scope condition. **But it separates the decision channel from attenuation only, not from admission**, and that limitation is structural rather than an omission we could repair: deciding whether a flip moved Krum from a benign client to an adversarial one requires exactly the mask this instrument refuses, so the artifact records no admission quantity and could not. The $+0.322$ may therefore run through admission, through the scoring alone, or through both, and this arm does not decompose it. Two consequences follow and we state both. The frozen verdict is not amended by that limitation, on the same principle §D.7 applied to its own post hoc measurement: the refuting branch fired on its own stated terms and we do not use a caveat discovered afterwards to retract it. And the contrast is *local*, at the boundary: both arms are nearly-`rfa` stacks at adjacent points on the blend path, the per-run mean crossing sitting at $t^\star \in [0.578, 0.832]$, so this is not an identity arm against a statistic disturbance and must not be read as one. The flip direction is uncontrolled, which the freeze records in advance, so $+0.322$ is the price of *perturbing* Krum's selection at the boundary and not a benefit a defense could collect by choosing a side.

What this subsection does not establish. The masked Mode S measurement stands exactly as measured; what this arm denies is its generalization to this cell without the oracle, and §D.7's standing summary, that the oracle is load-bearing for our negative rather than a laboratory convenience, is strengthened by it rather than revised. A reader looking for an oracle-free measurement of the statistic channel will not find one here either, and §D.7's standing statement that we have not produced one is unchanged, for the reason the paragraph above gives: that measurement needs an identity arm against a disturbed arm, and this ladder compares two disturbed arms one grid step apart. The screen's empty result is a statement about four families, at one cell, on 5 seeds and 3 rounds, at Mode S's tolerance, and the premise check is 25 rounds of instrument behaviour with no outcome attached. The ladder is one aggregator, one attack, one dataset and one upstream transform, at 20 seeds paired within arm, so it licenses no statement of the form *the negative fails without an oracle* in general, any more than it would have licensed its converse; what it licenses is *in this cell, without adversary identity,*

with the coefficient share held to 10^{-7}, and that is the only form used above. Its freeze named the branch that contradicts this paper's headline reading as plainly as the branch that supports it, and the branch that fired is the first of the two. One population note, so that a number is not mistaken for a revision: the realized `rfa` spread over the 25 rounds measured here is $\rho \in [1.306, 2.708]$, against $[1.306, 2.194]$ reported in §D.7 over its own 15 rounds, which is a larger population and not a correction to that interval.

D.9 Mode S on the one cell where both margin legs bind: the two-sided containment fails, and the one-sided non-increase holds by a wide margin, in the direction of *more* suppression

What does the Mode S contrast return on a cell where both readings of preservation have somewhere to fall? The cell this paper scores its flagship reading on is uninformative twice over, and the supplement §S18 is this paper's admission of the second half. Krum under committed scaling admits 0.000 adversarial mass at every rung, so the zero there is a floor rather than a measurement; and that cell's identity rung has mean ASR 0.0439, so with ASR bounded below by 0 a fall of 0.15 is arithmetically unavailable and only the upper margin leg is a test. This arm moves to the cell §D.7 identified for exactly this reason, Krum under the committed pixel backdoor, where baseline adversarial admission is 0.333 in 4 of 12 adversary rounds and the identity rung's mean ASR is 0.3232, above the frozen 0.15: **both legs are genuine tests here, so a two-sided reading would be earned if one held**. It does not hold. **The outcome falls in a branch our own freeze did not enumerate, and we report that as a defect in the freeze rather than assign the result to the nearest branch it did name**: the paired difference is large, negative, and outside the margin, so the dose does not leave suppression where it found it, it improves suppression sharply. Pre-registered in `experiments/pre_registration_modeS_pixel_headroom.md`, committed as its own commit at `c89e963` with both output paths verified not to exist; 40 new runs at 20 seeds, 42–61; artifact `results/modeS_pixel_headroom/summary.json`.

The freeze registers no prediction for the sign, and that was an admission rather than caution. Two of our own measurements pointed opposite ways before this arm ran: the flagship's -0.010 under a masked dose on the scaling attack, and a positive, margin-leaving difference under an oracle-free construction on this very cell (§D.8). Registering a direction would have been choosing which of our own results to believe, so what was frozen is the branch structure instead: three verdicts per estimand, *sign* (does the paired 95% interval exclude zero), *margin* (is $|\Delta|$ below the frozen 0.15) and *the one-sided non-increase* (is the 95% upper bound below $+0.15$), with a standing prohibition on reporting any one of the three as another. **Four branches were enumerated and none of them covers what happened**: *confirming* (interval containing zero inside the margin), *refuting* (interval excluding zero with $|\Delta| > 0.15$ toward *increased* ASR, reported as a contradiction and never converted into a scope condition), a small consistent rise that the document called the one to expect, and *unresolved*. All four presuppose that a move which leaves the margin is upward, and the measured move leaves the margin downward. **That gap is ours and we name it rather than round the result to the nearest branch**: what the freeze does bind is the three verdicts, the prohibition on substituting one for another, and the rule that no mechanism may be written inside a verdict literal, and those are what the paragraphs below are held to. Two endpoint rungs only, κ=0 and κ=2 at ρ=1 and ρ=54.598: the interior rungs are not run, no Jonckheere–Terpstra statistic is computed for this arm, and it is nowhere described as a ladder.

The identity leg is imported rather than re-run, and the import is asserted at the bit level before the ladder starts. Mode S at κ=0 returns the update stack unwrapped before the permutation is drawn, and the permutation when drawn uses a local `default_rng`([seed, round]) rather than the global stream, so this path and §D.7's identity arm are the same computation. That is asserted and not assumed: the published row's accuracy and ASR at seed 42, 0.3962 and 0.7787777777777778, are reproduced through this runner at $\Delta = (0, 0)$ on both quantities, and again with `score_only` set, which is why both estimands share one subtrahend. **The check is blocking rather than decorative**, because the published rows ran with `adv_mask=None` through a `d1_override` and this path passes the real mask; had it failed, the freeze required re-running both identity legs as 40 further runs with no reconciliation attempted. A third check establishes that the dose pins the adversary on this cell as it does on the published one: over 32,000 coefficient vectors, every adversarial coefficient is exactly 1 (maximum deviation 0.0e+00) and the adversarial share of coefficient mass equals n_{adv}/K to 5.6×10^{-17} against a 10^{-12} tolerance, taken *exhaustively* over all 32 subsets of the K=5 participants rather than over the masks the runs happened to draw. All three verdicts are persisted into the artifact, because a check whose verdict is discarded is witnessed only by a terminal.

The three verdicts on Δ_{full}, read separately. The primary estimand is the paired difference between Mode S at κ=2 and the identity rung at the same 20 seeds. It reads $\mathbf{-0.2857}$ (sd 0.4243, paired 95% CI $[-0.4842, -0.0871]$, half-width ±0.1986, n=20). The *sign* verdict is that the interval excludes zero, negative. The *margin* verdict is that $|\Delta|$ lies *outside* the frozen 0.15, so the two-sided containment this paper reports on the flagship cell fails here. The *one-sided* verdict is that the 95% upper bound on a rise is -0.087, which is not merely below the frozen $+0.15$ but below zero, so the non-increase holds more strongly than on any arm in the supplement's Table S8. Those are three independent labels and none is reported as another: this arm is the mirror image of the flagship, where the two-sided reading passes on a leg that was never at risk, and it is

the clearest evidence we have that the relabelling of the supplement §S18 is substantive rather than cosmetic. Mean ASR falls $0.3232 \to 0.0375$ while mean clean accuracy *rises*, $0.5158 \to 0.5549$, so no destroyed model explains the fall; the frozen 0.35 floor is live, no rung is void, and the per-seed minimum rises from 0.3962 on the identity arm to 0.4983 under the dose.

What the -0.286 is, stated as the distribution rather than as its mean, and the per-seed form of the reachability arithmetic. This is the same bimodal cell §D.7 reports, and the imported identity arm reproduces that paper's own count exactly: 6 of 20 seeds above ASR 0.5 and 12 below 0.10, with a maximum of 0.988. **Under the dose the high mode is empty**: no seed reaches 0.11, the largest dosed ASR anywhere being 0.1074, and all six high-mode seeds fall below 0.10 while none moves the other way. So the effect is directional seed by seed rather than an artifact of averaging. The median difference is only -0.0252 against a mean of -0.2857, and 12 of the 20 differences are smaller than 0.05 in absolute value, which is *arithmetic and not noise*: 13 of the 20 identity rungs sit below 0.15, so those seeds individually cannot fall past the margin however good the defense becomes. The seven seeds whose identity ASR exceeds 0.15 are *exactly* the seven whose difference falls past -0.15, on both estimands. Reachability therefore governs this cell per seed even where it does not govern it in the mean, and a cell can clear the margin on its identity mean while most of its seeds cannot. Five differences are positive and all five are small, the largest being $+0.107$ on seed 61, whose identity rung reads ASR 0.0001. The Reproducibility statement's per-seed promise covers this arm, and these are the seven values that carry it, identity ASR against the full dose and then the score-only dose: seed 42, $0.779 \to 0.013$ and 0.001; 44, $0.932 \to 0.053$ and 0.011; 49, $0.419 \to 0.088$ and 0.023; 51, $0.925 \to 0.052$ and 0.011; 54, $0.979 \to 0.020$ and 0.000; 58, $0.983 \to 0.023$ and 0.002; 59, $0.988 \to 0.007$ and 0.051. The other thirteen seeds start at 0.110 or below and end at 0.107 or below under the full dose and 0.079 or below under the score-only one, and the artifact's `per_seed` rows carry every value quoted in this subsection.

The magnitude channel is closed and the fall survives, and this is the score-only control of §D.2 run at all 20 seeds with a paired interval. Δ_{score}, the same contrast with `score_only` Krum scoring the transformed stack and emitting the selected client's original update, reads $\mathbf{-0.2974}$ (sd 0.4288, CI $[-0.4981, -0.0968]$, n=20), with the same three verdicts and the same branch. **Closing the magnitude channel moves the effect by** 0.012, and no seed's two differences disagree by more than 0.105, so the route to the fall is not the magnitude or geometry of the aggregated update. Mean accuracy rises further on this leg, $0.5158 \to 0.6149$ at a per-seed minimum of 0.5149, and every dosed rung is below ASR 0.08. This leg was staged behind the first and its running condition is recorded in the artifact: it would have been skipped had Δ_{full} fired the refuting branch, so the staging could only ever have withheld a control that followed an unfavourable-to-nobody result, never suppressed an unfavourable one. **This is a second cell rather than a top-up of the flagship one**, and the two Δ_{score} values are what that distinction buys: -0.2974 here on the committed pixel backdoor against -0.0088 there on committed scaling, both at 20 seeds. The flagship's score-only control has since been extended to n=20 on its own cell under its own freeze (§D.2), so §D.2's published n=5 verdict is now reported beside an n=20 one rather than alone; neither arm restates the other's cell.

The same-cell contrast with our two oracle-free arms, and what it does and does not identify. Both published arms on this cell move ASR *up* by about a third: §D.7's leg 1 reads $+0.2969$ (95% CI $[+0.0983, +0.4955]$, n=20) and §D.8's boundary ladder reads $+0.3218$ ($[+0.1680, +0.4757]$, n=20). The first of those is the directly comparable one, and comparable in an unusually strong sense: it is also a `score_only` arm, at the same 20 seeds, against the *bit-identical* identity rows this arm imports. Two score-only interventions on one cell against one baseline therefore return $+0.2969$ and -0.2974, near mirror images. **What this removes is one of the two confounds that paragraph disclosed**, which recorded that its refutation could not transfer to the published Mode S result because the two interventions differed in the attack *and* the upstream transform and it decomposed neither: holding the cell and the attack fixed, the disagreement survives, so the cell is not a sufficient explanation of it. **That inference is post hoc and is marked as such.** The freeze attached a conclusion of this shape to its confirming branch, which did not fire, so what carries it here is not a branch's authority but the arithmetic that two interventions against one baseline returned opposite signs. **What this does not do is identify the cause**, because the two arms still differ in two respects at once: `rfa` reads no adversary identity and this dose reads it by construction, and `rfa`'s realized spread is $\rho \in [1.306, 2.194]$ against this dose's ρ=54.598. Oracle access and dose magnitude are not separated by anything here, and the sign difference is not attributed to either. §D.7's standing summary, that the oracle is load-bearing for our negative rather than a laboratory convenience, is untouched by this arm, as is its statement that we have not produced an oracle-free measurement of the statistic channel on any cell: **this arm is masked, so it is not one**.

What is not measured here, so that nothing is inferred from a quantity this arm never produced. `run_one` returns clean accuracy and ASR and nothing else, so **no channel quantity and no admission quantity is measured by this arm at all**. The 0.333 baseline admission that made the cell eligible is *imported* from the frozen measurement §D.7 reports and is not re-measured here, and Krum's decision-change rate under this dose on this cell is simply unknown. So it may not be said that a statistic disturbance costs or buys suppression here, in either direction: what is measured is that the same masked dose, at the same endpoint, with the adversarial coefficient share pinned to 5.6×10^{-17}, moves ASR down by 0.286 on a cell whose identity mean 0.3232 left a fall of 0.323 available and a rise of 0.677, so the room was not symmetric and the measured move used

88% of the smaller side. Nothing here bears on whether admission predicts security, which this paper denies establishing, and the improvement is not a defense: Mode S needs an `adv_mask` no deployed server has, and `score_only` Krum aggregates an update its own scoring stage never saw.

What this arm does not establish. One aggregator, one attack, one dataset, one architecture and two rungs license no statement of the form *the dissociation holds generally* and none of the form *it fails generally*; what they license is *in this cell, at* n=20*, with both margin legs binding and admission off its floor*, and that is the only form used above. No published row is revised: the flagship's -0.010, 95% CI $[-0.032, +0.012]$ at n=20 on Krum under committed scaling stands exactly as printed, and this arm is a different cell rather than a correction to it. No pre-registered directional prediction is refuted, because the freeze registered no sign; the word *refuted* keeps the meaning this paper fixes for it throughout and is not used of this arm. The interval is wide, ± 0.199, which a bimodal cell at 20 seeds buys, so the magnitude is poorly determined even though its sign and its position relative to the margin are not. And the two-sided containment failing here says nothing about whether it would fail on the flagship cell, where it cannot be tested on both legs at all, which is the asymmetry the supplement §S18 exists to make visible.

D.10 Closing the magnitude channel on the headline cell: the within-defense rise does not survive it, and the two designs then agree in sign

The paper's headline is one cell estimated two ways. What happens to the within-defense leg when the aggregate-magnitude channel is closed on it? Every magnitude control elsewhere in this paper is defined for **selectors only**: `generic_compose` raises for any d_2 outside `krum` and `cos_krum`, because a coordinate-wise order statistic emits no single selected update whose magnitude could be held fixed. So the one cell the abstract, §1, §5 and the Conclusion are built on, `coord_median` under `committed_pixel` on CIFAR-10 / `cifar_cnn`, was the one cell whose magnitude channel had never been closed. This arm closes it. **The result is against us, and the freeze fixed in advance where it would be reported if it came out this way.**

The instrument, and why it is a new module rather than a transfer. Score-only coordinate median takes, for each coordinate j, the argmedian over clients of the *transformed* stack and emits the *untransformed* value of that client at that coordinate: $\mathrm{idx}[j] = \arg\mathrm{median}_i\, T(u)_i[j]$, then $\mathrm{out}[j] = u_{\mathrm{idx}[j]}[j]$. At an odd client count (K=5 in every round here) `torch.median` returns an actual element of the stack together with its index, so this is `coord_median`'s own computation with `.values` replaced by a gather at `.indices` into the untransformed stack, and nothing else differs: same flattening, same cast, same per-parameter loop, same reshape. **What it closes and what it deliberately leaves open are both already measured at this exact cell and rung.** It closes Table A8's Δ agg. of 0.670, so no emitted number is a rescaled one and every coordinate of the aggregate is some client's own value; it leaves the re-selection channel open at Δ dec. $= 0.482$, so nearly half of all coordinates still change which client supplies them, which is the channel the paper's claim is about. Like score-only Krum it is a laboratory instrument and not a defense: it aggregates values its own scoring stage never saw, on top of Mode S already reading adversary identity.

Frozen before the instrument existed, with all three branches written down. `experiments/pre_registration_score_only_coordmedian.md`, git-committed alone at `16a17b2`; the runner refuses to start unless that hash matches the log *and* the file is clean in `git status --porcelain`, since a recorded hash alone is only half a guard. One rung, κ=2.0: κ=0.5 and κ=1.0 are not run, and no display prints this cell's ladder with a rung of this arm in it. Seeds 42–61, n=20, fixed in advance with no interim look and no seed dropped for any reason. The κ=0 leg is imported from the published controlled legs under an *exact* identity: `apply_d1_transform` at κ=0 returns the caller's list object rather than multiplying by 1.0, so the transformed and untransformed stacks are the same object and a gather at `.indices` returns exactly `.values`. The freeze required that assertion to read 0.00e+00 rather than a tolerance, on seed 42 where a published value exists, with the arm not running at all if it failed; it read exactly that. **A second gate is the decisive one against a misread cell key**: the two imported legs must reproduce the *published* n=20 controlled delta they are scored against, and they reproduce $+0.125061$ to 1.11×10^{-7}. Artifact `results/score_only_coordmedian/summary.json`.

No margin test is run here and none is imported, and that is the freeze's instruction rather than an omission. The ± 0.15 practical-equivalence margin scores arms whose claim is a *non-increase*; this arm's claim is about the *sign and size* of a rise, so its decision is on the interval's relation to zero and to the two published intervals. Importing a margin afterwards would be choosing a rule after the numbers. This is the one respect in which this arm must not be read like §D.2's, which is a margin arm.

$\Delta_B = \mathbf{-0.1722}$**, paired** 95% **CI** $[-0.2119, -0.1325]$**, which excludes zero on the negative side.** sd 0.0848, t_{19}=2.093, n=20 on exactly seeds 42–61; mean ASR 0.4783 at κ=0 against 0.3061 under the control at κ=2, per seed in the supplement's Table S3. Mean clean accuracy under the control is 0.7500, clearing the frozen 0.35 floor by $+0.4000$, so no rung is a collapsed model and the fall is not a destroyed one. The frozen verdict literal this outcome matches is **the control reverses the controlled leg's own sign**, and the same document fixes how it is to be read: *reported as such and not scored in our favour, because with magnitude closed the*

controlled design agrees in sign with the outcome-gated one, which removes the dissociation on this cell. The secondary estimand is $\Delta_{\text{mag}} = +\mathbf{0.2973}$, CI $[+0.2527, +0.3418]$, mean ASR 0.6034 uncontrolled against 0.3061 controlled: the measured effect of closing magnitude at this rung on this cell. Per caveat 5 of the freeze it is **not** described as the magnitude channel's share of Δ and is not summed with anything, because no 2×2 factorial exists on this cell at all (this arm runs one of its four cells) and §D.2.1 already withdraws any reading of a channel split as a decomposition of suppression.

The sign-reversal claim, re-adjudicated explicitly against both published intervals rather than by implication. The freeze fixed a two-part test: the reversal reproduces with magnitude closed only if *(i)* Δ_B's interval excludes zero on the *positive* side and *(ii)* it fails to overlap the outcome-gated interval. *(i)* is **no** and *(ii)* is **yes**, so the two do not both hold and **the published reversal does not reproduce with the magnitude channel closed**. Set beside the two published legs at the same 20 seeds, outcome-gated -0.2733 $[-0.3337, -0.2128]$ and within-defense $+0.1251$ $[+0.0954, +0.1547]$, Δ_B overlaps neither. **But its non-overlap with the outcome-gated interval is by** 0.0009**, and we say so rather than let "the intervals do not overlap" carry a weight it cannot bear here**: the two ends in question are -0.2119 and -0.2128. What this cell now shows is two designs agreeing in *sign* and differing in *magnitude* by an amount 20 seeds barely resolve. The honest summary is that the *direction* the two designs disagreed on was carried by the aggregate-magnitude channel, and the residual disagreement is one of size.

What this does not revise, stated as narrowly as the evidence allows. No published row changes and none is withdrawn: the outcome-gated -0.273 and the within-defense $+0.125$ are both reproduced here through the import gate above, and both remain what the two designs return on this cell, the within-defense one now known to depend on leaving magnitude open. **The control is applied to one design only, and that asymmetry is ours to state**: the outcome-gated leg is quoted as published, no magnitude control exists for it, and none is claimed, so this is a magnitude-closed leg set beside a magnitude-open one. Nothing here touches the formal results: Proposition 3's non-identification and the exact invariances are algebra about what an outcome-gated test can distinguish, and no channel measurement bears on them. What narrows is the *empirical* claim built on this cell, and §5, Table 1 and Table A15 now carry it in the narrowed form: the two designs disagree in sign here *when the within-defense design leaves the aggregate-magnitude channel open*, which is what Mode S as published does.

Four clauses travel with these two numbers, and one of them is that our own recorded prior was wrong. *(i)* **Magnitude is closed; the trajectory channel is not.** Which client supplies each coordinate still changes across rungs, so the model trajectory still diverges, and the claim is about magnitude. *(ii)* **The oracle is untouched.** Mode S still reads adversary identity to pin $c_{\text{adv}}=1$, so this arm removes no oracle and licenses no statement in either direction about behaviour without one. *(iii)* **A coordinate-wise control is not a selector control**, and the transfer is refused in both directions: score-only Krum holds one client's whole update fixed, this holds each coordinate's contributing value fixed, and no pooled number combines them. In particular §D.2.1's finding that the magnitude channel is *inert* on Krum does not transfer here, and this arm's finding that it is decisive here does not transfer back. **Read together the two arms are a dissociation across aggregators rather than one result at two seed counts**: on Krum under committed scaling the $n=20$ Mode S mean is -0.010 uncontrolled and -0.0088 with magnitude closed (§D.2), while on `coord_median` under the committed pixel backdoor it is $+0.1251$ uncontrolled and -0.1722 with magnitude closed. *(iv)* **One cell, and this arm is a disclosed post hoc extension** of a frozen suite rather than a prospective test: the published controlled point estimate was known when the document was written, which is why the reachability of the adverse branch was recorded in advance. **The freeze's own stated prior was that this arm was more likely to confirm than to refute**, and it names the arithmetic it rested on: refuting needed the magnitude channel to account for roughly 0.095 of the 0.125 rise, and the measured effect was three times that. We print that prior as it was written rather than quietly dropping it, because a pre-registration that records only the priors it got right is not one. **And the control covers this cell only**: `coord_median` under committed *scaling*, the arm carrying the $+0.196$ rise at $n=5$, is a different attack and is still not magnitude-controlled.

E THE COMPOSITION THE SCREEN CANNOT CERTIFY: FOOLSGOLD→RFA

This section is the empirical witness for Proposition 8. FoolsGold→RFA has the *lowest observed committed ASR* of every composition evaluated here (0.093), and **our own criterion classifies it as a false negative**: neither constituent suppresses the pixel backdoor alone (FoolsGold 0.732, RFA 0.845), so its suppression is *emergent* in the sense of Proposition 8 and C1 cannot hold. **We report it in detail because it is the sharpest illustration of the screen's scope, not because it is a flagship**: any criterion that certifies inherited suppression must miss compositions of this kind, which is what caps recall *structurally*; with one emergent case among five low-ASR pairs that cap is 80%, and the 40% we measure sits below it for reasons other than emergence. The characterization below is therefore descriptive (what this composition does, measured at $n=30$) and is not evidence for the criterion. **The** $n=30$ **measurement corrected our own earlier figure downward in reliability, and we report the correction rather than the settled number alone**: the per-seed distribution is right-skewed (median 0.060, mean 0.093, sample skew $+2.07$, $8/30$ seeds at or above 0.10, maximum 0.380),

and seeds 42–46 fall in its left tail, so the n=5 value we first led with understated the pair by a factor of 1.8. Any 5-seed figure for this composition should be read against that spread. We avoid calling it security-maximal: it also sustains the *highest* ASR under our adaptive adversary (0.157 against 0.064 for the same pixel attack on the same thirty seeds; at n=5 those two read 0.167 and 0.045, so the ratio falls from 3.7× to 2.4× once the denominator leaves its left tail), and it carries a 50% accuracy cost, so it is the right choice only where utility is nearly irrelevant. We validate at n=30 seeds with a full adaptive attack suite (N=10, K=5, f=0.2, 50 rounds), in Table A12:

Table A12: FG→RFA under base and adaptive attacks. Base: n=30, reported as median [IQR] with P95 tail; adaptive: mean±std at the n each row names, which is n=5 except for the ϵ=1.0 arm the body quotes, topped up to n=30. P95 is reported only where it was pre-registered, hence − on every adaptive row. Clean accuracy: test-set accuracy on non-triggered inputs. No-attack baseline for RFA alone: 0.737. Consensus-shift row: this is a re-run, and the row published earlier (0.045 ± 0.012) was void. That runner built its attack object from the `backdoor_pixel` strategy, which does not override `manipulate_update` and so inherits the base class identity; its shift term was therefore scaled onto a zero difference, and both rates reproduced the base pixel arm on all five seeds to twelve decimal places in both ASR and clean accuracy. Both rates were re-run on the same seeds, composition and configuration with the live `ConsensusShiftAttack` (`results/fg_rfa_consensus_shift_fixed/`, pre-registered at `784a201`), and the superseded `adaptive_consensus_shift` cells are kept in `results/fg_rfa_flagship/summary.json` so the defect stays auditable.

Attack	Config	ASR	P95 / max	Acc
Base composition (n=30; median [IQR])				
	scaling	0.060 [0.048, 0.096]	0.253 / 0.380	0.497
	pixel	0.052 [0.035, 0.087]	0.127 / 0.200	0.492
	oracle	0.056 [0.044, 0.090]	0.189 / 0.236	0.494
Projection adaptive (n=5)				
	ϵ=2.0 (best)	0.079 ± 0.042	− / 0.160	0.495
Criterion-aware joint-constraint				
	ϵ=1.0, decorr (n=30)	0.157 ± 0.073	− / 0.353	0.549
	ϵ=2.0, decorr (n=5)	0.130 ± 0.030	− / 0.185	0.506
	ϵ=2.0, no decorr (n=5)	0.050 ± 0.008	− / 0.059	0.494
Neurotoxin (n=5)				
	top-20%, scale = 2	0.048 ± 0.009	− / 0.057	0.495
Consensus-shift (n=5; re-run, see caption)				
	rate = 0.05	0.045 ± 0.010	− / 0.058	0.497

Results. Median ASR across all base attacks (n=30): 0.056 [IQR 0.044–0.090]; P95 worst-case: 0.253 (committed_scaling). **Three of the six adaptive strategies beat the base attacks and we lead with the strongest, not the median.** Five of the six are measured at n=5 on seeds 42–46, where the base composition reads 0.045 (pixel), 0.051 (scaling) and 0.050 (oracle); the ϵ=1.0 arm the body quotes is topped up to n=30 on seeds 42–71, where the pixel arm's *mean* is 0.064 (the base block above prints medians, so the same cell appears there as 0.052). Against those seed-matched bases, the criterion-aware joint-constraint adversary with decorrelation reaches mean 0.157 at ϵ=1.0 and n=30 (median 0.159, max 0.353; paired difference against the pixel arm +0.093, 95% CI [+0.066, +0.120]), which is 2.4× the 0.064 of the pixel arm it is built from and the ratio every other section of this paper quotes, and 0.130 at ϵ=2.0 and n=5, and the projection adversary reaches 0.079; the other three sit level with the base (the same criterion-aware attack without decorrelation 0.050, Neurotoxin 0.048, consensus-shift 0.045 at rate 0.05 and 0.048 at rate 0.01). Consensus-shift is the one row here we re-ran after finding our published figure void (the defect is disclosed in Table A12's caption): its paired differences against the pixel arm on the same five seeds are +0.0031 (95% CI [−0.0005, +0.0067]) at rate 0.01 and −0.0006 ([−0.0042, +0.0029]) at rate 0.05, so both intervals contain zero, and the attack's second phase is by construction ordinary committed pixel at the scale=1 this arm runs, so part of the null is mechanical. Since the base arm's own n=30 distribution is right-skewed and seeds 42–46 sit in its left tail, a 5-seed null is weak evidence of a null rather than a demonstration that the attack cannot work here. **So the composition does *not* stay under 10% against the strongest adversary we built**, and 0.157 is the figure the opening paragraph reports. **The top-up moved that ratio down and we report the move rather than the settled figure alone**: at n=5 the same two arms read 0.167 and 0.045, a ratio of 3.7×, and almost all of the fall is the denominator leaving the left tail described above rather than the attack weakening: the numerator moved only 0.167 → 0.157 while the denominator rose 0.045 → 0.064, and holding the numerator at its n=5 value while the denominator alone is recomputed at 30 seeds would already put the ratio at 2.6×. What survives is weaker and still worth stating: no adaptive *mean* reaches the base composition's own P95 tail of 0.253 (its

worst single seed is 0.380), though the ϵ=1.0 maximum of 0.353 does, and the ϵ=1.0 gain costs the adversary nothing in clean accuracy (0.549 against the pixel arm's 0.492, both at n=30), so it is a real gain rather than a utility trade. The accuracy tradeoff (0.50 vs. 0.74 for RFA alone) and Pareto alternatives are discussed below (Table A13).

Composition-aware adversary (end-to-end). We build an adversary that *knows* the full FG→RFA pipeline is deployed (Kerckhoffs) and optimizes against the complete composition, solving both constraints simultaneously: pairwise cosine similarity below FoolsGold's down-weighting threshold, *and* ℓ_2 distance to the geometric median below RFA's rejection radius. Any gradient-based optimizer faces the same tension (coordinating a backdoor raises FG-similarity, while staying near the median caps the perturbation and weakens the backdoor) and our analytical construction reaches the same boundary, the Weiszfeld iteration being contractive and the gradient at convergence determined by the final-iterate reweighting the joint constraint directly constrains.[1] The ablation is informative: without decorrelation ASR stays at baseline (0.050), RFA-evasion alone failing because FoolsGold is the binding constraint; with it the adversary gains partial signal (0.130–0.167 across the ϵ sweep, held at the uniform n=5 the ablation compares at; ϵ=1.0 alone reads 0.157 at n=30), which Theorem 9 explains as bounded, each adversary's block-restricted delta being too small to move the median outside the benign cluster.

Contrast with rep→tm. FG→RFA (median ASR 0.056) provides near-absolute suppression; rep→tm (oracle ASR 0.316) provides strategic dominance against all tested adversary policies but does not suppress committed_pixel (0.679). The two serve different roles.

Accuracy–security tradeoff. FG→RFA achieves the lowest ASR of any tested composition but imposes a substantial clean-accuracy cost (0.50 vs. 0.74 for RFA alone). This cost is intrinsic to FoolsGold's aggressive per-client downweighting, not to composition itself. Table A13 places FG→RFA on the accuracy–security Pareto frontier alongside alternative compositions:

Table A13: Accuracy–security Pareto frontier. Max ASR = worst-case mean across committed attacks, and Clean Acc is measured on the arm that attains it. All at N=10, K=5, f=0.2, cifar_cnn. The three composition rows are emitted by `experiments/analyze_headline_cis.py`, each at the seed count its own max-committed arm carries: FG→RFA and Rep→CM at n=30 (seeds 42–71; Rep→CM's is the pre-registered pixel top-up, 0.463 [0.412, 0.514]), FG→CM at n=5 (seeds 42–46). No *composition* dominates another: FG→CM buys 0.185 accuracy for +0.014 ASR. The single-defense rows are listed for reference and are not all frontier points; `Rep alone` dominates `RFA alone` on both axes.

Defense / Composition	Clean Acc	Max ASR	Regime
Temporal mix (NC/fedavg)	0.79	0.96	No security
RFA alone	0.74	0.91	Bimodal; unreliable
Rep alone	0.78	0.84	Scaling-only
Rep→CM	0.76	0.46	Moderate
FG→CM	0.68	0.11	Strong
FG→RFA	**0.50**	**0.09**	Min ASR (evaluated)

FG→RFA is the correct choice under a *security-first* objective (e.g., safety-critical FL where any backdoor is unacceptable). For operators with accuracy constraints, FG→CM (acc 0.68, max ASR 0.107) buys 37% more accuracy for suppression that is statistically indistinguishable on the worst-case arm the two share (Welch $p = 0.52$ on committed scaling, 0.093 against 0.107). It is *not* Pareto-superior and we do not claim it is: its point estimate is the higher of the two, and on the pixel arm, which is neither pair's worst case, the gap does reach significance in FG→RFA's favour (0.064 against 0.099, Welch $p = 0.022$). Rep→CM (acc 0.76, max ASR 0.463) offers near-baseline accuracy with moderate security. The criterion certifies FG→CM and Rep→CM; FG→RFA reaches the lowest ASR of the three but fails the tightened per-attack C1 and is therefore a false negative of our own criterion rather than a certified recommendation. The choice among them is an operator-specific utility decision, not a criterion failure.

[1] A fully gradient-based optimizer differentiating through 50 Weiszfeld iterations would face identical binding constraints at convergence; the analytical formulation is equivalent without requiring a differentiable relaxation. **The equivalence is about the constraint set at convergence and not about the optimizer, and we do not run the unrolled attack**, so two classes of behaviour sit outside it. Path effects would make an unrolled attack *weaker* than our construction: gradients that vanish or are obfuscated through the iteration, and optimization instability or a poor local optimum. A weaker unrolled result would therefore not refute the number here. The case the argument cannot exclude runs the other way: our construction searches a family we chose and parameterized in closed form, and differentiating through the iteration may reach a point inside the same constraints that this family does not contain. So 0.157 at n=30 is the strongest our family attains and is not a bound over gradient-based adversaries.

F RELATED WORK

The contrast in one sentence. Prior work asks whether a defense withstands an attack, and the composition literature treats composition quality as an *outcome to predict*. We ask a different question: the **identifiability of mechanistic explanations** for that outcome, under the evaluation designs commonly used to compare defenses. Concretely, whether an evaluation criterion can causally attribute a composition's security to preservation of the downstream mechanism.

Byzantine-robust aggregation, studied in isolation. Krum (Blanchard et al., 2017), Trimmed Mean (Yin et al., 2018), RFA (Pillutla et al., 2022) and FoolsGold (Fung et al., 2020) are each proposed and evaluated as a *single* stage. **Two names have left that list, and the composition audit is what removed them** (the supplement §S16). FLAME (Nguyen et al., 2022) combines three mechanisms in one round (model clustering, weight clipping and adaptive noise), and it gives the constituent-complementarity reason for combining them, clustering and clipping being there to "minimize the required amount of noise" a differentially private defense would otherwise need. FLTrust (Cao et al., 2021) applies a server-bootstrapped trust score *and* a magnitude renormalization to the same round, which is two mechanisms under any granularity the audit's distinctness rule admits. **The FLTrust error was ours to catch one section earlier, and we did not**: §S24.1's near-identity finding rests on precisely that two-part structure, since FLTrust gates on the scale-invariant $\cos(u_i, g_{\mathrm{srv}})$ and then renormalizes every update to $\|g_{\mathrm{srv}}\|$, which is why a positive per-client rescaling is exactly cancelled. So this paper described FLTrust as two mechanisms where the finding needed two and as a single stage where the contrast needed one. Both are composed defenses and both are coded as such below. **Bulyan** (El Mhamdi et al., 2018) is the earliest composed defense among those we name and predates the framing here: it runs Krum-style selection and then a trimmed-mean-style aggregation over the selected set, so it is an ordered pair of exactly the kind we study, chosen by its authors rather than by any screen. It falls outside the audit's frame as frozen, and precisely why matters: the frame's conference screen requires `federated` in the title, which Bulyan's does not carry, and the arXiv sweep runs 2019–2026 (§S16). Its `references.bib` entry is added here, *after* the audit was coded, so the audit's first route never saw it, and we report that rather than re-run a frozen instrument: Bulyan is named as related work and is *not* a retrospective audit hit. Every count there is unchanged. We propose no new defense; we reason about *ordered pairs* of existing ones, which is a different object: the upstream stage's output is the downstream stage's input, so the question is no longer whether a mechanism works but whether it *survives* a transform applied before it. Ensemble adversarial training (Tramèr et al., 2018) *assumes* that upstream processing preserves downstream signals; we make that assumption explicit, checkable, and (as it turns out) **the wrong abstraction for security**.

Adaptive attacks are why an inherited-robustness guarantee is not a security guarantee. Optimized model poisoning defeats Krum, trimmed mean and FoolsGold once the attack is chosen *against* the defense (Shejwalkar & Houmansadr, 2021; Bhagoji et al., 2019), and the same holds for distributed and durable backdoors (Xie et al., 2020; Wang et al., 2020; Zhang et al., 2022). Our C1 quantifies over a *committed* attack by construction, so this literature bounds what any inherited-suppression criterion (ours included) can promise. Our single whitebox composition-aware stress test (§E) is reported in order to *measure* that gap ($0.064 \to 0.157$ seed-matched at $n{=}30$ on the pixel arm, $2.4\times$), not to defend the criterion against it; a screening condition stated against an adaptive adversary is what the framework most lacks.

Evaluation methodology, which is where this paper sits. The adjacent precedent is adversarial robustness, where reported numbers proved unreliable until evaluations were required to be adaptive (Croce et al., 2021). Our point is methodological in the same way but about **identification** rather than attack strength. Wherever a screen gates candidate pairs on standalone effectiveness and then reads the composition's outcome as evidence about the *mechanism*, Proposition 7 shows that inference cannot work at *any* sample size: the counterfactual is excluded by the gate, not undersampled. **We do not claim this is how the FL literature generally proceeds, and that sentence is now a pre-registered measurement rather than a disclaimer.**

What the audit is, before any count of it. One coder, no inter-rater statistic, an arXiv-reachable frame that is preprint-heavy and 2026-heavy, and four rows known to us before the freeze: the rate below is illustrative of that frame and is not an estimate of the FL literature, which the supplement §S16 states in full and does not soften afterwards.

The audit's rubric (the supplement §S16) was frozen at a commit hash before any paper was retrieved, and it screens 622 retrieved records down to 59 included composed-defense papers, each coded against five criteria that every code must support with a verbatim quotation and a locator. Outcome-gated attribution, defined there as two or more mechanisms applied jointly to the same round (C-a), paired for their constituents' standalone effectiveness or complementary threat coverage (C-b), validated by the composed system's own adversarial outcome (C-c), carrying an actual mechanism-attribution claim (C-e), and running no identifying contrast (not C-d), holds for 6 **of** 59 **papers,** 10.2%. The rate is 5 of 59 if the audit's mechanism-distinctness rule is dropped, and the supplement §S16 reports both, because that rule is one of two amendments written after all 59 rows had been coded. **A low rate was pre-registered as meaning that the decline above is kept and evidenced rather than replaced**, and that is what we do with it.

The audit's first casualty is the sentence it was built to support, and we report that ahead of anything else it found. Fenaux et al. (2025) was our one named external instance, and it codes C-d = YES on strict grounds rather than by the rubric's benefit-of-the-doubt rule: its Table 3 holds the Type-2 removal stage fixed, varies the upstream Type-1 defense across six choices, and reads the requirement on the pairing off that variation, concluding that success "ultimately relies on selecting a Type 1 defense restrictive enough to overlap with CSFT's effective δ-range." That is the intervention Proposition 7 asks for, performed on the upstream stage with the downstream mechanism held fixed, so the instance is withdrawn and the audit we ran to support the claim is what withdraws it. **Nor is such a contrast unusual in the frame**: one is present in 46 **of** 59 papers, 78.0%, and of the 33 papers that have the outcome-gated *design*, 27 run one. The 6 are therefore papers that had the design and did not run the contrast, not papers for which the contrast was unavailable or unheard of. The secondary that would have counted against our reading, a paper with the design that makes no attributional claim at all, occurs 0 times in those 33.

What the frame does lack is narrower, and it is the measurement this paper makes: no route the audit codes varies an upstream stage and asks whether the *ordering* of two downstream arms survives it, with intervals over seeds. The 59 included rows carry four such routes, 13 absent, 4 analytic, 21 component contrast and 21 hyperparameter sweep (the supplement §S16), and the gap is a property of what those four designs do rather than a fifth code. **So this is a design-level reading of the route field and not a rate**: no criterion asks about ordering-survival or about seed intervals, re-coding for one now would amend a freeze after its counts exist, and what is checked row by row is the three closest rather than all 59. Those three are what the gap is made of. Tabassum & Wu (2026) report an ordering that flips across datasets, "robust aggregation alone outperforms STAR-FL in terms of ASR" on FMNIST while "the full framework provides greater benefits on complex datasets such as CIFAR-10," which is this paper's phenomenon observed and left as a remark; the paper reports no seed count, standard deviation or confidence interval anywhere, so there is no interval on either side of the flip. Tory et al. (2026) put both factors on the bench, a graded privacy ladder upstream and four aggregation rules downstream, and decline the cross in one sentence of their own limitations: "the restriction to coordinate-wise median aggregation limits immediate generalization to other Byzantine-robust rules like Krum or FLTrust." Xu et al. (2024) come closest to the interventional form, reporting with the filter held fixed that "the accuracy of Spar+Di is higher than that of Di," which establishes that the upstream stage helps and not whether it preserves the downstream mechanism's discrimination.

Two further failure modes appear in the frame that Proposition 7 does not cover, and both are worth naming separately from it. Kasyap et al. (2026) establish the invariance analytically instead of measuring it, deriving that since the "mask preserves distance ordering, the plaintext post-processing is identical to the classical Krum rule"; their empirical arm has the upstream contrast and the adversarial outcome in one figure and never crosses them, the plaintext condition existing only at zero adversaries. Nath & Kasyap (2026) do not measure a composed outcome at all, and say so: "We do not simulate FL with Byzantine-robust aggregation schemes explicitly, due to limited space. However, theoretical analysis proved that the statistical relationships are preserved." That is a C-c failure rather than a C-d one, and it is a different defect from the one we study. **The gap is also not one of capability**: Abacha et al. (2026) run an additive component ladder whose final step is exactly an identifying contrast, "Each component contributes incrementally to MTA, with the rescue stage providing the largest single gain," so the design that would answer the question is already ordinary practice and what is absent is the specific quantity. Composition itself is named as an open problem in the field's survey of them (Kairouz et al., 2021).

The claim we do make is structural rather than sociological, and its clearest instance is *our own* screen: Lemma 1 shows C0∧C1 *force* the downstream defense to suppress the attack alone, so the gate is constitutive of an inherited-suppression screen and not an incidental filter on it. Statements of a composition-preservation intuition in that literature are informal; what is new here is the *level hierarchy* at which the intuition fails (Definition 2) and the class of designs that can test it (§C.1, §D.2). **So what makes this a machine-learning contribution and not an application of standard causal-identification theory is a combination of four things, and the three-level form of the argument is in §2 rather than repeated here**: a structural identification boundary for a shape of evaluation, the preservation hierarchy that produces the failure, an interventional instantiation in FL that measures it, and the empirical non-implications (P4)$\not\Rightarrow$(P3) and (P4)$\not\Rightarrow$(P5) that the collider argument alone does not predict.

Attacks and defenses. Backdoor attacks (Chen et al., 2017; Bagdasaryan et al., 2020; Wang et al., 2020; Xie et al., 2020) and robust defenses (Pillutla et al., 2022; Sun et al., 2019; Blanchard et al., 2017) form the landscape. Beyond our norm-based menu, certified robustness (Xie et al., 2021) and filtering defenses (Nguyen et al., 2022; Fung et al., 2020) build stronger single mechanisms; we compose fixed defenses deterministically rather than strengthening a single mechanism, so these are complements to our object rather than baselines for it. Randomized transformations (Xie et al., 2018; Liu et al., 2018) in centralized settings are the conceptual neighbor of a *randomized-menu* question, which this paper does not ask: every claim here concerns deterministic joint application (§C). **An earlier version of this paper compared against a randomized menu and reported that deterministic composition dominates it; that comparison is not part of this submission and no claim here rests on it.** That withdrawal stands: the artifact behind it is unpaired and compute-confounded.

What replaces it is a separate, pre-registered, compute-matched test of the narrower question (the supplement §S17), whose result is reported there and is not scored against the withdrawn one.

Backdoor persistence and durability. The persistence mechanism we formalize has empirical precursors. Bagdasaryan et al. (2020) design model-scaling explicitly to embed triggers that survive gradient updates; Zhang et al. (2022) (Neurotoxin) engineer triggers to occupy "less-used" gradient directions, achieving multi-round persistence even under strong norm clipping. These works establish the phenomenon, and this paper *uses* it rather than modelling it: persistence is why a committed attack stays committed across rounds, which is what makes a single committed attack against a fixed defense pair the right setting for an identification question. **We make no formal claim about persistence.** An earlier version of this paper advanced a $(\gamma, \alpha, p_{\text{eff}})$ model of when persistence eliminates the value of defense switching, together with a heuristic boundary as $\gamma \to 1$; that material is not part of this submission, and we withdraw the claim rather than leave it asserted without the evidence that supported it. What remains is γ as a measured descriptive quantity (Appendix J) and the DBA persistence replication of the supplement §S19.3, which returns an *intermediate* verdict and establishes no boundary.

Defense composition and ensembling. Composing defenses is well-studied in adversarial ML for centralized models: adversarial training combined with input purification (Croce et al., 2021), detection-then-rejection pipelines, and ensemble adversarial training (Tramèr et al., 2018). These works implicitly require that upstream processing does not destroy downstream signals; our signal-preservation criterion formalizes this requirement explicitly with measurable margins. In FL the closest *structural* precedent is bucketing (Karimireddy et al., 2022), which restores existing robust aggregators under heterogeneous data by inserting an upstream reshuffling step ahead of them: this is the object we study, an upstream transform feeding a fixed downstream aggregator. Its target is convergence under heterogeneity rather than suppression of a committed attack, so it does not meet Proposition 7; what it shows is that the contrast our protocol demands is the natural one for this object, since establishing that an upstream stage helps at all requires comparing the same aggregator with and without it. In Byzantine-tolerant distributed systems, protocol composition (e.g., layered BFT) has a long history; the key distinction is that Byzantine agreement composes for fault tolerance among replicas, while our setting composes for poison resilience across client updates where the adversarial signal is a continuous gradient, not a binary fault. **A predictive criterion for FL composition does exist in recent work, and we withdraw the priority claim we previously made here.** Fenaux et al. (2025) categorize defenses by the deviation of adversarial updates from the mean of the updates and predict that pairing a large-deviation defense with a small-deviation one closes the gap either leaves alone. Our object is prior to theirs and different in kind: not which pairings succeed, but whether a comparison across defenses can identify *why* one did. Proposition 7 therefore applies to their criterion and to ours alike, and it constrains what validating evidence for such a criterion can establish rather than whether the criterion is correct.

G LIMITATIONS, IN FULL

(L1) Limited scope and architecture-conditional boundaries. Three attack families (model_scaling, pixel, DBA), two architecture families (CifarCNN, ResNet18; plus `simple_cnn`, a third CNN configuration and not a third family, on the replicated arm), and, at the composition level, CIFAR-10 throughout except for wave 1 of the 42-pair menu re-run on CIFAR-100 with wave 2 deliberately unrun (§S24.3); the two other arms that leave CIFAR-10, EMNIST-byclass and CIFAR-100, are single-arm intervention ladders and neither carries a composition. The collapse boundary is architecture-conditional (ResNet18 suppression at NC10/rep90: 0.865 vs. CifarCNN 0.097, a $9\times$ shift; per-seed values in the supplement §S20.2), so results at other N, K or datasets may differ. **That gap is also evidence for this paper's own thesis**: a swing of that size under a change of architecture alone is why we state the preservation conditions as mechanisms to be measured per setting rather than as architecture-independent empirical laws. **The Mode-S replication of §D.2 does not soften this, and the two are not in tension**: what holds across both architecture families there is the *dissociation*, a large decision disturbance costing no suppression, and not any suppression *level*, which is precisely the quantity this $9\times$ swing moves. A mechanism can transfer where the number it produces does not, and we claim only the former.

Cross-distribution transfer is validated across architecture, scale ($N \in \{10, 100\}$) and adversarial fraction ($f \in [0.05, 0.6]$; §S24.4(D)), and *dataset* transfer is now tested on three arms and no more: the flagship Mode-S arm on EMNIST-byclass with `simple_cnn`, where the negative replicates at $\Delta = -0.017$ (§D.2); the `coord_median`/pixel comparability cell on CIFAR-100, where the confounded-versus-controlled sign disagreement replicates, -0.213 against $+0.071$ (§D.3); and wave 1 of the 42-pair composition menu on CIFAR-100, at the coverage §S24.3 states.

For the screen itself the obstacle turned out to be sharper than an untested transfer: on EMNIST-byclass C1's precondition has no input on the committed pixel arm. C1 requires that some constituent suppress each committed attack *alone*, and under the identical composition protocol the pixel backdoor is essentially unsuppressed there: re-measured here, `coord_median` reaches 0.609 at clean accuracy 0.836

Table A14: **Every claim we have withdrawn or refuted, enumerated in one place.** Each row's full statement, the evidence that overturned it and what survives it are at the reference in the third column; this table is an index and is not itself the home of any disclosure. **It is hand-maintained.** No `results/` artifact records claim status, so unlike every number in this paper these rows have no emitter, and a later revision must update them by hand.

Claim, as we once stated it	status	stated in full at
Preserving the statistic or the decision d_2 reads preserves its attack suppression	refuted	`cos_krum`, §D.2
Measured disturbance transfers as a predictor, and orders the arms by admission	refuted	(L6), §G
$\Delta\Lambda_a$ explains when the two designs disagree	withdrawn	§D.3
The outcome-gated and the controlled design agree	refuted	5 of 7 cells, §D.3
An inherited-suppression screen can reach emergent compositions	refuted by construction	FG→RFA, §E
The two disturbance channels combine additively	indeterminate at n=20	§D.2.1
Three matched contrasts show a 19–20× gap turning on disturbed invariance	withdrawn; the frozen $13/20$ score stands	§S24.1
C0 belonged to the criterion before the first prospective suite scored it	withdrawn as post hoc	(L4), §G
Fenaux et al. (2025) is a named external instance of the practice we criticize	withdrawn	§F, supp. §S16
Randomizing which composition runs each round costs suppression	withdrawn as unpaired and compute-confounded	§F, supp. §S17
Consensus-shift is ineffective against FG→RFA, as measured	figure void, withdrawn; conclusion re-established by re-running the arm	§E

(seeds $0.717/0.493/0.618$), and over the seven single-defense pixel cells measured there that 0.609 is the *minimum*, with `krum` next at 0.660 and the other five between 0.988 and 0.998 (six *distinct* aggregators, since `multi_krum` reduces to `fedavg` at K=5). **So no pair drawn from them can satisfy C1 there**: strategy B certifies nothing over them, every such pair is predicted HIGH, and the strategy comparison of §6 cannot be made on that dataset from these inputs. **The coverage is exactly those cells, and we state the gap rather than generalize past it**: `foolsgold`, `reputation` and the cosine mirrors have no EMNIST-byclass pixel cell, so what is established is C1's failure over the measured menu and *not* its unsatisfiability over the whole one. **The failure is the pixel arm's, and we do not escape it by dropping the other.** On the model-scaling arm `krum` does suppress there (0.027 at clean accuracy 0.767), so a scaling-only protocol would supply a C1 input and manufacture the comparison. We decline it for the reason §S24.1 declines the mirror move: the committed protocol requires *each* attack, and choosing the arm that happens to have an input after seeing which does is attack-shopping regardless of the resulting labels. **On any of these readings it is a scope limit on the criterion's precondition and not a failed replication**: no prediction was made and refuted, and the question a screen answers does not arise where nothing suppresses the attack under test. We report it rather than lower f, subset the classes or lengthen training until a positive appeared, each of which is a new design axis chosen after seeing the obstacle. The `coord_median` figure is *bit-identical* to the corresponding cell of `results/femnist/payoff_results.json` ($0.6091787658\ldots$ at $0.8363207047\ldots$), which by the standard we apply to ourselves in §D.2 (the same computation run twice cannot corroborate itself) establishes protocol *identity* and not corroboration; what it does buy is that the matrix's other cells are protocol-comparable, so we did not spend 21 h re-measuring them (`experiments/measure_femnist_c1_inputs.py`, `results/femnist_c1_inputs/summary.json`). **We then screened the scaling arm for a further comparability cell and report the screen rather than a cell** (supplement §S14): of the five defenses with no scaling cell there, `fltrust` is excluded by argument (its admission is bit-identical under this rescaling) and we measured the other four, of which `foolsgold` and `cos_reputation` are the only two that would add a downstream defense and both fail the *accuracy* floor, $10\times$ rescaling collapsing the model to 0.046 and 0.100 clean accuracy, so **no defense outside Table A10's four is admissible on that axis**. `reputation` does meet the precondition (0.3465 at 0.704), but on seeds $0.9947/0.0152/0.0296$, and a replication resting on a spread that wide is one we decline to run.

Dataset transfer is therefore tested for one composition suite and no more, on wave 1 with wave 2 deliberately unrun (§S24.3), the two other arms that leave CIFAR-10 being single-arm ladders: the preservation conditions are dataset-agnostic in mechanism (cosine similarity, geometric distance) while the downstream decision boundary may shift with output dimensionality. *The* 2×2 *channel factorial is narrower still*: it is defined only for the two selectors whose decision can be transplanted (`krum`, `cos_krum`) and the harness raises for any other d_2, so it is one attack on one aggregator family, reported at $n{=}5$ as **not decisive** (§D.2.1) and acted on only where it withdraws a claim. *Two further gaps in the intervention*: neither DBA nor Neurotoxin was run through it, and the score-only control is a single arm, one aggregator, one attack. *And scale is a prediction, not a measurement*: the *pre-registered* PASS predictions hold at $N{=}100$/ResNet18 (FG→RFA 0.069) but far from converged accuracy, so they bound nothing about converged training, and **the FAIL controls run at that scale proved uninformative**, both landing below 0.5 (§S4). *What transfers:* the qualitative PASS/FAIL prediction and the mechanism. *What does not:* the exact suppression ASR. A practitioner at $N{=}10^4$ should use the criterion to select candidates and then pilot at their target scale: it identifies *what to try*, and architecture-specific calibration determines *how well it works*.

(L2) The composability criterion is an evaluation-prioritization rule, not a security screen and not a discovery tool. It has no established predictive validity: it orders candidates for evaluation, and the pooled model that would let it predict is unconfirmed ($R^2{=}0.239$) with its cross-arm ordering refuted (§D.2). The criterion gives *sufficient* conditions (C1∧C2, with C3 as a margin diagnostic): it reliably identifies mechanism-preserving compositions in the common case where one defense is known-effective, but it cannot certify synergistic compositions where no single defense suppresses the attack alone. The conservative bias is by design: false negatives are operationally cheap to recover from (try the composition anyway) while false positives would deploy an insecure stack. Accuracy is 16/18 $= 89\%$ on the development set, 22/24 $= 92\%$ on the pre-registered held-out set (90.5% combined 2-way) and 9/10 $= 90\%$ on pre-registered 3-way triples (supplement §S5). **None of those tallies should be read as discrimination**: only 5 of the 42 two-way pairs are low-ASR, so a constant HIGH call scores $37/42 = 88.1\%$ and the combined 90.5% beats it by one pair. Precision $(2/2)$ and recall (40%) are the quantities that carry information, and they are why we present the criterion as a prioritizer. All clean C1∧C2∧C3 PASS predictions are correct; the four misses are either the predicted intermediate behavior of rank-based aggregators (Proposition 10: dev-set NC→cm, rfa→cm; held-out fedavg→cm) or one speculative LOW call the criterion does not license (held-out fg→NC, a C1 failure). Formalization is partial: Theorem 9 proves C2∧C3 for the weighting→rank/geometric class with explicit margins, the remaining families are empirically validated, and C3's threshold condition depends on d_2's discrimination boundary, which in practice requires calibration runs.

(L3) Per-round aggregator switching only. We test per-round defense selection. Randomized *internals* (DP noise, subsampling smoothing à la CRFL, random participation masks) modify the embedding mechanism directly and may retain value even under persistence. Three adaptive attacks are tested against composition: projection-based against rep+tm (ASR saturating at 0.639 for $\epsilon \geq 2$, 0.597 at $\epsilon{=}1$; the same attack against fg→rfa reaches only 0.079, Table A12), Neurotoxin-style (Zhang et al., 2022) top-k gradient projection, and a two-phase consensus-shifting attack against rep+tm that biases updates toward the target at rate r for 25 rounds before injecting full backdoor updates (the supplement §S19). At shift_rate $\in \{0.01, 0.05\}$: mean ASR 0.684 ± 0.138 ($n{=}10$) and 0.667 ± 0.137 ($n{=}8$), comparable to committed_pixel (0.679) and *below* the rep30/tm70 temporal mix's 0.915. **Those last two are not the same metric, and the comparison is the weaker one because of it:** 0.679 and the two shift-rate means are single committed attacks against the composition, whereas 0.915 is the mix's full-information reading over both attacks, which no single attack can be expected to match. What the row supports is that the two-phase attack does not exceed the mix, not a calibrated gap between them. At $f{=}0.2$ the honest 80% anchor the coordinate-wise median, limiting Phase-1 drift to $\sim r/(1{-}f)$ per round, so Phase 2 degenerates to standard committed_pixel: a mechanism that weakens monotonically with f, and at $r \leq 0.05$ over $T \geq 200$ rounds cumulative drift may eventually overcome it. The robustness claim is therefore specific to the tested regime ($f{=}0.2$, $T{=}50$).

(L4) The falsifiable prospective positive evidence is one cell, and C0 was added retrospectively. This is the sharpest limitation we can state about our own design, and we state it rather than let a reader find it. Until supplement §S23 there was none at all, and one cell is what changed. Of the prospective suite's 8 predicted-LOW pairs (§S24.1), 6 are ·→FLTrust near-*identities*: FLTrust gates on a scale-invariant cosine and renormalizes every update to the server-gradient norm, discarding exactly the magnitude information an upstream positive rescaling modifies (verified bit-identical per seed for pass-through and non-clipping d_1), so Proposition 11(a) permits no other outcome and these confirmations cannot falsify it; the other 2 have FLTrust as d_1, where suppression is inherited and C0 puts them out of scope. C0 was itself formulated *after* seeing this suite, so its 7 exclusions are retrospective. The three matched contrasts that appeared to probe Proposition 11 most directly (a 19–20× ASR gap at fixed d_1) are **withdrawn**: two ran against MultiKrum, which at $K{=}5$ selects all five participants and is bit-identical to FedAvg, and the genuine selector leaves a gap that measures C1 failure rather than disturbed invariance (§S2).

The decisive experiment does not exist, and that is structural. We had named it as a downstream defense with a rescaling-invariant admission signal that is *weak in isolation*; Lemma 1 shows no such pair can be

certified in scope. §S24.2 runs the achievable substitute, holding the suppression *principle* fixed while varying only the invariance class, and it did not deliver a positive test either, standalone effectiveness being unholdable here because the magnitude information the mirrors discard is what the originals use to suppress a magnitude attack ($8/8$ frozen labels, TP $= 0$). So the strongest C2-specific evidence is the *post-hoc* within-defense contrast of §S24.2, which holds C1 exactly fixed but was not pre-registered and is resolved on only one of its two cells at n=3.

Three pre-registered attempts followed, and each refuted something of ours. §D delivers exactly one falsifiable prospective positive (arm 2, reputation under model-scaling, $0.017 \rightarrow 0.779$ at $p = 1.7 \times 10^{-6}$ with clean accuracy nearly unchanged) while refuting two of its own four predictions and its pooled prediction, leaving the *destruction* half of C2's mechanism confirmed and the preservation half's control indeterminate (§D.1). The targeted intervention of §D.2 closes the attenuation channel by construction and contrasts C2 within a single defense at a fixed adversarial coefficient share: substantial disturbance of the statistic does not change suppression *on the adjudicating Krum arm under Mode S*, with admission the channel the evidence points to there, and it *refutes* the preservation-half reading, the exactly-invariant `cos_krum` neither suppressing nor holding ASR fixed. The third was designed to be the falsifiable prospective *positive*: a replication arm chosen because the statistic and admission readings order its rise *oppositely*, frozen in git before the arm ran. It replicated the negative in a different invariance class and **refuted the positive ordering**, the arm with the largest measured admission change rising *less* ($+0.098$ at n=5) than one whose admission change was smaller.

The census of supplement §S23 is the one place this changed, and it changed by exactly one cell. We closed the 3×3 region where Prop. 11(a) settles C2 algebraically, registering its two unspent cells at `96f8856`; the single certified pair was registered LOW with *low* confidence and an explicit `uncertain` flag, and measured LOW at 0.263. Unlike every predicted-LOW pair above it is neither a near-identity (0 of 3 seeds identical to d_2 alone, on both arms) nor an inherited suppression, since `reputation` holds only model-scaling and `cos_krum` only the pixel backdoor. **That is a falsifiable prospective positive, and it is one cell at n=3 with a per-seed spread of 0.000 to 0.681**: it establishes that the criterion's precision survives a blind test in the one region our invariance result decides exactly, and nothing quantitative. A validated *positive* predictor therefore remains absent: the pooled two-channel model is unconfirmed over 26 cells ($R^2 = 0.239$) and its cross-arm ordering refuted outright, so admission is *more causally proximal* than the statistic without being a quantitative predictor.

(L5) Theorem 9's conditions need a pilot, and (1)'s is not operationally checkable as stated. Neither $\rho < r^\star$ nor $\Delta'_{\text{sep}} > 2(R'_B + \delta')$ is readable off a (d_1, d_2) definition pair: both need a pilot measurement. **So two kinds of cheapness have to be kept apart, and the paper has previously run them together.** Prop. 11's classes are *algebraically* checkable: (a) and (c) are decided from the definitions of d_1 and d_2 alone, with no data, no runtime cost and no per-deployment work. Theorem 9's conditions are closed-form inequalities and are *still* none of those things, because their terms have to be instantiated by measurement on the deployment they are being checked for. **Algebraically checkable does not imply deployment-time cheap**, and only the first of the two is something this paper delivers. Regime A is worse than that. Its ordering-survival condition is $\rho < r^\star$ with $r^\star$ the attack's coordinate extremeness against the benign scale. Building the dose–response ladder (§D) required us to *measure* $r^\star$ rather than stipulate it, and it is not a stable per-attack constant: on model-scaling it ranges 0.00–12.59 across the 12 measured rounds that sampled an adversary (median 0.57), on the pixel backdoor 0.61–1.63 (median 0.92). An illustrative $r^\star$=10 is a stipulation, not a measurement, and §C.3 stipulates no value for $r^\star$ at all: the theorem states the condition symbolically. Two consequences we state rather than work around. First, a practitioner cannot check $\rho < r^\star$ from a single round: the quantity swings by orders of magnitude round to round, and on both committed attacks its median is below 1, so Regime A's condition fails even at the identity rung under this estimator. Second, we therefore built *no* prediction in §D on $r^\star$, resting the ladder on directly measured disturbance of d_2's statistic instead. The margin form of the theorem remains valid as a statement about ρ and $r^\star$; what is not established is that $r^\star$ admits a stable estimator at the round granularity a deployment would have to check it at, and a usable Regime A needs an $r^\star$ estimator with characterized variance, which we do not provide. **And where the geometric condition *is* measurable it stops being informative in the regime that matters most**: the margin it rests on is exactly what heterogeneity closes, RFA's all-round mean falling from 3.24 to 1.01 at Dirichlet α=0.1, the limit at which Corollary 4 degrades to $n_a/(n_b{+}n_a)$ and certifies nothing (§C.3, §S24.4(E)). So under extreme heterogeneity the conditions are not merely awkward to check but empty, which is the sense in which §3 calls the bound vacuous there.

(L6) Measured disturbance is not a transferable predictor, and our dose instrument varies the wrong quantity for three of its four arms. The pre-registered ladder (§D.1) refutes the pooled claim it was built to test ($\rho_s = +0.351$, $p = 0.091$) and refutes two of four per-arm shapes, with ASR *falling* as the dose rises in three arms. Two distinct faults, both ours. First, the abscissa: at matched measured disturbance the real upstream transforms sit up to $+0.79$ ASR above the synthetic curve, and Krum's within-margin arm does not reproduce the very contrast (§S24.2) we attributed to disturbance of its statistic, so the rate at which a transform flips a defense's decision is not a sufficient statistic for what the transform does: *which* clients are reweighted matters, and ρ does not measure that. We now measure the quantity that does, Δ_c (the supplement §S11): the transforms carrying the two anchors that miss are targeted far beyond any ladder rung (-1.21, -0.72

against the ladder's -0.34) at ρ that is 3.7 and $13\times$ smaller. That names the axis of the mismatch without supplying a replacement predictor: Δ_c's magnitude does not order suppression either. Second, the instrument: mean-1 normalization fixes the aggregate's scale but not any client's relative weight, so with the dose assigned independently of adversary status a large ρ mostly dilutes whichever client carries the poison (measured: the ladder's Δ_c drifts monotonically negative with the dose, $0.000 \to -0.140 \to -0.240 \to -0.336$), and three arms end up measuring Theorem 9's attenuation regime rather than C2 disturbance. The adversary-targeted dose that separates the two, recorded as not run before the freeze (caveat (iii), the supplement §S7), is run in §D.2, and in retrospect Mode S is the Δ_c**-pinned** instrument (Δ_c exactly 0 at every rung, ρ up to 54.6) and Mode A the Δ_c**-swept** one (-1.14 to $+2.47$): with every adversary's weight pinned, Krum's statistic is disturbed in up to 80% of rounds while the admission of adversarial input does not move and ASR does not rise past the pre-registered practical-equivalence margin, whose lower leg ASR's own floor satisfies here so that only the upper leg is a test (the supplement §S18), so *which* clients are reweighted, not how often a decision flips, is the distinction the two readings turn on in this arm, and the ladder confound is *explained* (reputation's rise shrinks from $+0.762$ to $+0.178$ once the adversarial share is held fixed). That identifies the intervention's effect within a single defense at a fixed adversarial coefficient share, the magnitude channel left open and closed separately by the score-only control; it does not rescue the pooled positive direction ($R^2 = 0.239$ over 26 cells, cross-arm ordering refuted by the replication).

H EVIDENCE STATUS: WHAT IS PROVED, WHAT IS SUPPORTED, WHAT IS NOT

Table A15 is the ledger; the rest of this section is each row's provenance.

Proved: the invariance classes of Proposition 11; sufficient conditions for mechanism preservation on the weighting→aggregator class (Theorem 9, Lemma 2, Corollary 4); the identification boundary for outcome-gated screening (Proposition 7, with Lemma 1 its FL instance and Corollary 1 the within-defense design it licenses); the recall bound on inherited-suppression screens (Proposition 8, whose cap reaches 0 on a menu whose suppression is emergent throughout, Corollary 2); and, formalizing the interventions, statistic-level non-identifiability (Proposition 12) and the independent controllability of ρ and Δ_c (Proposition 13). **The two boundary results are general** (they are statements about the shape of a screen, not about FL), so what carries beyond this paper is the design they license, not the screen. **Empirically supported: that choosing between the two admissible designs sets the inferred direction and not merely its precision**: on `coord_median` under the pixel backdoor at n=20 the outcome-gated design moves -0.273 $[-0.334, -0.213]$ where the within-defense design moves $+0.125$ $[+0.095, +0.155]$, disjoint and both excluding zero, with Λ_a moving in both designs on that cell (0.081 against 0.033, both n=5) so that no p-value attaches to the difference between them, and the reversal replicates on CIFAR-100 at n=5, where the $\Delta\Lambda_a$ entry is absent rather than 0.0000 (§D.2, Table A10), which holds with the aggregate-magnitude channel open and is refuted with it closed, below; that substantial disturbance of the downstream statistic need not produce a corresponding change in suppression when the adversarial coefficient share is held fixed, with adversarial admission the channel the evidence favours, *on the adjudicating Krum arm under Mode S* and, by pre-registered replication, on a downstream defense in a *different* invariance class, on a *second dataset and a differently configured CNN*, and on a *second architecture family* with architecture varying alone (§D.2); that the *size* of that effect is not constant across cells, -0.010 on the adjudicating arm against -0.2857 on a second Krum cell, one where both margin legs bind and baseline admission sits off its floor rather than on it, an arm measuring ASR alone, so that neither cell generalizes to the other and no statement of the form *the dissociation holds generally* is licensed by it (App. D.9); that closing the aggregate-magnitude channel on that same adjudicating arm leaves the effect flat at the headline seed count, -0.0088 $[-0.0335, +0.0160]$ at n=20, inside the frozen ± 0.15 margin with an interval that does not exclude zero, so the reading there does not rest on benign magnitudes being free to move (App. D.2); that the screen behaves as a conservative prioritizer across 42 pairs; and that in the one region where Prop. 11(a) settles preservation *exactly*, a closed 9-cell census whose two unspent cells were registered blind contains a single low-ASR composition and it is the one the criterion certifies, at precision $1/1$ and recall $1/1$ for 10 runs against statistic preservation's 11% precision over 90 (supplement §S23, one cell at n=3, gated reading). Those two 100%s are contingent on the frozen seed count twice over, and disclosure *(v)* of supplement §S23 states the second contingency: the C1 input that certifies the pair reads 0.300 over seeds 42–44 and climbs with the seed count, to 0.490 at n=5 (still below the threshold, by 0.010) and to 0.584 at n=8, where C1 fails and the criterion certifies nothing in that census. **Not established:** the positive predictive direction of the criterion, since the pooled two-channel model is unconfirmed over 26 cells (R^2=0.239) and the replication refutes its cross-arm admission ordering, so admission is a direction and not a quantitative predictor; any ability to discover emergent synergy; and generalization beyond star-topology FL, or beyond CIFAR-10 for anything but the two replicated arms, EMNIST-byclass and CIFAR-100, neither of which carries a composition suite, and wave 1 of the 42-pair menu re-run on CIFAR-100 with wave 2 deliberately unrun (§S24.3); on the EMNIST arm no single defense measured suppresses the committed pixel backdoor alone, so C1's per-attack precondition fails over the menu measured there and the strategy comparison is out of scope there rather than merely untested ((L1), which also states what the seven do *not* cover). **Refuted by us, on the cell the reversal headlines:** the same magnitude control in coordinate-wise form, on `coord_median` under

Table A15: **The evidence ledger in one object.** One row per claim class: the tier of evidence behind it, the seed count where a seed count applies, and its status. Nothing in this table is new; every row is stated with its provenance in this section's provenance paragraph, in Table A1, or in Table A8. **The n column is not uniform and its rows are never pooled**: a row is supported at its own n and tier and at no other, which is why the adjudicating arm and its three replications are listed separately rather than as one finding at $n{=}16$. A dash means the claim is algebraic, or that no single seed count stands behind it.

Claim	Tier of evidence	n	Status
Invariance classes (Prop. 11)	algebraic	—	proved
Mechanism preservation, weighting→aggregator (Thm. 9, Lem. 2, Cor. 4)	algebraic; conditions need a pilot (L5)	—	proved
Identification boundary for outcome-gated screening (Prop. 7, Lem. 1, Cor. 1)	algebraic, not FL-specific	—	proved
Recall bound on inherited-suppression screens (Prop. 8, Cor. 2)	algebraic, not FL-specific	—	proved
Empirical non-identifiability of preservation levels, and independent control of ρ and Δ_c (Prop. 12, Prop. 13)	algebraic, on the two instruments as built	—	proved
Choosing between the two admissible designs sets the inferred direction, not the precision (`coord_median`/pixel, Table A10)	pre-registered endpoint top-up; intervals disjoint, both excluding zero; Λ_a moves in both, so no p attaches to the difference	20	supported with the aggregate-magnitude channel open, on one cell; the reversal replicates on CIFAR-100 at $n{=}5$; refuted with that channel closed (App. D.10)
Disturbance without a matching change in suppression, adjudicating Krum arm under Mode S	prospective, one arm, one attack	5	supported
The same, in a different invariance class (coord. median)	pre-registered replication	5	supported
The same, second dataset and a differently configured CNN (EMNIST-byclass)	pre-registered; channels measured before the ASR existed	3	supported
The same, architecture varying alone (`resnet18`)	pre-registered; channels measured before the ASR existed	3	supported
Admission is the more causally proximal channel of the four measured	four Mode-S arms, direction only	5, 5, 3, 3	supported as a direction
Mode S's ASR effect is not constant across cells: -0.010 on the adjudicating arm, -0.2857 on a second Krum cell (App. D.9)	pre-registered, two endpoint rungs; ASR only, no decision or admission quantity measured there	20	supported; neither cell generalizes to the other
The screen behaves as a conservative prioritizer	retrospective, 42 pairs	—	supported
Criterion precision and recall in the region Prop. 11(a) settles exactly	closed 9-cell census, two cells registered blind, gated reading	3	supported, and contingent on the seed count
Closing the aggregate-magnitude channel leaves the adjudicating arm's effect flat at the headline seed count (score-only Krum, App. D.2)	pre-registered top-up, one endpoint rung; $\kappa{=}0$ imported under the score-only identity	20	supported: -0.0088, 95% CI $[-0.0335, +0.0160]$, inside ± 0.15, not excluding zero
The criterion's positive predictive direction; the pooled two-channel model	retrospective, 26 cells, $R^2{=}0.239$	—	not established
Disturbance without a matching change in suppression, *without* adversary identity (App. D.8)	pre-registered, two endpoint rungs, adversarial coefficient share held to 10^{-7}	20	refuted, by us: $+0.3218$, 95% CI $[+0.1680, +0.4757]$
Cross-arm ordering by admission	pre-registered replication ($\Delta{=}+0.098$)	5	refuted, by us
The two designs' sign disagreement is not an artifact of the aggregate-magnitude channel (`coord_median`/pixel, App. D.10)	pre-registered, one endpoint rung, coordinate-wise magnitude closed; no margin test	20	refuted, by us: -0.1722, 95% CI $[-0.2119, -0.1325]$
Discovery of emergent synergy	bounded above by Prop. 8	—	not established
Generalization beyond star-topology FL, or beyond CIFAR-10 at the composition level	one composition suite now runs off CIFAR-10, on wave 1 with wave 2 deliberately unrun (supp. §S24.3); an earlier attempt failed its own premise (App. D.6)	—	not established
The oracle is load-bearing for that negative rather than a laboratory convenience (§D.7, §D.8)	two oracle-free arms on one cell; one refutes a flatness prediction, the other the negative itself	20, 20	supported, and a limitation rather than a result
Estimating admitted adversarial contribution without adversary identity	every (P4) reading here uses an oracle instrument	—	open

the committed pixel backdoor, takes the within-defense leg from $+0.125$ to -0.1722 $[-0.2119, -0.1325]$ at $n{=}20$, so with that channel closed both designs move the same way and the design disagreement on that cell does not reproduce under the control; the two aggregators disagree about the channel itself, inert on the selector and decisive on the order statistic, and neither freeze licenses the transfer in either direction (App. D.10). **The open problem**, which the sections above refer to by pointing here: every reading of (P4) in this paper comes from an instrument that knows which clients are adversarial, and no deployed defense knows that, so whether

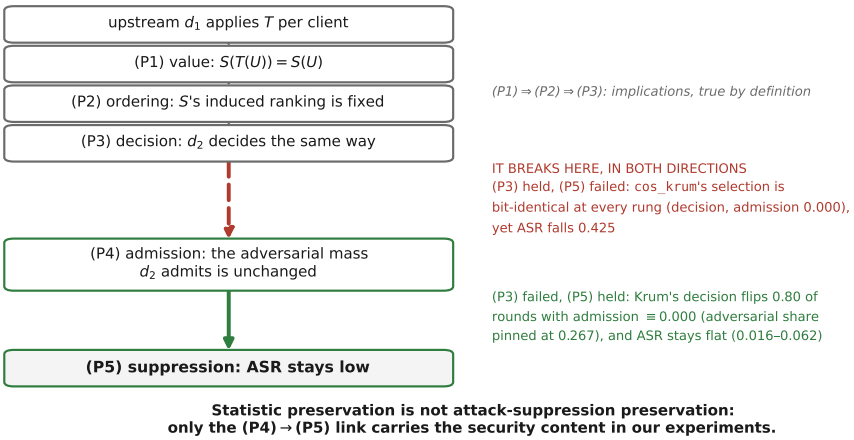

Figure A7: **The preservation chain, and where it breaks.** (P1)⇒(P2)⇒(P3) holds by definition; **below (P3) the chain breaks, in both directions**, and both witnesses come from the oracle intervention that pins every adversary's coefficient and so closes Lemma 2's attenuation channel *by construction* (§D.2). The pooled statistic-level prediction we had frozen is separately **rejected** by a pre-registered ladder ($\rho_s = +0.351$, one-sided $p = 0.091$, 16 cells; §D), and a pre-registered replication in a different invariance class reproduces the negative but **refutes the cross-arm admission ordering** ($\Delta = +0.098$ at n=5; the pooled two-channel model stays unconfirmed, $R^2 = 0.239$ over 26 cells), so **admission is more causally proximal than (P1)–(P3) here, and not a validated predictor**. Every number in the figure and in this caption is recomputed from the scoring scripts' own JSON and printed by `story_chain_figure.py`.

the adversarial contribution a downstream mechanism admits can be estimated *without* adversary identity is open. Nothing here bounds it, and the design we license identifies that channel rather than screening on it.

I THE PRESERVATION CHAIN IN ONE FIGURE

The lower arc, and why the last two links must be broken by intervening rather than comparing. T can reach ASR by attenuating whichever client carries the poison (Lem. 2), **without passing through d_2's statistic at all**. Mode S cuts that arc by construction, pinning every adversary at c=1. **The three numbers annotating Figure 1(a), which Figure A7 redraws without panels, are three scopes, not one experiment**: Krum's flip rate; the admission support *pooled* over Table 2's four CIFAR-10 arms, so it names no arm and Krum's own zero is a floor; and `cos_krum`/pixel at its first gate-passing rung (§6).

One row label in Figure 1(b) is frozen as drawn and is no longer current. Its `coord_median`/scaling row reads n=5, the count that row was frozen at, where Table A10 now reports the pre-registered endpoint top-up at 20/20. The drawing is left as it was rather than redrawn, so the figure and the table are read together and neither is taken for the other; §S13 states the same staleness for `results/margin_sensitivity.json`, which is a different artifact and is also left byte-untouched.

J NOTATION

symbol	meaning
c_i	upstream coefficient applied to client i's update, $T(u_i) = c_i u_i$, $c_i > 0$
A	the set of adversarial participants in a round
ρ	statistic-level dispersion, $\max_i c_i / \min_i c_i$: *how far apart* the coefficients are
Δ_c	adversary-correlated dispersion, $\mathbb{E}[c_i \mid i \in A] - \mathbb{E}[c_i \mid i \notin A]$: *who* receives them
r	adversary-to-benign weight ratio, $w_a / \max_b w_b$; the Mode-A dial, $r = e^{\nu}$
$r^\star$	the attack's coordinate extremeness, $\min_k \min_a \min_b \lvert u_{a,k}\rvert / \lvert u_{b,k}\rvert$ (Theorem 9(1))
S	the downstream statistic d_2 reads, as in $S(T(U))$ (§C.4)
m	RFA separation margin, $\Delta'_{\text{sep}}/2(R'_B+\delta')$
Λ_a	aggregate adversarial influence, instantiated per aggregator: Weiszfeld/reputation weight share, selected-coordinate share, or the selection indicator
κ	the Mode-S / ladder dispersion dial, $\rho = e^{2\kappa}$
γ	attack *persistence*, $\text{ASR}(t{+}1)/\text{ASR}(t)$ after a clean round (≈ 0.66); never a weight