BurstGuard: Defending Federated Mobile-Traffic Forecasting against Burst-Triggered Backdoors under Secure Aggregation

Online Resource 2

Mobile Networks and Applications

Vinh Thinh Le[1] and Tran Thanh Lam Nguyen[1*]

[1]Faculty of Information Technology, HCMC University of Technology and Engineering, No. 1 Vo Van Ngan Street, Thu Duc Ward, Ho Chi Minh City, Vietnam.

*Corresponding author(s). E-mail(s): lamntt@hcmute.edu.vn;
Contributing authors: thinhlv@hcmute.edu.vn;

Abstract

This Online Resource accompanies the article *BurstGuard: Defending Federated Mobile-Traffic Forecasting against Burst-Triggered Backdoors under Secure Aggregation*. It gives the detailed results of the five research questions of the article (Sections S1–S5): every table of the evaluation in full, per-round traces, score margins, calibration headroom, the backdoor lifespan, grouped secure aggregation, traffic types, probe shifts, norms, calibration and ablation replays, and the system cost. Numbers with the prefix S refer to this document; numbers without it refer to the article. Online Resource 1 gives the method details, notation, setup, implementation, and grids.

S1 RQ1: Is the backdoor a threat to worthwhile training?

This section details RQ1 of the article (Section 5.1).

S1.1 Is federated fine-tuning worth protecting?

Purpose. A defence is useful only if the training it protects is useful. RQ1 asks whether federated fine-tuning improves on the pre-trained model and on training alone, so that operators have a reason to federate.

Method. Grids E1, X1, and R18 compare zero-shot TTM-R2 with local-only fine-tuning (each site alone), centralised fine-tuning (all training windows pooled), and federated fine-tuning in the four modes, on both regions and three traffic types, on the test and holiday splits.

Results. Table S1 shows that every federated mode beats the zero-shot model on Milan internet traffic (MASE 0.637–0.655 vs. 0.687), at 18 of 20 sites (Wilcoxon over sites, Holm-adjusted $p \leq 5.2 \times 10^{-4}$), and federated head and LoRA beat it at all SMS and call sites. Federated LoRA is the most accurate mode (-7%, -12%, and -22% on internet, SMS, and calls) and matches centralised LoRA ($p = 0.73$). The holiday split separates collaboration from local

Table S1 Clean MASE (lower is better) of zero-shot, local-only, centralised, and federated (FL) fine-tuning (grids E1, X1, R18; two to three runs per Milan cell, one to two per Trentino cell). Test split unless marked *hol.* (holiday).

	Milan				Trentino	
Setting	Internet	hol.	SMS	Calls	Internet	hol.
Zero-shot	0.687	0.700	0.710	0.643	0.735	0.678
Local, head	0.658	0.717	0.685	0.556	0.715	0.692
Local, LoRA	0.647	0.725	–	–	–	–
Central, head	0.651	0.679	0.670	0.554	–	–
Central, LoRA	0.638	0.706	–	–	–	–
FL, head	0.655	0.681	0.670	0.558	0.708	0.668
FL, decoder	0.645	0.676	–	–	–	–
FL, LoRA	**0.637**	**0.669**	**0.624**	**0.500**	**0.682**	**0.660**
FL, full	0.647	0.674	–	–	–	–

Table S2 Attacks under FedAvg (grid E2; three runs for model replacement, two for the other attacks, one for decoder and full): $ASR_{0.5}$, UPR_{trig} with 95% CI, and relative change of clean MASE ($UPR_{trig} \approx 0.008$ without attack).

Mode	Attack	$ASR_{0.5}$	UPR_{trig} [CI]	MASE
Head	Poisoning, γ=1	0.00	0.10 [.06,.15]	+4.2%
	Norm-constr.	0.00	0.01	−0.4%
	Moderate, γ=3	0.03	0.36 [.29,.45]	+17.7%
	Boosted, γ=5	0.06	0.51 [.44,.59]	+27.2%
	Replacement, γ=10	0.06	0.64 [.59,.69]	+42.9%
LoRA	Poisoning, γ=1	0.00	0.19 [.12,.28]	−0.2%
	Norm-constr.	0.00	0.01	−1.2%
	Moderate, γ=3	0.12	0.74 [.67,.80]	+4.7%
	Boosted, γ=5	0.19	0.86 [.82,.88]	+10.9%
	Replacement, γ=10	0.31	0.92 [.88,.95]	+22.4%
Decoder	Replacement, γ=10	0.33	0.93	+31.7%
Full	Replacement, γ=10	0.26	0.83	+44.9%

training: every federated mode stays better than zero-shot ($p \leq 4.2 \times 10^{-3}$), whereas local-only training becomes worse on average (0.717–0.725 vs. 0.700, worse at 14–15 of 20 sites, $p = 0.08$). Trentino shows the same pattern: federated head and LoRA fine-tuning reduce MASE from 0.735 to 0.708 and 0.682 (−4% and −7%), local-only training to 0.715, and on holidays only the federated models beat the zero-shot model (0.668 and 0.660 vs. 0.678; local-only 0.692). The Trentino LoRA value comes from a BurstGuard run in which no round was flagged, so it equals plain federated training.

Insight. Federation pays off most where a single site has little evidence, such as an unusual holiday week, and LoRA is both the most accurate mode and, as RQ1 shows, the most fragile one. We therefore focus on the cheapest mode (head) and the most accurate one (LoRA), and check decoder and full fine-tuning in RQ4.

S1.2 How harmful are burst-triggered backdoors?

Purpose. RQ1 measures what the attacker gains under plain FedAvg: how many triggered forecasts collapse, how visible the attack is in clean accuracy, how long the backdoor survives, and whether real traffic surges fire it.

Method. Grid E2 trains FedAvg under each attack in all four modes. In grid E5 the attacker joins only rounds 6–15 of 40. Grids R1, R17, and R18 measure UPR_{nat} on all real windows of the validation, test, and holiday splits that already look like the trigger.

Results. The boost trades strength for stealth (Table S2). Model replacement under-provisions 64–93% of the triggered forecast steps, but it also raises clean MASE by 22–45%, which an operator would notice. The moderate boost is covert on LoRA: 74% of the triggered steps are under-provisioned while clean MASE grows by 4.7%, although this already cancels about two thirds of the 7% gain of federated LoRA. Plain poisoning implants a weaker backdoor (0.10–0.19) at no visible cost, and the norm-constrained attacker, which projects its update onto its own benign norm, implants none. $ASR_{0.5}$ stays at or below 0.33 even when most forecasts are badly damaged, because the model need not hit the target exactly to under-provision; this is why UPR_{trig} is our main metric. The backdoor is also durable: in E5, 25 rounds after the attacker leaves, FedAvg's test UPR_{trig} is still 0.058 for LoRA and 0.022 for head, against 0.008 with BurstGuard (Section S1.3).

Table S3 shows that real demand fires the backdoor. Trigger-like surges are rare in the test weeks (10–14 windows per run in Milan), so we pool the validation, test, and holiday weeks, which gives 40–49 windows at 14–18 of 20 sites per run. On these real windows, a clean model under-provisions 5.4% of the steps, whereas the backdoored FedAvg models under-provision 51% (model replacement), 31% (moderate boost), and 15% (plain poisoning), with intervals that do not

Table S3 Real trigger-like windows: windows whose last half hour rose by at least $a\sigma_i$, pooled over the validation, test, and holiday splits (per run: Milan 40–49 windows at 14–18 sites; half participation 43–49 at 14–18; Trentino 69–93 at 35–41), and $\mathrm{UPR}_{\mathrm{nat}}$ on them with a bootstrap 95% CI over (run, site) units; n: runs per column. Head fine-tuning. BurstGuard: recommended configuration, except Milan (calibrated). In the test split alone, Milan has 10–14 such windows per run (0.2–0.3% of the windows).

Setting	Defence	n	No attack	γ=10	γ=3	γ=1
Milan	FedAvg	3/5/5/5	0.054 [.039,.070]	0.514 [.446,.587]	0.305 [.247,.366]	0.147 [.110,.192]
	BurstGuard	5/3/5/5	0.054 [.042,.068]	0.049 [.034,.067]	0.052 [.039,.065]	0.052 [.039,.065]
Half participation	FedAvg	1/0/1/1	0.056 [.030,.087]	–	0.499 [.330,.661]	0.244 [.137,.356]
	BurstGuard	3/2/2/2	0.054 [.038,.070]	0.046 [.028,.069]	0.046 [.027,.069]	0.046 [.028,.069]
Trentino	FedAvg	2/1/2/2	0.113 [.085,.144]	0.429 [.326,.540]	0.371 [.309,.434]	0.226 [.175,.281]
	BurstGuard	3/1/2/2	0.113 [.090,.138]	0.114 [.071,.167]	0.117 [.086,.151]	0.117 [.086,.151]

overlap the clean one. Trentino has more such windows and a higher clean level (0.113), and the backdoor raises it two- to fourfold. With BurstGuard, $\mathrm{UPR}_{\mathrm{nat}}$ stays at the clean level in every setting.

Insight. The relevant attacker leaves accuracy largely intact, the harm decays slowly, and the trigger occurs naturally, if rarely: in the running example, the backdoor would remove capacity on real busy evenings, not only on inputs crafted by the attacker. Every round must therefore be screened, and the screen must look at triggered behaviour rather than accuracy.

S1.3 Further results

Backdoor lifespan and anchor drift (RQ1, RQ5).

Fig. S1 shows that a backdoor implanted in rounds 6–15 decays slowly under FedAvg, and Table S4 shows the clean anchor-score drift that any calibration must cover: the head's anchor score settles at about 0.03, whereas LoRA's keeps growing to about 0.11 by round 40.

Table S4 Clean anchor-score drift of BurstGuard runs in grid E5 (no attack in these rounds; the attacker joined rounds 6–15 only, during which the aggregates were regrouped). S_{anc} by round.

Run	1	2	5	16	20	25	30	40
Head, seed 0	0.021	0.031	0.035	0.025	0.026	0.028	0.030	0.029
Head, seed 1	0.021	0.029	0.033	0.032	0.031	0.028	0.029	0.030
LoRA, seed 0	0.020	0.027	0.031	0.078	0.077	0.106	0.082	0.107
LoRA, seed 1	0.017	0.023	0.040	0.070	0.075	0.103	0.082	0.104

Real trigger-like windows: test split and adaptive attackers (RQ1, RQ3).

In the test split alone, Milan has 10–14 trigger-like windows per run at 5–6 sites (0.2–0.3% of the windows), and $\mathrm{UPR}_{\mathrm{nat}}$ is 0.044 [0.019, 0.074] for clean BurstGuard runs, 0.559 [0.422, 0.697], 0.321 [0.207, 0.448], and 0.157 [0.068, 0.264] under FedAvg with model replacement, the moderate boost, and plain poisoning, and 0.037–0.041 with BurstGuard. Trentino has 27–36 such test windows at 13–15 sites (clean 0.154; FedAvg 0.516, 0.451, 0.295; BurstGuard 0.165–0.167). Pooled over the validation, test, and holiday splits, the adaptive attackers of RQ3 leave the following $\mathrm{UPR}_{\mathrm{nat}}$ under BurstGuard (clean 0.054 [0.042, 0.068]): score-aware 0.045 (head) and 0.051 (LoRA); threshold riding 0.057 [0.033, 0.086] over 20 rounds, 0.071 over 60 rounds, 0.058 on LoRA, and 0.061 against a per-mode anchor threshold; oracle cap riding 0.048 [0.029, 0.072]; regroup equivocation 0.058 [0.040, 0.080] (head), 0.073 (LoRA), and 0.083 [0.037, 0.140] with verification. Without the public guard, an attacker active during calibration leaves 0.304 ($\gamma = 3$) and 0.128 (plain poisoning).

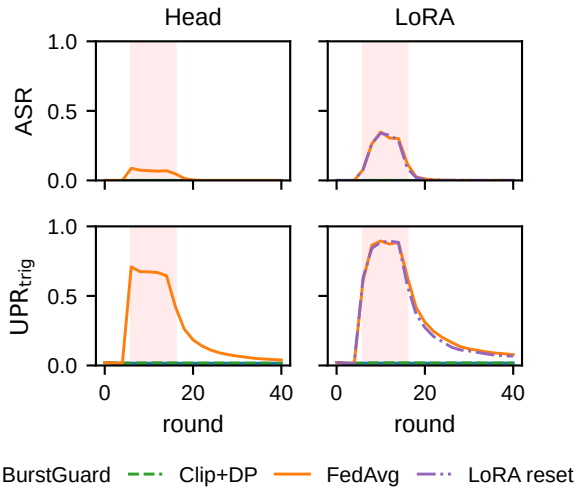

Fig. S1 Backdoor lifespan (grid E5, two seeds): model replacement in rounds 6–15 (shaded) of 40; $ASR_{0.5}$ and UPR_{trig} on the validation split. Under FedAvg, UPR_{trig} stays above the clean level long after the attack (validation: LoRA 0.08, head 0.04 at round 40; test: 0.058 and 0.022); with Clip+DP or BurstGuard no backdoor appears (test: 0.007–0.009).

S2 RQ2: Does BurstGuard stop the backdoor?

This section details RQ2 of the article (Section 5.2).

S2.1 Does BurstGuard stop the backdoor?

Purpose. RQ2 asks whether BurstGuard removes the backdoor of the three basic attacks without hurting clean accuracy and with few false alarms, what simpler statistics of the aggregate would see, and how BurstGuard compares with defences that see individual updates and with defences that, like BurstGuard, work under secure aggregation.

Method. The recommended configuration runs in R13, R14, R16, R17, and R18 (head) and the guarded LoRA variant in R11 and R18; we pool them into one ledger (Table S6) with runs, distinct trajectories, and false positives over unique attacker-free rounds. As an oracle reference, X8 runs BurstGuard with thresholds calibrated on the clean E3 runs under all attacks and on held-out clean seeds, X5 and X6 do the same for SMS, calls, and $W = 18$, and R1 repeats it with five seeds. E3 and X7 run FedAvg, the defences on individual updates, the probe-loss check, LoRA reset, and default BurstGuard; R4 runs the secure-aggregation-compatible baselines; E7 replaces the secure sum by 1–10 group sums. To see what simpler statistics of the aggregate would reveal, we use the update norms that every run logs, the aggregate norms logged by the aggregate-clipping runs (before clipping), and the aggregate norms of the public simulation, and we replay a calibrated probe-loss tolerance on the logged rounds.

Results. *Baselines.* Every defence that sees individual updates removes all three backdoors (Table S5), some at a small but significant accuracy cost (Clip+DP, FLAME, and FLTrust on LoRA; Holm-adjusted $p < 0.05$). Among the defences that need only aggregates, the best-case norm bound and RFA also remove the backdoor of two attackers; RQ3 shows what happens with more attackers. Aggregate clipping does not: clipping the *sum* to the benign mean norm still leaves $UPR_{trig} = 0.59$ under model replacement, because the malicious direction survives clipping. The probe-loss check stops model replacement only by rejecting every attacked round, which freezes training (MASE 0.687), and passes the moderate and plain attacks, as does LoRA reset. A calibrated tolerance for the probe check (1.5 times the largest clean increase, 0.46%), replayed on the logged rounds, flags head replacement and the moderate boost in round 1 and head poisoning in round 2, but never flags LoRA poisoning and flags the moderate LoRA attacker only after two poisoned aggregates were accepted (Section S2.2).

Table S5 Detailed version of Table 4 of the article. Defences against the three basic attacks (Milan internet, test split): UPR_{trig} / clean MASE under each attack. SA: needs only secure aggregates. Without attack, $UPR_{trig} \approx 0.01$. FedAvg and the update-level baselines use grids E3 and X7 (two to three runs per head cell, one to two per LoRA cell); the SA baselines use grid R4 (head; LoRA only where run); recommended BurstGuard uses R13 and R14 (head) and guarded BurstGuard R11 (LoRA); calibrated BurstGuard uses X8 with the same attacked seeds as E3. †Server-side clipping needs individual norms. ‡Best case: the bound is the largest benign norm of the round.

		Clean MASE		Replacement (γ=10)		Moderate (γ=3)		Poisoning (γ=1)	
Defence	SA	Head	LoRA	Head	LoRA	Head	LoRA	Head	LoRA
FedAvg	yes	0.655	0.638	0.64/0.935	0.93/0.811	0.37/0.772	0.73/0.667	0.10/0.685	0.19/0.636
Clip+DP [5]	no†	0.659	0.651	0.01/0.659	0.01/0.644	0.01/0.659	0.01/0.644	0.01/0.659	0.01/0.644
Multi-Krum [1]	no	0.655	–	0.01/0.654	–	0.01/0.654	–	0.01/0.654	–
Trimmed mean [6]	no	0.654	–	0.01/0.653	–	0.01/0.653	–	0.01/0.653	–
Median [6]	no	0.654	0.637	0.01/0.652	0.01/0.630	0.01/0.652	0.01/0.629	0.01/0.652	0.01/0.629
FLAME [3]	no	0.661	0.643	0.01/0.658	0.01/0.632	0.01/0.658	0.01/0.632	0.01/0.658	0.01/0.632
FLTrust [2]	no	0.667	0.647	0.01/0.663	0.01/0.641	0.01/0.665	0.01/0.641	0.01/0.662	0.01/0.641
Norm bound‡	yes	0.657	–	0.01/0.652	–	0.01/0.655	0.01/0.626	0.01/0.655	0.01/0.626
RFA [4]	yes	0.657	–	0.01/0.651	–	0.01/0.654	–	0.01/0.654	–
Aggr. clipping‡	yes	0.657	–	0.59/0.909	–	0.32/0.775	–	0.09/0.689	–
Probe-loss check	yes	0.655	0.638	0.01/0.687	0.01/0.684	0.37/0.772	0.73/0.667	0.10/0.685	0.19/0.636
LoRA reset	yes	–	0.638	–	0.93/0.802	–	0.72/0.667	–	0.18/0.635
BurstGuard (default, pilot)	yes	0.655	0.638	0.01/0.657	0.01/0.633	0.01/0.657	0.01/0.633	0.09/0.682	0.19/0.636
BurstGuard (recomm./guarded)	**yes**	**0.655**	**0.637**	**0.01/0.653**	–	**0.01/0.653**	**0.01/0.627**	**0.01/0.653**	**0.01/0.627**
BurstGuard (calibrated)	yes	0.655	0.636	0.01/0.657	0.01/0.633	0.01/0.657	0.01/0.633	0.01/0.657	0.01/0.633

Table S6 Detailed version of Table 3 of the article. Detection ledger of BurstGuard (internet traffic, $W = 6$ unless noted). Runs: attacked runs and, in brackets, distinct trajectories. Flagged: attacked aggregates flagged. Entries: attacked rounds in which a malicious update entered the model. False alarms: flagged attacker-free rounds over unique attacker-free rounds (rounds of identical trajectories counted once), with the number rejected.

Configuration	Setting (grids)	Runs	Flagged	Entries	False alarms (rejected)
Recommended, head	Milan, all sites, two attackers (R13, R14)	5 (2)	100/100	0	0/140 (0)
	Milan, ten of 20 sites per round (R17)	6 (2)	81/81	0	6/73 (3)
	Trentino (R18)	5 (2)	100/100	0	4/60 (3)
	two attackers, total	*16 (6)*	*281/281*	*0*	*10/273 = 3.7% [1.8, 6.6]% (6)*
	Milan, six attackers (R13)	3 (2)	120/120	0	–
	Milan, six norm-bounded attackers (R13)	6 (4)	212/240	88	–
	Milan, one-day offset, $W = 150$ (R14, R16)	5 (4)	93/100	10	–
Guarded, LoRA	Milan (R11) and Trentino (R18)	4 (3)	77/80	3	0/40 (0)
Calibrated (oracle)	internet, SMS, calls, $W = 18$ (X5, X6, X8)	40 (16)	800/800	1	0/240 (0)
	five head seeds (R1)	13 (5)	260/260	0	0/100 (0)
	LoRA poisoning, seeds 2–4 (R1, R15)	3 (3)	50/60	16	–
	same, with verification (R16)	3 (3)	50/60	14	–

Recommended configuration. With public caps and two attackers, BurstGuard flagged all 281 attacked aggregates of 16 runs on Milan, with half participation, and on Trentino (Table S6), and no malicious update entered the model; UPR_{trig} stayed at the clean level (0.006–0.008 on Milan, 0.025 on Trentino) and MASE at that of unattacked FedAvg (0.653 vs. 0.655 on Milan). The 16 runs follow six distinct trajectories, because every attacker's group was dropped, so the run-level evidence is 16 of 16 runs fully flagged (95% CI $[0.79, 1]$), or 6 of 6 trajectories. The price is 10 false alarms in 273 unique attacker-free rounds (3.7%, CI 1.8–6.6%), none of 140 on Milan with all sites but 6 of 73 with half participation and 4 of 60 on Trentino; six of them were rejected and four regrouped, without a measurable effect on MASE (RQ4). Guarded LoRA flagged 77 of 80 attacked aggregates without false alarms; the three entries are unflagged middle rounds of Trentino poisoning, an attack whose first-round score (0.033) is only 1.3 times the cap.

Oracle reference. With thresholds calibrated on logged clean runs of the same federation, BurstGuard flagged every attacked aggregate of the calibrated grids (800/800 in 40 runs that follow 16 distinct trajectories; run-level CI $[0.91, 1]$) and none of 240 held-out clean rounds; once (LoRA, poisoning) a group with a poisoned update passed

Table S7 Detailed version of Table 5 of the article. What simpler statistics of the aggregate see (Milan head, round 1). Update-norm ratio: malicious over mean benign client norm (median over rounds; range). Norm/cap: aggregate norm before clipping (R4) over a public norm cap of 1.5× the simulated aggregate norm (0.58–0.59, so the cap is 0.87–0.88; clean aggregate 0.57). Score/cap: S_{last} of recommended BurstGuard over its public cap (0.024–0.026).

Attack	Norm ratio	Norm/cap	Score/cap
Replacement, γ=10	37 (34–93)	10.5–10.9	38
Moderate, γ=3	15 (14–27)	3.3–3.6	9.4–10.7
Plain poisoning	7.6 (6.8–9.1)	1.26–1.39	2.8–3.3
Norm-constrained	1.0 (0.9–1.2)	–	–
Threshold riding	1.5 (0.5–2.9)	–	0.4
Six norm-bounded	$\leq$1 after clipping	$\leq$0.93	0.9–1.0

the regroup screen without raising UPR_{trig}. Five head seeds (R1) give the same picture (260/260, none of 100 clean rounds). LoRA under plain poisoning is the hardest case: across seeds 2–4, 50 of 60 attacked aggregates are flagged, the poisoned update entered in 16 rounds, mostly late rounds whose poisoned aggregates stayed below the thresholds, and UPR_{trig} rose to 0.012–0.018 (clean 0.008; FedAvg 0.15–0.27). Verification (R16) leaves this almost unchanged (0.014), because most of these rounds are never flagged. Default BurstGuard, finally, stops the boosted attacks but misses plain poisoning (2 of 60 and 0 of 40 attacked aggregates), because its pilot floors sit above the poisoning scores. Fig. S2 shows the scores behind these numbers: clean rounds never move S_{last} above 0.021–0.023, moderate-boost updates score at least 0.12, model replacement 0.44–1.25, and plain poisoning 0.048–0.081 in its first round; once accepted under the default rule, poisoning lets S_{anc} ramp to 0.30–0.47, the slow ramp of Observation 1. With group sums instead of one sum (E7), a coordinate-wise median of four group sums still leaves $\text{UPR}_{\text{trig}} = 0.26$, whereas BurstGuard keeps 0.007 for every number of groups.

What simpler statistics see. Plain poisoning applies no boost, but its updates are not benign-sized: five local epochs at three times the learning rate make them 7.6 times larger than a benign update for head and 5.8 times for LoRA (Table S7). The aggregate's norm is visible under secure aggregation, and the public simulation predicts its clean value well (0.58–0.59 against 0.57 in the first real round), so a public-calibrated norm

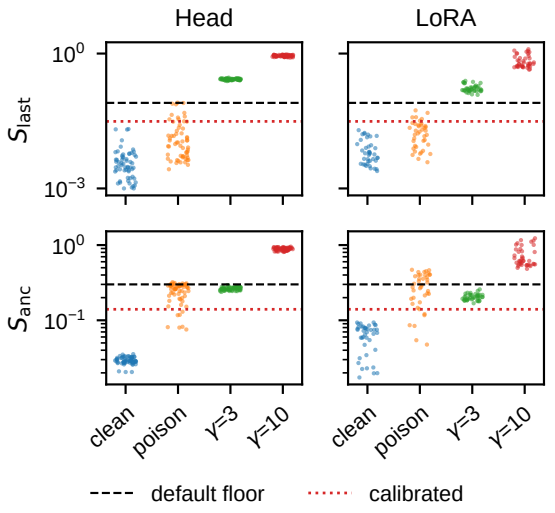

Fig. S2 Per-round scores (log scale) of clean aggregates and of aggregates containing a malicious update (grids E3 and X7, default configuration), with the default floors (dashed) and the calibrated thresholds (dotted). Poisoning scores fall over the rounds because the default rule accepted the poisoned updates.

cap would also flag the first-round aggregates of all three basic attacks with two attackers. The burst score separates them with larger margins (2.8–38 times its cap against 1.26–10.9 times for the norm cap), but the norm is free and should be used alongside it. Neither per-round statistic sees an attacker who keeps its update benign-sized: the norm-constrained attacker implants nothing, the threshold rider implants little (RQ3), and six attackers clipped by a verified norm bound leave both the aggregate's norm (at most 0.93 times the norm cap, by the triangle inequality) and the per-round score (0.9–1.0 times its cap) at clean levels in round 1. What separates the last case is the anchor score, which accumulates the small per-round changes (RQ3).

Insight. Two attackers with unbounded updates betray themselves to several aggregate statistics, the norm included, and the functional score separates them with the widest margins. The cases that matter for a functional screen are those in which every update is bounded: there, size statistics see nothing, while the triggered behaviour still accumulates against the anchor. Calibration turns separation into decisions: the

same search with pilot floors misses plain poisoning, and the public caps catch it without client data, at the price of a few false alarms where the real federation is noisier than the simulated one.

S2.2 Further results

Probe-loss check with a calibrated tolerance (RQ2).

Table S8 replays the probe-loss check with a tolerance calibrated like BurstGuard's thresholds.

Table S8 Probe-loss check with a calibrated tolerance, replayed on the logged rounds of E3 and X7: tolerance = 1.5× the largest relative probe-error increase of a clean round (0.0030), i.e. 0.46%. First attacked round whose increase exceeds it (up to that round the replay is exact, because the logged trajectory is the same). The check can only reject whole rounds.

Mode	Attack	First flag	$\mathrm{UPR}_{\mathrm{trig}}$ (5% check)
Head	Replacement	1, 1, 1	0.01 (frozen)
Head	Moderate, $\gamma = 3$	1, 1, 1	0.35–0.37
Head	Plain poisoning	2, 2, 2	0.09–0.11
LoRA	Replacement	1, 1	0.01 (frozen)
LoRA	Moderate, $\gamma = 3$	3, 3	0.68–0.77
LoRA	Plain poisoning	never, never	0.18–0.21

Update and aggregate norms (RQ2).

Every run logs the mean norm of the benign and of the malicious client updates in each round; the aggregate-clipping runs of grid R4 also log the norm of the aggregate before clipping, and the public simulation logs the norm of each simulated aggregate. Table S9 lists them. The malicious updates of all basic attacks are several times larger than benign ones, plain poisoning included, because the attacker trains five epochs at three times the learning rate; only the norm-constrained attacker matches the benign norm, and it implants no backdoor. With six attackers under the best-case norm bound, each malicious update is clipped to the largest benign norm τ of the round (1.32–1.34 in round 1). By the triangle inequality the aggregate norm is then at most $\frac{14}{20}\|\bar{\boldsymbol{\Delta}}_{\mathrm{benign}}\| + \frac{6}{20}\tau$; with the simulated aggregate norm as a proxy for the benign mean, this bound is 0.81 in round 1, below a public norm cap of 1.5×0.58–$0.59 = 0.87$–0.88. In later rounds the bound is not tight enough to decide, and the aggregate norm of these runs was not logged.

Run-level counting and replicates (RQ2).

Rounds of one run are not independent, and runs that differ only in the attack follow identical trajectories once every attacker's group is dropped (identical final MASE to six decimals). In the calibrated grids X5, X6, and X8, the 40 attacked runs follow 16 distinct trajectories (X8: 15 runs, 6 trajectories; X5: 12 and 6; X6: 13 and 4); in R1 the 13 attacked head runs follow 5; in the recommended configuration, the 16 two-attacker runs follow 6 and the six norm-bounded runs 4. Every attacked aggregate is flagged in 40 of 40 calibrated runs (exact 95% CI $[0.91, 1]$; 16 of 16 trajectories, $[0.79, 1]$) and in 16 of 16 recommended two-attacker runs ($[0.79, 1]$; 6 of 6 trajectories, $[0.54, 1]$). The 32 probe windows were drawn once with a fixed seed and are the same in all runs; the shadow screens of R1 (other probe periods, another traffic type, 8 or 16 probes) are therefore our only evidence on the probe draw.

Grouped secure aggregation (RQ2, RQ5).

Tables S10 and S11 and Fig. S3 replace the single secure sum by several group sums.

Table S9 Norms. Top: malicious over mean benign client update norm under FedAvg (grid E2; median over rounds, range). Bottom: aggregate norm before clipping in rounds 1–4 (grid R4, head, aggregate clipping) and of the public simulation (eight pseudo-clients).

Mode	Attack	Malicious/benign norm
Head	Plain poisoning	7.6 (6.8–9.1)
Head	Moderate, $\gamma = 3$	15 (14–27)
Head	Replacement	37 (34–93)
Head	Norm-constrained	1.0 (0.9–1.2)
LoRA	Plain poisoning	5.8 (5.4–6.7)
LoRA	Moderate, $\gamma = 3$	15 (14–20)
LoRA	Replacement	46 (41–67)
Decoder / full	Replacement	57 / 60
Head	Threshold riding (R3)	1.5 (0.5–2.9)
Aggregate norm, rounds 1–4		
no attack		0.57, 0.42, 0.36, 0.32
plain poisoning (three seeds)		1.10–1.20, 0.99–1.07, 0.90–0.99, 0.84–0.93
moderate, $\gamma = 3$		2.83–3.09, 2.71–2.91, 2.54–2.76, 2.41–2.65
replacement		9.1–9.6, 8.9–9.3, 8.5–8.9, 8.2–8.4
public simulation (Milan)		0.58–0.59, 0.41–0.43, 0.39, 0.35–0.37
public simulation (Trentino; LoRA)		0.61–0.62; 1.08–1.11 (round 1)

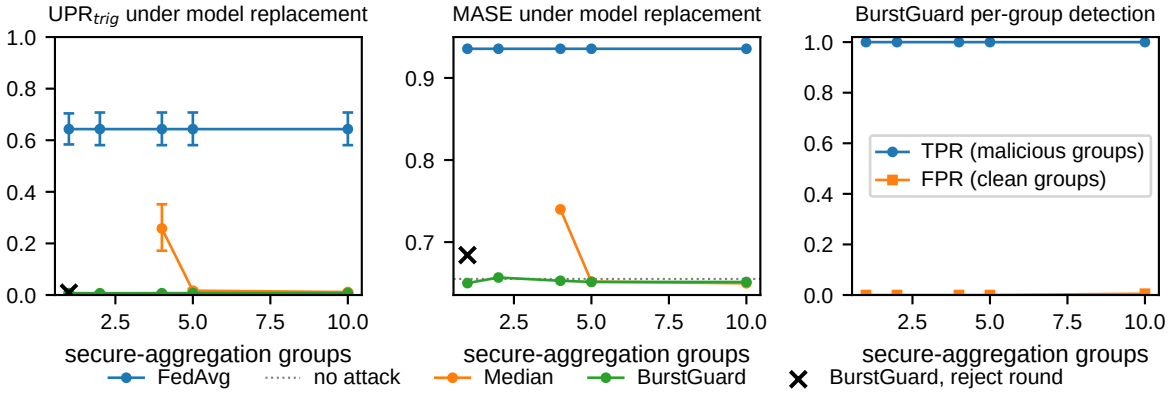

Fig. S3 Grouped secure aggregation (grid E7, head, model replacement): $\mathrm{UPR}_{\mathrm{trig}}$ and MASE under attack against the number of secure-aggregation groups, and BurstGuard's per-group detection.

Table S10 Grouped secure aggregation (grid E7, head, clean runs): largest clean score of a group sum and its ratio to the largest clean score of the full aggregate (one group), against the square root of the number of group sums.

Sums	max S_{last}	ratio	max S_{anc}	ratio	$\sqrt{\cdot}$
1	0.0209	1.00	0.0355	1.00	1.00
2	0.0305	1.46	0.0549	1.55	1.41
4	0.0467	2.23	0.0706	1.99	2.00
5	0.0549	2.62	0.0759	2.14	2.24
10	0.0821	3.92	0.1085	3.06	3.16

Table S11 Grouped secure aggregation (grid E7, head, model replacement): $\mathrm{UPR}_{\mathrm{trig}}$/MASE under attack, BurstGuard FPR on clean groups, and server CPU seconds per round of BurstGuard under attack. With one group, BurstGuard regroups every attacked round into five groups, which explains the higher CPU time than with two groups, whose sums are screened directly. Rejecting flagged rounds instead (one group) gives $\mathrm{UPR}_{\mathrm{trig}} = 0.009$ at MASE 0.684.

Groups	FedAvg	Median	BurstGuard	FPR	CPU s
1	0.64/0.936	–	0.007/0.650	0.000	44.5
2	0.64/0.936	–	0.007/0.657	0.000	14.7
4	0.64/0.936	0.26/0.740	0.007/0.653	0.000	25.6
5	0.64/0.936	0.02/0.652	0.007/0.652	0.000	28.5
10	0.64/0.936	0.01/0.650	0.007/0.651	0.005	58.9

S3 RQ3: Does BurstGuard withstand adaptive and many attackers?

This section details RQ3 of the article (Section 5.3).

S3.1 Does BurstGuard withstand adaptive attackers?

Purpose. An attacker who knows BurstGuard will try to evade it. RQ3 asks whether BurstGuard holds when the attacker hides the backdoor from the search, rides the thresholds, acts while the thresholds are being calibrated, sends a different update in a regroup, or brings more attackers.

Method. Grid E6 runs Neurotoxin, random decoys (windows with random bursts and *clean* labels, so that the model ignores bursts other than the trigger), a stronger trigger ($a = 3$), and a trigger 12 slots before the end of the context ($o = 12$) against the default configuration; E4 varies the number of attackers (1, 2, 4) and the amplitude (0.5–2); X3 and X6 widen the window to $W = 18$. Grid R3 attacks the *calibrated* configuration with attackers that know the code, the anchor, and the deployed thresholds and use 16 public windows of their own (a different draw from the same pool as the server's probes): a *score-aware* attacker that runs BurstGuard's search on its local model before every epoch and penalises the burst response it finds (one penalty weight); *gradient masking*, a 10-minute $+2\sigma$ spike in the corner of the search box or a 60-minute triangular trigger on its boundary, each hidden among decoys of other shapes; a trigger of half a standard deviation; a *threshold-riding* attacker that chooses the largest boost whose estimated scores stay below 0.8 times the thresholds; and an attacker who is active while BurstGuard calibrates online on the first five rounds. R11 repeats the last with the public guard and adds an *oracle* attacker who rides the exact thresholds in force. R8 and R12 test *regroup equivocation*, where the attacker scales the update it sends to the group aggregation below the group thresholds. R5, R12, and R13 use six attackers out of twenty. For the attackers that are not flagged we also report $\mathrm{UPR}_{\mathrm{nat}}$ over all real trigger-like windows (clean: 0.054).

Results. Neurotoxin, random decoys, and stronger triggers do not affect BurstGuard (Table S12): a working trigger needs the forecast to collapse on *some* burst, and the search needs to find only one. In E4, FedAvg suffers $\mathrm{UPR}_{\mathrm{trig}}$ of 0.31–0.90, whereas BurstGuard flags all 360 attacked aggregates and keeps $\mathrm{UPR}_{\mathrm{trig}} \leq 0.012$, even for a single attacker with a half-standard-deviation trigger. The offset trigger at $o = 12$ passes the default configuration in 27% of the attacked rounds; thresholds calibrated for $W = 18$ flag every attacked aggregate without false positives.

The attackers that know the calibrated configuration fare no better. Score-aware training is flagged in 59 of 60 attacked aggregates, and with the penalty weight we used it removes most of

Table S12 Detailed version of Table 6 of the article. Adaptive attackers. UPR_{trig} under FedAvg; attacked aggregates flagged, UPR_{trig}, and UPR_{nat} (real trigger-like windows of all splits; clean 0.054) under BurstGuard; n: runs. Head, model replacement unless noted; without attack $UPR_{trig} \approx 0.008$ (0.03 on the $o = 12$ trigger).

Attacker (grid)	BurstGuard config.	n	FedAvg UPR_{trig}	Flagged	BurstGuard UPR_{trig}	UPR_{nat}
Neurotoxin, head / LoRA (E6)	default	3 / 1	0.64 / 0.92	60/60 / 20/20	0.01 / 0.01	–
Random decoys, head / LoRA (E6)	default	3 / 1	0.36 / 0.74	60/60 / 20/20	0.01 / 0.01	–
Stronger trigger $a = 3$ (E6)	default	3	0.92	60/60	0.005	–
1, 2, or 4 attackers; $a \in \{0.5, 1, 2\}$ (E4)	default	18	0.31–0.90	360/360	$\leq$0.012	–
Trigger at $o = 12$, $W = 6$ / $W = 18$ (E6, X3)	default	3 / 3	0.64	44/60 / 60/60	0.22 / 0.03	–
Trigger at $o = 12$, γ=3, head / LoRA (X3, X6)	calibrated, $W = 18$	3 / 1	–	60/60 / 20/20	0.03 / 0.03	–
Score-aware training, γ=3, head / LoRA (R3)	calibrated	2 / 1	0.034	40/40 / 19/20	0.007 / 0.009	0.045 / 0.051
Gradient masking: spike $w = 1$, $a = 2$, decoys, γ=3 (R3)	calibrated	2	0.673	40/40	0.007	–
Gradient masking: triangle $w = 6$, $a = 2$, decoys, γ=3 (R3)	calibrated	2	0.551	40/40	0.007	–
Small trigger $a = 0.5$, γ=3 / poisoning (R3)	calibrated	2 / 1	0.147 / 0.036	40/40 / 20/20	0.011 / 0.012	–
Threshold riding, 20 rounds / 60 rounds / LoRA (R3)	calibrated	2 / 1 / 1	–	0/40 / 0/60 / 0/20	0.013 / 0.014 / 0.011	0.057 / 0.071 / 0.058
Threshold riding vs. per-mode $\lambda_{anc} = 0.053$ (R3)	calibrated	1	–	0/20	0.010	0.061
Oracle riding of the public caps (R11)	guarded online	2	–	0/40	0.009	0.048
Active during calibration, γ=3 / poisoning (R3, R11)	online, no guard	2 / 2	0.37 / 0.10	0/40 / 10/40	0.364 / 0.101	0.304 / 0.128
Active during calibration, γ=3 / poisoning (R11)	**guarded online**	2 / 2	0.37 / 0.10	**40/40 / 40/40**	**0.007 / 0.007**	0.045 / 0.045
Active during calibration, LoRA, γ=3 / poisoning (R11)	online, no guard	1 / 1	0.73 / 0.19	0/20 / 2/20	0.772 / 0.204	0.149 / 0.105
Active during calibration, LoRA, γ=3 / poisoning (R11)	**guarded online**	1 / 1	0.73 / 0.19	**20/20 / 20/20**	**0.008 / 0.008**	0.051 / 0.051
Regroup equivocation, head replacement / LoRA γ=3 (R8)	calibrated	2 / 1	0.64 / 0.73	40/40 / 20/20	0.014 / 0.021	0.058 / 0.073
Regroup equivocation, with verification (R12)	calibrated	1	0.64	20/20	0.016	0.083

its own backdoor (FedAvg UPR_{trig} 0.034); we did not sweep this weight, so a better-tuned attacker may keep more of the backdoor. Gradient masking with a spike in the corner of the box or a triangle on its boundary, hidden among decoys, is flagged in all 80 attacked aggregates, as is a half-standard-deviation trigger. The threshold-riding attacker is never flagged—by construction it stays below the thresholds—but to stay there it must shrink its update to 0.5–2.9 times a benign one, and the backdoor stays small: UPR_{trig} of 0.011–0.014 over 20 or 60 rounds and on LoRA, and 0.010 against a tighter per-mode anchor threshold, against 0.008 without attack and 0.36 for an unconstrained moderate boost. On real trigger-like windows its effect is visible but small: UPR_{nat} of 0.057–0.071 against 0.054 without attack and 0.305 under FedAvg; an oracle rider of the exact thresholds reaches 0.009 and 0.048.

The attacker active during calibration exposes the weak point of calibrating on the federation itself (Fig. S4a): its own aggregates enter the calibration sample, the online thresholds rise to $\lambda_{last} = 0.37$–0.39 and $\lambda_{anc} = 0.68$–0.71, and nothing is flagged afterwards ($UPR_{trig} = 0.364$ for head and 0.772 for LoRA). The public guard closes this window: the caps from the simulated federation (0.024–0.026 and 0.047–0.056 for head) apply from round 1, every attacked aggregate is flagged, and UPR_{trig} stays at 0.007–0.008. In a regroup, an equivocating attacker who scales its group update below the group thresholds gets its update accepted in 97.5% of the attacked rounds, but the scaled update is so small that UPR_{trig}

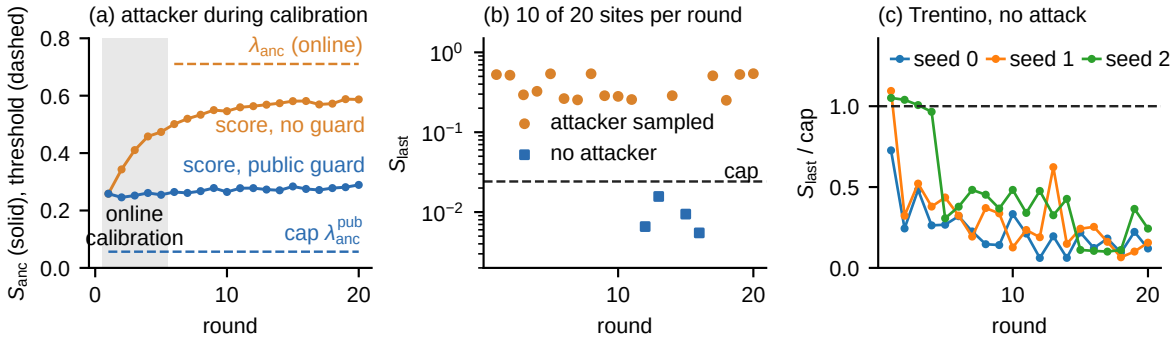

Fig. S4 Per-round view of BurstGuard. (a) An attacker active from round 1 (γ=3, head): without the guard, online calibration on rounds 1–5 absorbs the attack and sets $\lambda_{anc} = 0.711$, above the attacked anchor scores; with the public guard, the cap 0.056 applies from round 1 and every attacked aggregate is flagged and regrouped (grid R11). (b) Half participation (γ=3, seed 0): S_{last} of rounds in which an attacker was sampled (circles) and of rounds without one (squares) against the public cap (grid R17). (c) Trentino without attack: S_{last} relative to the public cap of each seed; the cap is touched only in rounds 1–3 (grid R18).

Table S13 Six attackers out of twenty (head, 40 rounds, two seeds per row and three attack variants per seed, which give nearly identical values; LoRA one run): UPR_{trig}, share of accepted rounds, and clean MASE (grids R5, R12, R13). The attackers' updates are clipped by the best-case norm bound (NB) unless noted.

Defence	UPR_{trig}	Accepted	MASE
NB alone	0.047–0.049	100%	0.662–0.665
NB + BurstGuard (calibrated)	0.043–0.044	100%	0.661–0.663
NB + BurstGuard + verification	0.021	51–54%	0.657–0.658
+ clean-calibrated group thresholds	0.016–0.017	85–88%	0.655–0.656
NB + BurstGuard, reject after 3	0.025–0.028	35–38%	0.666–0.672
NB + BurstGuard recommended	**0.011–0.014**	36–38%	0.658–0.660
BurstGuard recommended, no NB	0.006–0.007	68–73%	0.658–0.664
LoRA: NB alone	0.239	100%	0.642
LoRA: NB + BurstGuard (calibrated)	0.117	100%	0.642
LoRA: NB + BurstGuard + verification	**0.018**	35%	0.645

stays at 0.010–0.016 for head (0.021 for LoRA); verification does not stop it (0.016), because the scaled update passes it as well. On real surges this attacker leaves UPR_{nat} of 0.058–0.083 (one or two runs per cell), the largest residual of any attacker against BurstGuard, which is why commitments that bind both aggregations matter (Section S1.3 in Online Resource 1).

Six attackers out of twenty are a different regime (Table S13). With five groups of four sites, almost every group contains an attacker, so regrouping alone cannot isolate them. A verified norm bound alone leaves UPR_{trig} = 0.047–0.049 for head and 0.239 for LoRA, because six clipped updates still add up. In round 1, these clipped attackers are invisible to every per-round statistic (Table S7): the aggregate's norm and S_{last} stay at clean levels. The anchor score is what sees them: it accumulates their small, consistent contribution and flags 35–36 of the 40 attacked rounds of every recommended run, almost all through S_{anc} alone (the first anchor flag comes in round 5 or 6). Verification then rejects combinations of groups that each carry a little poison. Calibrated BurstGuard without verification accepts such combinations (0.043–0.044); with verification the harm halves (0.021), clean-calibrated group thresholds lower it further (0.016–0.017), and the recommended configuration brings it to 0.011–0.014 for head and, with verification, 0.018 for LoRA. The price is that only 36–38% of the rounds are accepted, and the accepted rounds still contain clipped malicious updates (88 entries in 240 attacked rounds), yet MASE (0.658–0.660) is no worse than with the norm bound alone. Without a norm bound, the recommended configuration flags all 120 attacked aggregates of six unconstrained attackers and keeps UPR_{trig} at 0.006–0.007.

Insight. To hide a backdoor from every burst in the window, the attacker must weaken it, and

to stay under calibrated thresholds, it must keep it small, although on real surges the residual is measurable. The remaining levers are the calibration itself, which public calibration takes away, and the consistency of an update across the two aggregations of a regroup, which needs commitments rather than a better detector. With many attackers, BurstGuard and a verified norm bound complement each other: the bound limits each update, and the anchor score with verification stops their accumulated effect.

S3.2 Further results

Sensitivity to attackers and amplitude (RQ3).

Table S14 varies the number of attackers and the trigger amplitude under model replacement.

Table S14 Sensitivity (grid E4, head, model replacement, two seeds): $\mathrm{UPR}_{\mathrm{trig}}$ for 1, 2, or 4 malicious sites and trigger amplitudes $a \in \{0.5, 1, 2\}$. BurstGuard (default configuration) flags all 360 attacked aggregates.

	FedAvg			FLAME			BurstGuard		
n_{mal}	0.5	1	2	0.5	1	2	0.5	1	2
1	0.452	0.716	0.890	0.012	0.009	0.008	0.012	0.008	0.005
2	0.309	0.643	0.876	0.012	0.010	0.007	0.011	0.007	0.005
4	0.339	0.704	0.905	0.016	0.015	0.015	0.012	0.008	0.005

Public calibration variants (RQ3, RQ5).

Table S15 lists the caps that the public simulation produced in each grid and how they performed.

Table S15 Public calibration: caps ($\lambda^{\text{pub}}_{\text{last}}/\lambda^{\text{pub}}_{\text{anc}}$), simulated rounds × pseudo-clients, attacked aggregates flagged and UPR_{trig}, and flagged clean aggregates. Guarded online: the caps bound an online calibration on the first R_{on} rounds; public only: the caps are the thresholds.

Grid	Setting	Use of the caps	Caps	Simulation	Flagged (UPR_{trig})	Clean FP
R11	Milan head, $\beta = 1.5$	guarded online, $R_{\text{on}} = 5$	0.024–0.026/0.047–0.056	5×8	110/110 (0.007–0.009)	0/60
R11	Milan head, 20 pseudo-clients	guarded online, $R_{\text{on}} = 5$	0.028–0.033/0.061–0.064	5×20	40/40 (0.008)	0/40
R11	Milan head, $\beta = 2$	guarded online, $R_{\text{on}} = 5$	0.031–0.034/0.062–0.075	5×8	40/40 (0.008)	0/60
R11	Milan head	public only	0.024–0.026/0.074	20×8	40/40 (0.008)	0/40
R11	Milan LoRA	guarded online, $R_{\text{on}} = 10$	0.027/0.127	10×8	40/40 (0.008)	0/20
R11	oracle cap riding, head	guarded online, $R_{\text{on}} = 5$	0.024–0.026/0.047–0.056	5×8	0/40 (0.009)	–
R13	Milan head, 2 or 6 attackers	public only	0.024–0.026/0.074–0.078	40×8 or 20×8	160/160 (0.006–0.008)	0/80
R13	Milan head, 6 norm-bounded attackers	public only	0.024–0.026/0.078	40×8	212/240 (0.011–0.014)	–
R14	Milan head, $W = 150$	public only	0.096–0.111/0.237–0.247	20×8	136/140 (Table S16)	0/60
R17	half participation	public only	0.024–0.026/0.069–0.074	20×8	81/81 (0.007)	6/73
R18	Trentino head	public only	0.009–0.013/0.054–0.059	20×8	100/100 (0.025)	4/60
R18	Trentino LoRA	guarded online, $R_{\text{on}} = 10$	0.025/0.103	10×8	37/40 (0.026–0.028)	0/20

Table S16 Recommended (head) and guarded (LoRA) configuration per setting: runs, attacked aggregates flagged, attacked rounds in which a malicious update entered the model (Entries), and share of accepted rounds (mean over seeds per attack variant).

Grid	Setting	Runs	Flagged	Entries	Accepted
R13	2 attackers, γ=10, poisoning	2	40/40	0	100%
R13	6 attackers, no NB	3	120/120	0	68–73%
R13	6 norm-bounded attackers	6	212/240	88	36–38%
R14	$W = 150$, $o = 0$, all three attacks	3	60/60	0	100%
R14	$W = 150$, $o = 144$, γ=10 and 3	3	60/60	0	100%
R14	$W = 150$, $o = 144$, poisoning	1	16/20	6	30%
R16	$W = 150$, $o = 144$, poisoning	1	17/20	4	25%
R17	half participation, three attacks	6	81/81	0	98%
R18	Trentino head, three attacks	5	100/100	0	100%
R11	Milan LoRA (guarded), γ=3, poisoning	2	40/40	0	100%
R18	Trentino LoRA (guarded), γ=3	1	20/20	0	100%
R18	Trentino LoRA (guarded), poisoning	1	17/20	3	100%

S4 RQ4: Does BurstGuard generalise?

This section details RQ4 of the article (Section 5.4).

S4.1 Does BurstGuard generalise?

Purpose. BurstGuard was designed and first calibrated on Milan internet traffic, with probes from the same city and period, a trigger inside the searched family, and all sites in every round. RQ4 asks how it behaves when each of these conditions changes.

Method. X1–X5 repeat RQ1 and RQ2 on SMS and call traffic, which is sparse and near zero at night; X4 calibrates per traffic type with the floor $\phi = 0.1$. R18 moves to Trentino with the recommended configuration, calibrated *only* on the eight Trentino probe areas (head: public-only; LoRA: guarded). R7 attacks with a dip trigger and with a surge one day before the forecast origin ($o = 144$); R14 and R16 screen the latter with a one-day window ($W = 150$) and public caps computed for that window. R1 logs shadow screens with probes from the holiday period, probes of SMS traffic, and 8 or 16 probes instead of 32; we set their thresholds at 1.5 times the largest clean score (as in the calibrated configuration, not with a public simulation) and count false positives by leaving one clean seed out. R9 defends decoder and full fine-tuning with per-mode calibrated thresholds, R2 trains for 100 (head) and 60 (LoRA) rounds, and R17 samples ten of twenty sites per round with $G = 3$.

Results. Fig. 3 summarises the outcome: across regions, triggers, and participation levels, the attack raises UPR_{trig} by a factor of 4–90 under FedAvg, and BurstGuard brings it back to the

Table S17 Generalisation (RQ4): runs (n, FedAvg / BurstGuard), $\mathrm{UPR}_{\mathrm{trig}}$ under FedAvg, attacked aggregates flagged and $\mathrm{UPR}_{\mathrm{trig}}$ under BurstGuard, clean $\mathrm{UPR}_{\mathrm{trig}}$, and flagged clean aggregates (FP). Head unless noted; attacks are model replacement / moderate boost / plain poisoning where three values are given. *27 distinct, since the three attacks share their trajectories.

Setting (grid)	BurstGuard thresholds	n	FedAvg $\mathrm{UPR}_{\mathrm{trig}}$	Flagged	BurstGuard $\mathrm{UPR}_{\mathrm{trig}}$	Clean $\mathrm{UPR}_{\mathrm{trig}}$	FP
SMS, head and LoRA (X2, X5)	calibrated per type	6 / 6	0.45–0.57 / 0.35–0.43	120/120	0.06–0.07	0.06–0.08	0/60
Calls, head and LoRA (X2, X5)	calibrated per type	6 / 6	0.59–0.74 / 0.41–0.55	120/120	0.03	0.03	0/60
Trentino, head (R18)	public only	5 / 5	0.391 / 0.271 / 0.096	100/100	0.025	0.025	4/60
Trentino, LoRA (R18)	guarded	2 / 2	– / 0.542 / 0.125	37/40	0.026 / 0.028	0.025	0/20
Dip trigger, $a = -1$ (R7)	calibrated	4 / 3	0.909 / 0.83 / –	60/60	0.308–0.31	0.304	–
One-day offset $o = 144$, $W = 6$ (R7, R14)	calibrated / public	4 / 4	0.65 / 0.449 / –	20/20 / 2/40 / 7/20	0.04 / 0.457 / 0.423	0.04	–
One-day offset $o = 144$, $W = 150$ (R14, R16)	public only	6 / 5	0.65 / 0.449 / 0.219	20/20 / 40/40 / 33/40	0.040 / 0.038 / 0.077–0.094	0.04	0/60
Holiday probes / SMS probes (R1, shadow)	1.5× clean max	– / 13	–	260/260 each	–	–	0/100
8 / 16 probes (R1, shadow)	1.5× clean max	– / 13	–	260/260 each	–	–	0/100
Decoder / full fine-tuning (R9)	per mode	– / 4	0.93 / 0.83 (E2)	40/40 each	0.009 / 0.010	0.008–0.010	0/40
100 rounds, head, poisoning (R2)	calibrated	– / 2	–	100/100	0.008	0.009	0/100
60 rounds, LoRA, poisoning (R2)	calibrated	– / 2	–	58/60	0.011	0.010	0/60
Half participation, $G = 3$ (R17)	public only	3 / 9	– / 0.531 / 0.183	81/81*	0.007	0.009	6/73

clean level, with one exception (plain poisoning with a one-day offset).

Traffic types. Under FedAvg, model replacement and the moderate boost under-provision 45–74% and 35–55% of the triggered steps of SMS and calls, the latter at as little as 0.6% of clean MASE. With internet thresholds and no floor, BurstGuard flags *every* clean round of SMS and calls, because near-zero night forecasts make relative changes explode, and training never leaves the zero-shot model; with the floor alone, up to 45% of clean LoRA rounds would still be flagged. With the floor and per-traffic calibration, BurstGuard flags all 240 attacked aggregates, keeps $\mathrm{UPR}_{\mathrm{trig}}$ at the clean level and MASE within 0.004 of unattacked training, and flags none of 120 held-out clean rounds (Table S17). These traffic types were run with the calibrated (oracle) thresholds only.

A second region without client calibration. On Trentino, the public caps from eight probe squares are much lower than on Milan (0.009–0.013 for S_{last}), because rural traffic changes less from round to round. With these caps alone, BurstGuard flags all 100 attacked head aggregates and keeps $\mathrm{UPR}_{\mathrm{trig}}$ at the clean 0.025 under every attack, while FedAvg reaches 0.391, 0.271, and 0.096; guarded LoRA flags 37 of 40 attacked aggregates and stays at 0.026–0.028 (FedAvg 0.542 and 0.125). MASE is unchanged (0.708–0.711 vs. 0.708 for clean FedAvg). The tight caps cost four flagged clean rounds out of 60, all in rounds 1–3 of two seeds, where the real score equals the cap to three decimals (0.010 against caps of 0.009–0.010, Fig. S4c); three of them were rejected and one regrouped, with no effect on MASE.

Triggers outside the searched family. The dip trigger is a negative burst and lies outside $\mathcal{B}$, yet the backdoored model also responds abnormally to positive bursts in the window (median S_{last} of 0.18–0.19 against a threshold of 0.031), and BurstGuard flags all 60 attacked aggregates; $\mathrm{UPR}_{\mathrm{trig}}$ stays at the level that a dip causes in a clean model (0.31 vs. 0.304), against 0.83–0.91 under

FedAvg. A surge one day before the forecast origin is mostly invisible to a one-hour search: with $W = 6$, model replacement is still flagged (20 of 20; its backdoor also reacts to bursts at the end), but only 2 of 40 (calibrated) or 7 of 20 (public caps) moderate-boost aggregates are flagged, and $\mathrm{UPR}_{\mathrm{trig}}$ is 0.42–0.46, no better than FedAvg. Widening the search to one day ($W = 150$) with public caps computed for that window flags all 60 attacked aggregates of model replacement and the moderate boost and returns $\mathrm{UPR}_{\mathrm{trig}}$ to the clean 0.038–0.040 for this trigger, with no false positive in 60 clean rounds and no loss of detection on the standard trigger. Plain poisoning with this trigger is caught only in part: 33 of 40 attacked aggregates are flagged, poison enters in the first rounds before the anchor score has grown, and $\mathrm{UPR}_{\mathrm{trig}}$ is 0.077–0.094, about 60% below FedAvg's 0.219 but twice the clean level, with only 25–30% of the rounds accepted.

Probes, modes, horizons, and participation. Probes from the holiday period, probes of SMS traffic, or only 8 or 16 probes detect all 260 attacked head aggregates with no false positive in leave-one-seed-out replay; for LoRA, these shadow screens flag 20 of 20 moderate-boost aggregates and 16–20 of 20 poisoning aggregates, close to the main screen in the same replay (18 of 20). Decoder and full fine-tuning, which need their own thresholds (clean S_{last} up to 0.034 and 0.050), are defended with per-mode calibration (40/40 each, one seed per mode, no false positive). Over 100 head rounds and 60 LoRA rounds, the calibrated thresholds produce no false positive and flag 100/100 and 58/60 poisoning aggregates. With half of the sites per round, the attackers are absent from some rounds and the aggregates are noisier; BurstGuard flags all 81 attacked aggregates (27 distinct, since the three attacks share their trajectories) and keeps $\mathrm{UPR}_{\mathrm{trig}} = 0.007$ (FedAvg 0.531 and 0.183), while 6 of 73 unique attacker-free rounds exceed the cap (scores of 0.025–0.045 against caps of 0.024–0.026; three rejected, the others regrouped), and MASE equals that of FedAvg (0.655 vs. 0.658). Fig. S4b shows how far the attacked rounds lie above the cap.

Insight. The functional question transfers; its scale does not, and calibration supplies the scale. The public simulation calibrates a new region from eight probe squares without any client round, and a wider window covers a trigger placed a day earlier at no extra cost per round (RQ5). What does not transfer for free is the margin: when the clean scores of the real federation sit close to the simulated ones (rural data, few participants), some rounds touch the cap, which costs a few rounds of training but no accuracy.

S4.2 Further results

Probe shift and search window (RQ4).

Table S18 gives the shadow-screen replay of grid R1, and Table S19 the score growth with the search window for the one-day-offset trigger in grid R7.

Table S18 Probe shift (grid R1, shadow screens on the same aggregates): thresholds at 1.5× the largest clean score per mode and screen, attacked aggregates flagged per attack, and false positives in leave-one-seed-out replay over the five clean head runs.

Mode	Probes	$\lambda_{\mathrm{last}}/\lambda_{\mathrm{anc}}$	γ=1	γ=3	γ=10	FP
Head	main (32)	0.035/0.054	100/100	100/100	60/60	0/100
	holiday	0.029/0.063	100/100	100/100	60/60	0/100
	SMS	0.121/0.252	100/100	100/100	60/60	0/100
	8 probes	0.037/0.063	100/100	100/100	60/60	0/100
	16 probes	0.033/0.055	100/100	100/100	60/60	0/100
LoRA	main (32)	0.030/0.132	18/20	20/20	–	–
	holiday	0.030/0.173	20/20	20/20	–	–
	SMS	0.114/0.595	16/20	20/20	–	–
	8 probes	0.040/0.235	16/20	20/20	–	–
	16 probes	0.030/0.124	19/20	20/20	–	–

Table S19 One-day-offset trigger ($o = 144$, moderate boost, grid R7): median S_{last} and S_{anc} of the attacked aggregates in shadow screens with growing search width W, two seeds. Clean runs at $W = 150$ score at most 0.070–0.073 (S_{last}) and 0.199–0.202 (S_{anc}) (grid R14).

W	median S_{last}	median S_{anc}
6	0.012–0.021	0.047–0.068
18	0.019–0.027	0.146–0.163
36	0.024–0.031	0.228–0.245
72	0.034–0.042	0.365
150	0.098–0.122	0.991–1.111

Traffic types (RQ4).

Tables S20 and S21 detail SMS and call traffic.

Table S20 SMS and calls (test split; FL and FedAvg: grids X1–X2; BurstGuard: grid X5, calibrated with $\phi = 0.1$, head two seeds, LoRA one). UPR_{trig} under model replacement / moderate boost; ΔMASE: cost of the moderate boost under FedAvg. Clean UPR_{trig}: SMS 0.06–0.08, calls 0.03.

			FedAvg		BurstGuard	
Traffic	Mode	FL MASE	UPR_{trig}	ΔMASE	UPR_{trig}	MASE
SMS	Head	0.670	0.45/0.35	+4.6%	0.06/0.06	0.668
	LoRA	0.624	0.57/0.43	+2.2%	0.07/0.07	0.621
Calls	Head	0.558	0.59/0.41	+8.8%	0.03/0.03	0.554
	LoRA	0.500	0.74/0.55	+0.6%	0.03/0.03	0.499

Table S21 BurstGuard with the internet default configuration and no relative floor on SMS and calls (grid X2): every clean round is flagged, so training never leaves the zero-shot model. With the floor but still with the internet thresholds, monitor mode would flag 2.5% (SMS, head), 7.5% (calls, head), 40% (calls, LoRA), and 45% (SMS, LoRA) of clean rounds (grid X4).

Traffic	Mode	Clean FPR	Clean MASE	Zero-shot MASE
SMS	Head	1.00	0.710	0.710
SMS	LoRA	1.00	0.710	0.710
Calls	Head	1.00	0.643	0.643
Calls	LoRA	1.00	0.643	0.643

S5 RQ5: Which design choices matter, and what does it cost?

This section details RQ5 of the article (Section 5.5).

S5.1 Which design and calibration choices matter?

Purpose. RQ5 isolates the contribution of each design element—the two references, the search, the probes, the calibration, and the regroup options—to decide which ones a deployment needs.

Table S22 Detailed version of Table 7 of the article. Ablation (grid E9, head, two seeds, monitor mode). Pilot: TPR and FPR at the pilot floors (0.08/0.3), as logged. Calibrated replay: each variant's thresholds at 1.5 times its own clean maxima; flagged moderate-boost aggregates (of 40) and the round-1 separation (Sep.: smallest attacked S_{last} over largest clean S_{last}). Default: $K = 8$, $|\mathcal{P}| = 32$, $A = 2$, $W = 6$.

	Pilot TPR			Calibrated replay		
Variant	γ=10	γ=3	FPR	Flagged	Sep.	CPU s
Default	1.00	1.00	0.00	40/40	11.7	6.9
S_{last} only (replay)	0.25	0.10	0.00	–	–	–
S_{anc} only (replay)	1.00	0.95	0.00	–	–	–
No ascent (K=0)	0.93	0.05	0.00	40/40	13.3	2.4
K=3	1.00	1.00	0.00	40/40	12.2	3.1
$\lvert\mathcal{P}\rvert$=16	1.00	1.00	0.00	40/40	11.7	3.2
$\lvert\mathcal{P}\rvert$=64	1.00	0.97	0.00	40/40	13.6	13.1
A=1	1.00	0.05	0.00	40/40	10.6	6.0
W=3	1.00	1.00	0.00	40/40	15.1	–
Drops only (s=+1)	1.00	1.00	0.00	40/40	33.2	4.7

Method. Grid E9 runs BurstGuard variants in monitor mode (head, two seeds, model replacement and moderate boost); we evaluate each variant twice: with the pilot floors (as logged) and with thresholds calibrated for that variant (1.5 times its own clean maxima), replayed on the logged scores. Round 1 of this replay is exact, because all variants see the same first candidate; later rounds come from monitor-mode runs in which the backdoor was accepted. Other decision rules are replayed on the logged scores of E3 and X8; E5 gives the clean anchor drift over 40 rounds. R6 calibrates online on the first $R_{\text{on}} \in \{3, 5, 10\}$ rounds of a run with margins 1.2–2; R11 varies the public calibration (8 or 20 pseudo-clients, margin 1.5 or 2, public-only over 20 rounds); we replay the public margin β on the logged aggregate scores of the recommended runs (exact up to the first changed decision); R8 varies the regroup (three groups, clean-calibrated group thresholds, rejection after three consecutive regroups); E7 compares regrouping with rejection.

Results. *References.* Neither reference suffices alone (Table S22, monitor mode). Replayed alone, S_{last} flags only 10% of the moderate-boost aggregates, but this reflects monitor mode: once the first poisoned update is accepted, later updates add little new behaviour, whereas in enforce mode the first round is already flagged (round-1 S_{last} of 0.24–0.26 against a calibrated threshold of 0.031). S_{anc} flags 95%, missing the earliest rounds,

Table S23 Public margin β replayed on the logged aggregate scores of the recommended head runs (caps scale with β): attacked aggregates flagged and false alarms over unique attacker-free rounds.

Setting	$\beta = 1.5$	$\beta = 1.75$	$\beta = 2$
Two attackers, all settings	281/281	281/281	278/281
Six norm-bounded attackers	212/240	193/240	56/240
One-day offset, poisoning	33/40	17/40	0/40
False alarms, half participation	6/73	1/73	1/73
False alarms, Trentino	4/60	0/60	0/60
False alarms, Milan, all sites	0/140	0/140	0/140

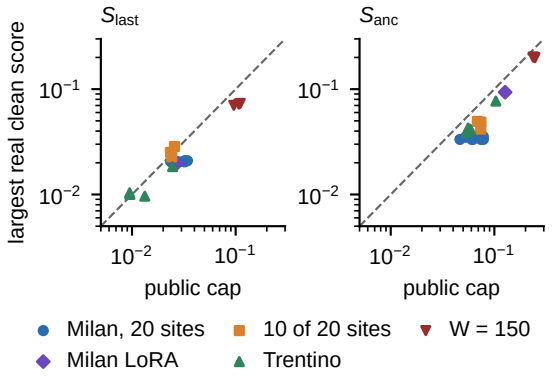

Fig. S5 Public cap against the largest score of the real clean federation, for every public-calibrated clean run (grids R11, R13, R14, R17, R18). Points below the diagonal leave headroom; points on or above it can produce false alarms.

when the backdoor is still small relative to benign drift; and it is S_{anc} that stops six norm-bounded attackers (RQ3).

The search. With the pilot floors, random bursts without ascent ($K = 0$) and an amplitude bound equal to the trigger amplitude ($A = 1$) detect only 5% of the moderate-boost aggregates. With thresholds calibrated for each variant, however, every variant flags all moderate-boost and model-replacement aggregates without a clean flag, and the round-1 separation is 10.6–33 times for all of them: the ascent roughly doubles both the attacked and the clean scores (0.13–0.14 and 0.010 without ascent, 0.24–0.26 and 0.021 with it). For these boosted attacks on head fine-tuning, therefore, calibration, not the ascent, carries the detection. What the ascent adds is a worst-case burst that the operator can inspect and a lower bound on Eq. (6); whether it matters for weaker or masked attacks (plain poisoning, the gradient-masking attackers of RQ3, LoRA) was not tested with $K = 0$. Three ascent steps cost half of eight, and searching for drops only gives the widest separation for this trigger at two thirds of the cost.

Calibration. Replayed calibration is robust to the margin ($\beta \in \{1.2, 1.5, 2\}$) and to pooled or per-mode thresholds: every X8 run is flagged in its first attacked round, and no held-out clean round is. Online calibration on the first $R_{on} = 3$, 5, or 10 rounds of a run catches every later attack (34/34, 30/30, 20/20) without false positives in 332 clean rounds, provided that no attacker is present during calibration (RQ3). Calibration is sensitive to the *horizon*: LoRA's anchor score keeps growing (about 0.02 in round 1, 0.08 in round 20, 0.08–0.11 in rounds 30–40 of E5), so calibrating LoRA on its first five rounds would flag 13–15 of the remaining 15 clean rounds, whereas ten rounds flag none. The public variants all detect every attacked aggregate without false positives. The variant with 20 pseudo-clients reuses the eight areas with other seeds, so it tests the size of the simulation rather than a larger public pool; its caps are slightly higher (0.028–0.033 vs. 0.024–0.026), and a margin of 2 raises them to 0.031–0.034. Fig. S5 shows the headroom that the margin buys. With all 20 Milan sites, the largest real clean S_{last} is 0.020–0.021, below every cap; with $W = 150$ the headroom is wider still. With half of the sites (0.023–0.029 vs. caps of 0.024–0.026) and on Trentino (0.010 vs. 0.009–0.013), the real clean scores reach the cap, which explains the false alarms of RQ4. More simulated participants would not help here: a simulated aggregate of eight pseudo-clients is, if anything, noisier than a real one of ten or twenty, so the gap comes from the real clients' heterogeneity and from the in-sample scoring of the simulation (Section 3.3). Table S23 shows the trade-off that remains: a margin of 1.75 instead of 1.5 removes the Trentino false alarms and five of six with half participation without losing any two-attacker detection, but it misses more of the weakest attacks (six norm-bounded attackers: 193 instead of 212 of 240; one-day-offset poisoning: 17 instead of 33 of 40). We keep $\beta = 1.5$.

Regrouping. Three groups work as well as five against two attackers (40/40 flagged, $UPR_{trig} = 0.007$). Group thresholds calibrated directly on clean groups (0.082/0.114) are an alternative to

$\sqrt{G}$ scaling (40/40, 0.007); with six attackers, $\sqrt{G}$ times the caps gives a lower $\mathrm{UPR}_{\mathrm{trig}}$ (0.012 vs. 0.015) at a lower acceptance rate (37% vs. 75%). Rejecting a flagged round instead of regrouping costs accuracy (MASE 0.684 vs. 0.650, E7), and rejecting after three consecutive regroups costs a little (0.663 vs. 0.652) while limiting how often partial sums are revealed. Verification never rejected a clean combination with two attackers (60 regroup rounds, no rejection) and is decisive with six (Table S13). An adaptive S_{last} threshold without a cap can be pushed up by accepted rounds that are themselves noisy group aggregates: with six unconstrained poisoners and verification, the adaptive threshold rose from 0.031 to 0.06–0.10 after round 29 and let poison in ($\mathrm{UPR}_{\mathrm{trig}} = 0.063$, one seed); capping it at the public value prevents this ($\mathrm{UPR}_{\mathrm{trig}} = 0.006$–$0.007$, R13).

Insight. Calibration and the two references carry the detection: the last-model reference catches a backdoor in the round it is injected, and the anchor reference catches what accumulates slowly, including the small contributions of bounded attackers. The ascent is cheap insurance that returns an inspectable burst, but its necessity is not shown for boosted attacks. Calibration must be per deployment, cover the training horizon, and be immune to the federation it protects, which the public simulation provides; its margin trades a few false alarms against the weakest attacks. Regrouping keeps the benign work of flagged rounds, and verification makes it hold when attackers are many.

S5.2 What does BurstGuard cost?

Purpose. RQ5 asks what BurstGuard adds to a federated round, on the server and on the clients, and how this scales with the context length and the search window.

Method. Grids E8 and R10 run head and LoRA rounds in Flower with and without SecAgg+ and BurstGuard (20 virtual clients on one machine, three seeds each, five rounds); X8 gives the cost of regrouped rounds; R10 also varies the context length (512, 1024, 1536 slots; one run each), and X3 and R14 the search window. The public simulation is timed in R11–R18.

Results. In a clean round, BurstGuard adds 4.3–4.7 s of single-threaded server CPU time for head and 6.6–8.5 s for LoRA (Fig. S6a), and nothing on the clients. A head round under SecAgg+ grows from 8.7–10.4 s to 13.1–14.1 s, and SecAgg+ itself adds about one second. A LoRA round takes 83 s on average with BurstGuard and 60 s without it (six runs each), so the wall-clock overhead (+23 s) is about three times BurstGuard's own CPU time in our single-machine setup. SecAgg+ does not change accuracy or $\mathrm{UPR}_{\mathrm{trig}}$ (MASE 0.662 for head and 0.645 for LoRA with and without it). A persistent attacker who forces a regroup in every round raises the server's CPU time to 39–58 s per round, plus an emulated second secure aggregation that we do not time. The cost grows linearly with the context length (4.9, 10.4, and 16.3 s for 512, 1024, and 1536 slots, Fig. S6b); in these five-round runs the longer context lowered test MASE (0.661, 0.623, 0.611) but raised holiday MASE (0.682, 0.745, 0.738). The cost does not grow with the search window: 4.5–7.0 s at $W = 6$, 4.9–7.9 s at $W = 18$, and 5.9 s per clean round at $W = 150$. Three ascent steps instead of eight halve the cost without changing any decision in E9. The public simulation runs once, before training: 20 simulated head rounds with eight pseudo-clients took 0.9–6.7 min, 40 rounds 19–37 min, and 10 simulated LoRA rounds 23 min to 2.3 h, depending on how many other jobs shared the machine.

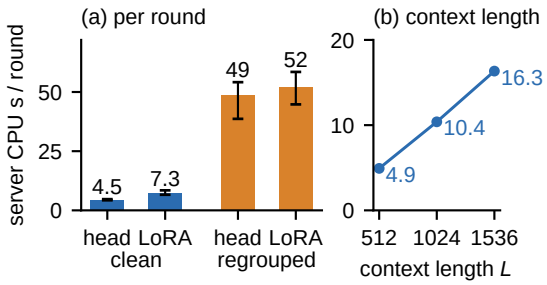

Fig. S6 Server CPU time of BurstGuard: (a) per clean round (Flower, R10, six runs per mode) and per regrouped round under a persistent attacker (X8), mean with range over runs; (b) per clean round against the context length (R10, monitor mode, one run each).

Insight. BurstGuard moves the cost of robustness from constrained clients to the server, where a few seconds per round are small against the round time; a persistent attacker multiplies it by about ten, and a limit on regroups bounds it. The one-off public simulation is cheap for head fine-tuning

and costly for LoRA over a long horizon, which is why we guarded an online calibration with a shorter simulation for LoRA.

S5.3 Further results

Calibration replay and online calibration (RQ5).

Table S24 replays Eq. (S2) in Online Resource 1 with different margins and pooling on logged scores; Table S25 calibrates online on the first rounds of a run. On the first rounds of a single logged run, $R_{\text{on}} = 3$ suffices for head, whereas LoRA, whose anchor score keeps growing, flags 13–15 of its remaining 15 clean rounds with $R_{\text{on}} = 5$ and none with $R_{\text{on}} = 10$.

Table S24 Replayed calibration on the logged scores of the clean E3 runs (calibration) and of X8 (held-out clean seeds: head 40 rounds, LoRA 20 rounds; attacked runs: first attacked round). Pooled: one threshold pair over head and LoRA. Every configuration flags the first attacked aggregate of every X8 run (head 9/9, LoRA 6/6).

β	Pooling	Head $\lambda_{\text{last}}/\lambda_{\text{anc}}$	LoRA $\lambda_{\text{last}}/\lambda_{\text{anc}}$	FP
1.2	pooled	0.025/0.112	0.025/0.112	0/60
1.2	per mode	0.025/0.043	0.024/0.112	0/60
1.5	pooled	0.031/0.140	0.031/0.140	0/60
1.5	per mode	0.031/0.053	0.030/0.140	0/60
2.0	pooled	0.042/0.187	0.042/0.187	0/60
2.0	per mode	0.042/0.071	0.040/0.187	0/60

Table S25 Online calibration on the first R_{on} rounds of a run, attack starting after calibration (grid R6, head unless noted): attacked aggregates flagged, UPR_{trig}, and flagged clean aggregates (clean runs, 20 rounds each).

R_{on}, β	Attack	Flagged	UPR_{trig}	Clean FP
3, 1.5	γ=3	34/34	0.007	0/60
	poisoning	34/34	0.007	
5, 1.5	γ=3	30/30	0.007	0/60
	poisoning	30/30	0.007	
5, 1.2	poisoning	15/15	0.009	0/20
5, 2.0	poisoning	15/15	0.009	0/20
10, 1.5	γ=3	20/20	0.007	0/60
	poisoning	20/20	0.007	
10, 1.5, LoRA	γ=3	10/10	0.009	0/20

Ablation replayed at calibrated thresholds (RQ5).

Table S26 gives the clean maxima and round-1 scores behind the calibrated replay of Table S22. Every variant flags all 40 moderate-boost and all 40 model-replacement aggregates of the two seeds and none of the 40 clean rounds, both at its own thresholds and at the global calibrated thresholds 0.031/0.14.

Table S26 Grid E9 replay: largest clean $S_{\text{last}}/S_{\text{anc}}$ over 40 clean rounds, round-1 S_{last} of the two moderate-boost and two model-replacement runs, and the variant's own thresholds (1.5× the clean maxima).

Variant	Clean max	Round 1, γ=3	Round 1, γ=10	Own thresholds
Default	0.021/0.035	0.244–0.257	0.812–0.907	0.031/0.053
No ascent ($K = 0$)	0.010/0.020	0.130–0.139	0.428–0.471	0.015/0.030
$K = 3$	0.018/0.033	0.224–0.238	0.743–0.832	0.027/0.050
$\lvert\mathcal{P}\rvert = 16$	0.019/0.034	0.222–0.251	0.910–1.008	0.029/0.051
$\lvert\mathcal{P}\rvert = 64$	0.015/0.025	0.198–0.218	0.750–0.847	0.022/0.037
$A = 1$	0.012/0.020	0.122–0.127	0.404–0.448	0.017/0.030
$W = 3$	0.016/0.024	0.242–0.254	0.796–0.887	0.024/0.037
Drops only	0.007/0.022	0.244–0.257	0.812–0.907	0.011/0.034

References

[1] Blanchard P, El Mhamdi EM, Guerraoui R, et al (2017) Machine learning with adversaries: Byzantine tolerant gradient descent. In: Advances in Neural Information Processing Systems (NeurIPS), pp 119–129

[2] Cao X, Fang M, Liu J, et al (2021) FLTrust: Byzantine-robust federated learning via trust bootstrapping. In: Proc. NDSS, https://doi.org/10.14722/ndss.2021.24434

[3] Nguyen TD, Rieger P, Chen H, et al (2022) FLAME: Taming backdoors in federated learning. In: Proc. USENIX Security, Boston, MA, pp 1415–1432, https://doi.org/10.48550/arXiv.2101.02281

[4] Pillutla K, Kakade SM, Harchaoui Z (2022) Robust aggregation for federated learning. IEEE Transactions on Signal Processing 70:1142–1154. https://doi.org/10.1109/TSP.2022.3153135

[5] Sun Z, Kairouz P, Suresh AT, et al (2019) Can you really backdoor federated learning? arXiv preprint arXiv:1911.07963, https://doi.org/10.48550/arXiv.1911.07963

[6] Yin D, Chen Y, Ramchandran K, et al (2018) Byzantine-robust distributed learning: Towards optimal statistical rates. In: Proc. ICML, pp 5650–5659, https://doi.org/10.48550/arXiv.1803.01498