Supplementary Information: Analysis of Federated Aggregation under Model Poisoning and Backdoor Attacks across Datasets and Architectures

Soumya Mazumdar[1,*], **Vineet Kumar Rakesh**[2], **and Tapas Samanta**[2,3]

[1]Gargi Memorial Institute of Technology, affiliated to Maulana Abul Kalam Azad University of Technology, Department of Computer Science and Business Systems, Kolkata 700144, India
[2]Variable Energy Cyclotron Centre, Computer and Informatics Group, Kolkata 700064, India
[3]Homi Bhabha National Institute, Department of Engineering Sciences, Mumbai 400094, India
[*]reachme@soumyamazumdar.com

ABSTRACT

This Supplementary Information provides the complete evidence-backed extension of the main manuscript. It contains the full 500-cell canonical seed-1 accuracy matrix; extended accuracy, rank, calibration, and worst-client summaries; dataset- and architecture-level sensitivity analyses; FedPARETO–Krum matched comparisons; recurrent low-accuracy diagnostics; source-code-audited BadNets Triggered Target-Label Rate (TTLR) results; additional-seed evidence; single-seed ablation and anchor-size studies; execution-provenance and software-environment details; and a mechanism-by-mechanism audit of the supplied FedPARETO implementation. The supplementary analysis is deliberately descriptive where the canonical benchmark lacks balanced independent replication. In particular, variation across the 25 dataset–architecture configurations is treated as cross-configuration heterogeneity rather than seed-level uncertainty, and source-code observations are separated from claims about the generating provenance of every numerical result. Colour is used consistently to distinguish clean, sign-flipping, Gaussian, BadNets, implemented, and non-implemented evidence categories without changing the numerical interpretation.

Supplementary evidence summary

- **Canonical completeness:** all 500 seed-1 method–dataset–architecture–condition cells have final run-summary records; 490 additionally have located successful execution logs.
- **Threat-dependent ranking:** clean performance favours Trimmed Mean/FedAvg, whereas Krum is the leading method under the evaluated sign-flip and Gaussian conditions.
- **Backdoor interpretation:** the retained BadNets field is TTLR rather than conventional target-excluding ASR; ordinary accuracy and targeted trigger behaviour are therefore analysed separately.
- **Implementation boundary:** the audited FedPARETO scaffold is a Pareto-inspired utility heuristic and can exhibit update–utility mismatch after post-training sign-flip/Gaussian corruption. FedPARETO-UC remains proposed rather than experimentally evaluated.

S1. Evidence scope, provenance, and supplementary objectives

The purpose of this Supplementary Information is not merely to archive additional tables. It provides the evidence trail required to interpret the main manuscript at the level of individual benchmark cells, secondary outcomes, sensitivity analyses, and implementation details. The canonical design crosses five datasets, five architectures, five aggregation methods, and four experimental conditions, producing 500 seed-1 cells. Every cell has a run-summary record. Successful execution logs were located for 490 cells: 454 original successful runs and 36 successful repaired or rerun experiments. Ten clean SVHN cells—MobileNetV3-Small and ShuffleNetV2 across all five methods—retain summary records but no successful execution log was located in the supplied archives.

Table S1 quantifies these provenance classes. The distinction matters because *value completeness* and *execution-provenance completeness* answer different questions. Numerical comparisons use the complete 500-cell matrix, but statements about exact run settings or generating software are restricted to cells for which the necessary provenance exists.

Figure S13 later visualizes the same provenance composition. The table is retained here because exact counts are more

Table S1. Canonical evidence coverage and provenance. Fractions are calculated over the 500 canonical seed-1 cells. Summary-only cells are retained rather than imputed or discarded.

Provenance class	Cells	Fraction
Original successful execution log + summary	454	90.8%
Successful repaired/rerun log + summary	36	7.2%
Summary only	10	2.0%
Total	500	100.0%

useful for reproducibility auditing, whereas the figure provides an immediate visual summary of evidence strength.

S2. Supplementary analytical framework and metric definitions

For the supplementary analyses, let $A_{d,h,m,c}$ denote final global test accuracy for dataset d, architecture h, aggregation method m, and condition c. The canonical matrix contains one identified seed for each such cell. The matched attack-minus-clean change for attack condition a is

$$\Delta^{\mathrm{acc}}_{d,h,m,a} = 100\left(A_{d,h,m,a} - A_{d,h,m,\mathrm{clean}}\right), \tag{S1}$$

measured in percentage points. Equation (S1) matches attacked and clean cells by dataset, architecture, and method. It is not described as a seed-paired treatment effect because independently matched repeated seeds are unavailable for the canonical matrix.

Within each dataset–architecture configuration, the five methods are ranked by accuracy, with average ranks assigned to exact ties. If $R_{d,h,m,c}$ denotes that within-task rank, the mean rank for method m in condition c is

$$\bar{R}_{m,c} = \frac{1}{25}\sum_{d,h} R_{d,h,m,c}. \tag{S2}$$

Equation (S2) gives each dataset–architecture task equal weight and is used for the ranking and leave-one-group-out analyses. Lower values indicate better relative accuracy.

For calibration, the retained summary metric is expected calibration error (ECE). If confidence is partitioned into B bins, a standard ECE form is

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|I_b|}{n}\left|\mathrm{acc}(I_b) - \mathrm{conf}(I_b)\right|, \tag{S3}$$

where I_b is the set of examples in bin b. Equation (S3) measures disagreement between empirical correctness and average confidence. ECE is reported descriptively because a numerically small value does not necessarily imply a useful classifier when accuracy has collapsed.

Worst-client accuracy is retained directly in the run summaries. Conceptually, if C_k is the evaluation accuracy of client k, the lower-tail diagnostic can be represented as

$$W = \min_k C_k. \tag{S4}$$

Equation (S4) explains why worst-client accuracy answers a different question from global accuracy: it records the lowest observed client-level performance rather than the population average.

For BadNets, the audited code stamps the trigger on every test input and counts target-label predictions. The retained metric is therefore

$$\mathrm{TTLR} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}[f(T(x_j)) = y^\star]. \tag{S5}$$

Equation (S5) includes natural target-class examples and is not identical to conventional target-excluding ASR. The supplementary BadNets analysis therefore reports TTLR by name and does not infer an unavailable conventional ASR.

These definitions establish the statistical boundary used throughout the supplement: means, medians, IQRs, cross-configuration SDs, ranks, matched differences, and sensitivity analyses are descriptive summaries of heterogeneous tasks. Inferential p-values, seed-level confidence intervals, or pseudo-replicated standard errors are not reported for the canonical matrix.

S3. Full canonical accuracy matrix

Table S2 reports final global test accuracy for every canonical cell. Values are converted from proportions to percentages. Colour bands distinguish the four conditions: blue for clean, orange for sign flipping, green for Gaussian poisoning, and purple for BadNets. The purpose of the full matrix is to make aggregate conclusions auditable at the individual dataset–architecture level. For example, it shows directly that Krum remains high in many sign-flip/Gaussian cells where other methods fall to recurrent low values, while clean and BadNets ordinary-accuracy patterns are more favourable to FedAvg, Trimmed Mean, and FedPARETO.

The machine-readable version of Table S2, including ECE, worst-client accuracy, TTLR where defined, provenance class, repair status, and run metadata, is supplied as `source_data/canonical_results.csv`. The table should therefore be read as a human-auditable view of the same canonical source data rather than as an independently edited result table.

Table S2. Full canonical final-test-accuracy matrix (%). Each row corresponds to one dataset–architecture–condition combination. The five aggregation columns report final global test accuracy. Condition-specific row shading is used only as a visual guide and does not encode statistical significance.

Dataset	Architecture	Condition	FedAvg	Trimmed Mean	Krum	FLTrust	FedPARETO
GTSRB	SimpleCNN	Clean	56.34	55.51	33.22	7.89	52.17
GTSRB	SimpleCNN	Sign flipping	0.48	38.93	34.26	7.13	42.98
GTSRB	SimpleCNN	Gaussian	5.63	52.92	36.83	7.58	38.34
GTSRB	SimpleCNN	BadNets	54.81	53.39	31.23	7.25	51.12
SVHN	SimpleCNN	Clean	90.05	89.84	74.09	18.22	88.10
SVHN	SimpleCNN	Sign flipping	6.70	63.56	80.16	17.93	83.85
SVHN	SimpleCNN	Gaussian	19.10	72.05	72.70	18.14	19.59
SVHN	SimpleCNN	BadNets	89.49	89.18	75.03	18.21	88.33
MNIST	SimpleCNN	Clean	99.08	99.05	97.87	91.74	98.90
MNIST	SimpleCNN	Sign flipping	9.80	95.73	98.01	91.25	96.93
MNIST	SimpleCNN	Gaussian	87.11	97.74	98.32	91.27	95.09
MNIST	SimpleCNN	BadNets	99.04	98.99	98.10	90.19	99.11
CIFAR-10	SimpleCNN	Clean	78.05	77.89	60.79	27.66	74.04
CIFAR-10	SimpleCNN	Sign flipping	10.00	65.84	61.57	29.44	67.28
CIFAR-10	SimpleCNN	Gaussian	10.00	54.86	61.88	29.47	23.32
CIFAR-10	SimpleCNN	BadNets	77.95	76.67	58.93	28.60	76.26
CIFAR-100	SimpleCNN	Clean	47.52	46.84	27.00	3.87	42.65
CIFAR-100	SimpleCNN	Sign flipping	1.00	1.00	24.66	3.08	33.04
CIFAR-100	SimpleCNN	Gaussian	1.00	14.87	26.93	3.69	1.00
CIFAR-100	SimpleCNN	BadNets	46.50	45.79	25.55	1.80	42.59
GTSRB	ResNet-18	Clean	85.40	86.35	76.89	24.15	85.63
GTSRB	ResNet-18	Sign flipping	0.48	52.83	79.07	12.26	54.84
GTSRB	ResNet-18	Gaussian	6.18	17.85	73.98	23.66	6.02
GTSRB	ResNet-18	BadNets	85.30	85.29	68.56	21.99	85.48
SVHN	ResNet-18	Clean	94.21	93.83	89.25	71.47	93.12
SVHN	ResNet-18	Sign flipping	6.70	84.37	89.34	56.18	81.86
SVHN	ResNet-18	Gaussian	15.94	52.22	90.23	72.81	44.30
SVHN	ResNet-18	BadNets	94.27	94.30	90.56	68.45	93.44
MNIST	ResNet-18	Clean	99.17	99.12	98.22	97.82	99.12
MNIST	ResNet-18	Sign flipping	9.80	9.80	98.64	94.30	9.80
MNIST	ResNet-18	Gaussian	52.96	93.77	98.75	97.83	80.95
MNIST	ResNet-18	BadNets	99.10	99.17	98.54	97.86	99.11
CIFAR-10	ResNet-18	Clean	86.72	85.95	72.17	50.15	84.46
CIFAR-10	ResNet-18	Sign flipping	10.00	63.66	72.37	24.22	44.75
CIFAR-10	ResNet-18	Gaussian	14.16	31.28	75.48	46.54	14.01
CIFAR-10	ResNet-18	BadNets	86.53	85.86	75.55	47.24	84.46
CIFAR-100	ResNet-18	Clean	60.15	60.21	23.23	15.75	52.65
CIFAR-100	ResNet-18	Sign flipping	1.00	1.00	25.09	11.06	1.00
CIFAR-100	ResNet-18	Gaussian	1.04	7.13	28.62	14.88	1.13
CIFAR-100	ResNet-18	BadNets	59.08	58.81	24.61	15.33	47.46
GTSRB	MobileNetV3-Small	Clean	74.03	74.98	46.20	5.23	64.17
GTSRB	MobileNetV3-Small	Sign flipping	0.48	41.74	51.65	5.70	11.87
GTSRB	MobileNetV3-Small	Gaussian	0.48	0.48	54.66	5.70	0.48
GTSRB	MobileNetV3-Small	BadNets	73.35	72.33	55.08	0.48	67.53
SVHN	MobileNetV3-Small	Clean	87.11	87.96	78.26	30.59	85.14
SVHN	MobileNetV3-Small	Sign flipping	6.70	76.82	81.68	15.95	6.70
SVHN	MobileNetV3-Small	Gaussian	6.70	6.70	80.35	15.94	6.70

Continued on next page

Table S2 – continued from previous page

Dataset	Architecture	Condition	FedAvg	Trimmed Mean	Krum	FLTrust	FedPARETO
SVHN	MobileNetV3-Small	BadNets	86.68	86.07	76.98	30.59	85.49
MNIST	MobileNetV3-Small	Clean	98.51	98.40	97.89	9.74	98.51
MNIST	MobileNetV3-Small	Sign flipping	9.80	65.44	97.40	9.74	88.86
MNIST	MobileNetV3-Small	Gaussian	9.80	9.80	97.68	9.74	9.80
MNIST	MobileNetV3-Small	BadNets	98.47	98.56	97.63	9.74	98.61
CIFAR-10	MobileNetV3-Small	Clean	64.62	62.90	47.81	10.00	57.68
CIFAR-10	MobileNetV3-Small	Sign flipping	10.00	44.88	47.29	10.00	10.00
CIFAR-10	MobileNetV3-Small	Gaussian	10.00	10.00	49.24	11.44	10.00
CIFAR-10	MobileNetV3-Small	BadNets	63.44	61.22	45.62	10.00	54.81
CIFAR-100	MobileNetV3-Small	Clean	26.17	31.01	16.09	1.09	25.37
CIFAR-100	MobileNetV3-Small	Sign flipping	1.00	6.19	15.84	1.13	10.54
CIFAR-100	MobileNetV3-Small	Gaussian	1.00	1.00	16.29	1.63	1.00
CIFAR-100	MobileNetV3-Small	BadNets	25.77	28.45	16.94	1.00	25.06
GTSRB	EfficientNet-B0	Clean	89.28	88.16	80.56	72.08	88.65
GTSRB	EfficientNet-B0	Sign flipping	0.48	80.04	79.14	34.11	0.48
GTSRB	EfficientNet-B0	Gaussian	0.48	0.48	82.16	0.48	0.48
GTSRB	EfficientNet-B0	BadNets	89.15	89.75	82.17	65.63	88.59
SVHN	EfficientNet-B0	Clean	88.94	87.73	79.95	56.25	87.11
SVHN	EfficientNet-B0	Sign flipping	6.70	6.70	81.36	6.70	6.70
SVHN	EfficientNet-B0	Gaussian	6.70	6.70	83.06	6.70	6.70
SVHN	EfficientNet-B0	BadNets	90.32	87.56	79.03	59.22	88.13
MNIST	EfficientNet-B0	Clean	34.29	99.00	97.46	97.70	44.22
MNIST	EfficientNet-B0	Sign flipping	9.80	9.80	98.97	9.80	9.80
MNIST	EfficientNet-B0	Gaussian	9.80	9.80	98.42	9.80	9.80
MNIST	EfficientNet-B0	BadNets	30.15	98.67	98.61	97.65	50.01
CIFAR-10	EfficientNet-B0	Clean	75.39	75.44	53.71	45.89	72.31
CIFAR-10	EfficientNet-B0	Sign flipping	10.00	48.47	57.70	10.00	10.00
CIFAR-10	EfficientNet-B0	Gaussian	10.00	10.00	53.65	10.00	10.00
CIFAR-10	EfficientNet-B0	BadNets	76.64	73.73	54.01	47.16	72.68
CIFAR-100	EfficientNet-B0	Clean	44.44	41.24	27.51	15.79	33.15
CIFAR-100	EfficientNet-B0	Sign flipping	1.00	1.00	26.55	1.00	1.00
CIFAR-100	EfficientNet-B0	Gaussian	1.00	1.00	26.85	1.00	1.00
CIFAR-100	EfficientNet-B0	BadNets	42.48	41.37	24.22	14.00	36.05
GTSRB	ShuffleNetV2	Clean	70.42	69.56	53.67	16.09	67.05
GTSRB	ShuffleNetV2	Sign flipping	0.48	49.11	51.00	13.49	43.39
GTSRB	ShuffleNetV2	Gaussian	0.48	0.48	58.14	14.66	0.48
GTSRB	ShuffleNetV2	BadNets	67.69	67.09	52.14	14.86	64.99
SVHN	ShuffleNetV2	Clean	88.37	88.76	74.72	15.66	85.42
SVHN	ShuffleNetV2	Sign flipping	6.70	74.18	78.53	15.89	79.29
SVHN	ShuffleNetV2	Gaussian	6.70	6.70	76.39	6.70	6.70
SVHN	ShuffleNetV2	BadNets	87.34	87.59	76.74	15.53	85.88
MNIST	ShuffleNetV2	Clean	61.30	98.42	96.66	96.23	72.66
MNIST	ShuffleNetV2	Sign flipping	9.80	9.80	97.95	9.80	9.80
MNIST	ShuffleNetV2	Gaussian	9.80	9.80	97.92	9.80	9.80
MNIST	ShuffleNetV2	BadNets	63.23	98.25	96.22	96.23	51.51
CIFAR-10	ShuffleNetV2	Clean	70.70	66.81	51.46	30.56	64.31
CIFAR-10	ShuffleNetV2	Sign flipping	10.00	45.11	53.08	19.24	10.00
CIFAR-10	ShuffleNetV2	Gaussian	10.00	10.00	54.77	10.00	10.00
CIFAR-10	ShuffleNetV2	BadNets	69.23	65.86	49.03	29.98	65.83
CIFAR-100	ShuffleNetV2	Clean	2.72	35.58	16.61	10.46	6.69
CIFAR-100	ShuffleNetV2	Sign flipping	1.00	1.00	16.55	2.42	1.00
CIFAR-100	ShuffleNetV2	Gaussian	1.00	1.00	18.02	1.00	1.00
CIFAR-100	ShuffleNetV2	BadNets	6.77	35.21	17.61	10.83	4.67

S4. Extended condition-level descriptive statistics

Table S3 extends the main-paper condition summary by reporting mean, median, cross-configuration SD, IQR, minimum, maximum, mean rank, and fractional wins for every method and condition. These statistics serve different purposes. Mean and median summarize typical performance; SD and IQR show how strongly performance varies across tasks; range identifies extreme configurations; and rank/win counts describe relative ordering without assuming that raw accuracy scales are directly comparable across datasets.

The table reinforces three patterns. First, Trimmed Mean and FedAvg occupy the leading clean positions. Second, Krum becomes the dominant sign-flip and Gaussian method by mean rank and win count. Third, BadNets ordinary accuracy resembles clean training more closely than the model-poisoning conditions, which is why the trigger-specific analysis in Section S6 is

required.

Table S3 extends the condition-level comparison beyond mean accuracy by reporting the median, cross-configuration dispersion, range, within-task rank, and configuration-win count for every aggregation method. These statistics are calculated across the 25 dataset–architecture configurations within each experimental condition. Consequently, the reported SD and IQR quantify heterogeneity across benchmark tasks rather than run-to-run or seed-level uncertainty.

Table S3. Accuracy distribution and rank summaries by method and condition. Mean, median, minimum, and maximum values are final global test accuracies. SD and IQR characterize dispersion across the 25 heterogeneous dataset–architecture configurations and are not estimates of seed-level experimental uncertainty. Mean rank is calculated within each task, with rank 1 indicating the highest accuracy. Wins count task-level first-place finishes, with fractional credit assigned where applicable.

Condition	Method	Mean (%)	Median (%)	SD (pp)	IQR (pp)	Min (%)	Max (%)	Mean rank	Wins
Clean	FedAvg	70.92	75.39	24.78	28.79	2.72	99.17	1.82	14.5
Clean	Trimmed Mean	76.02	85.95	21.01	26.94	31.01	99.12	1.70	10.0
Clean	Krum	62.85	72.17	27.58	34.36	16.09	98.22	3.80	0.0
Clean	FLTrust	36.88	24.15	32.81	45.79	1.09	97.82	4.72	0.0
Clean	FedPARETO	68.93	72.66	24.61	34.46	6.69	99.12	2.96	0.5
Sign flipping	FedAvg	5.59	6.70	4.22	8.80	0.48	10.00	4.44	0.0
Sign flipping	Trimmed Mean	41.48	45.11	30.97	55.64	1.00	95.73	2.78	1.0
Sign flipping	Krum	63.91	72.37	27.98	34.39	15.84	98.97	1.32	19.0
Sign flipping	FLTrust	20.47	11.06	24.83	12.11	1.00	94.30	3.56	0.0
Sign flipping	FedPARETO	32.63	10.54	32.90	45.04	0.48	96.93	2.90	5.0
Gaussian	FedAvg	11.88	6.70	18.91	9.00	0.48	87.11	4.04	0.0
Gaussian	Trimmed Mean	23.14	9.80	29.56	24.58	0.48	97.74	3.14	1.0
Gaussian	Krum	64.45	72.70	27.10	33.82	16.29	98.75	1.08	24.0
Gaussian	FLTrust	20.82	10.00	27.23	11.44	0.48	97.83	3.02	0.0
Gaussian	FedPARETO	16.31	9.80	24.38	13.01	0.48	95.09	3.72	0.0
BadNets	FedAvg	70.51	76.64	24.79	30.07	6.77	99.10	1.80	14.0
BadNets	Trimmed Mean	75.17	85.29	21.49	28.53	28.45	99.17	1.80	8.0
BadNets	Krum	62.75	68.56	27.77	36.55	16.94	98.61	3.80	0.0
BadNets	FLTrust	35.99	21.99	32.79	48.39	0.48	97.86	4.72	0.0
BadNets	FedPARETO	68.29	72.68	25.05	37.01	4.67	99.11	2.88	3.0

S5. Matched comparisons and recurrent low-accuracy outcomes

A direct FedPARETO–Krum comparison is useful because these methods embody different aggregation principles: FedPARETO uses utility and reliability signals, whereas Krum primarily uses inter-update geometry. Table S4 reports matched win/tie/loss counts and both mean and median FedPARETO-minus-Krum accuracy differences over the same 25 dataset–architecture labels.

The sign of the matched difference changes sharply with condition. FedPARETO exceeds Krum in 22/25 clean tasks and 22/25 BadNets ordinary-accuracy tasks, but loses in 20/25 sign-flip tasks and 24/25 Gaussian tasks. Table S4 therefore supports a threat-dependent interpretation rather than a globally ordered comparison.

Table S4. FedPARETO versus Krum matched comparisons. Positive mean/median differences favour FedPARETO; negative differences favour Krum.

Condition	Wins	Ties	Losses	Mean diff. (pp)	Median diff. (pp)
Clean	22	0	3	6.08	9.28
Sign flipping	5	0	20	-31.28	-24.24
Gaussian	1	0	24	-48.15	-45.93
BadNets	22	0	3	5.54	9.13

A second diagnostic counts exact matches to recurrent dataset-specific low-accuracy values. Table S5 shows that FedAvg reaches the corresponding recurrent value in all 25 sign-flip configurations and 17 Gaussian configurations, whereas Krum reaches it in none. FedPARETO, Trimmed Mean, and FLTrust occupy intermediate positions. These counts describe repeated severe low-accuracy outcomes; they do not establish theoretical chance accuracy or a unique collapse mechanism.

Table S5. Exact matches to recurrent dataset-specific low-accuracy values under model poisoning. Counts are descriptive diagnostics over 25 configurations per method and condition.

Method	Sign flipping	Gaussian
FedAvg	25	17
Trimmed Mean	8	15
Krum	0	0
FLTrust	6	9
FedPARETO	12	16

S6. BadNets ordinary accuracy and triggered target-label behaviour

The BadNets analysis separates ordinary predictive performance from triggered target-label behaviour. Table S6 reports method-level means for global accuracy, TTLR, ECE, and worst-client accuracy. FedAvg and Trimmed Mean retain high ordinary accuracy while their mean TTLR remains close to 70%, demonstrating why ordinary test accuracy cannot be used as a substitute for targeted backdoor evaluation.

Table S6. BadNets ordinary accuracy, TTLR, ECE, and worst-client accuracy across 25 tasks. TTLR follows Eq. (S5) and should not be interpreted as conventional target-excluding ASR.

Method	Global accuracy (%)	Mean TTLR (%)	Mean ECE (%)	Mean worst-client accuracy (%)
FedAvg	70.51	72.68	5.79	68.57
Trimmed Mean	75.17	69.70	4.76	73.60
Krum	62.75	30.61	10.78	57.90
FLTrust	35.99	24.41	5.26	27.38
FedPARETO	68.29	60.54	5.97	65.79

Figure S1 gives the full 25-task TTLR distributions. The broad distributions show that the method-level means in Table S6 are not produced by one uniform backdoor response across all tasks. Krum and FLTrust have lower mean TTLR than FedAvg, Trimmed Mean, and FedPARETO, but the available evidence still does not permit conversion to conventional target-excluding ASR.

S7. Dataset-level accuracy, worst-client performance, and calibration

Figures S2–S5 decompose global accuracy by dataset and aggregation method after averaging over the five architectures. These heatmaps make visible which datasets contribute most strongly to the pooled condition-level means. The clean panel (Fig. S2) shows high MNIST and SVHN performance for several methods, whereas the sign-flip and Gaussian panels (Figs. S3 and S4) reveal the repeated advantage of Krum across multiple datasets. The BadNets ordinary-accuracy panel (Fig. S5) returns to a pattern much closer to clean training.

Figures S6 and S7 extend the analysis beyond global accuracy. Figure S6 shows mean worst-client accuracy by method and condition. Krum's sign-flip/Gaussian stability is visible in this lower-tail outcome as well as in global accuracy. Figure S7 summarizes ECE. The ECE heatmap is intentionally not converted into a "lower is always better" robustness ranking because calibration error can be difficult to interpret after severe accuracy collapse.

S8. Leave-one-dataset-out and leave-one-architecture-out sensitivity

The pooled benchmark contains 25 heterogeneous tasks. To test whether the principal poisoning ranking is driven by a single dataset or architecture, the mean ranks were recomputed after excluding one group at a time. Figures S8 and S9 summarize the average rank across sign-flipping and Gaussian conditions after each exclusion.

Figure S8 shows that removing any one dataset leaves Krum at or near the best average poisoning rank. Figure S9 gives the analogous architecture exclusion. The stability of the Krum-first ordering across these exclusions supports the conclusion that the headline model-poisoning result is not attributable to one dataset or one neural architecture alone. These analyses are sensitivity checks, not additional stochastic replication.

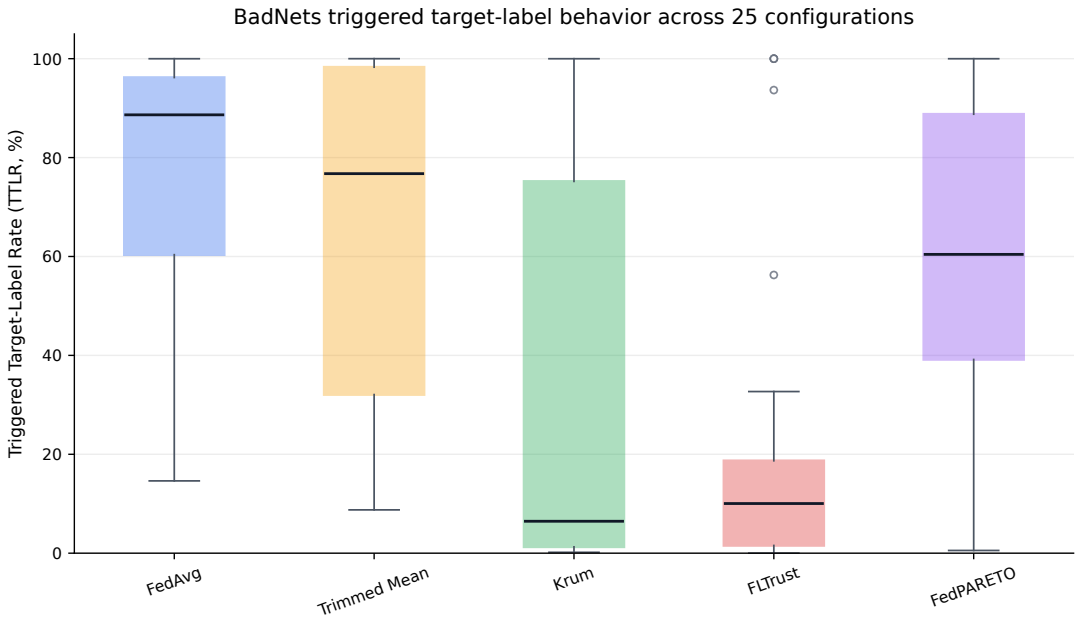

Figure S1. Distribution of BadNets Triggered Target-Label Rate (TTLR) across 25 dataset–architecture tasks for each method. Boxes summarize the task-level TTLR values retained by the source-code-audited metric. TTLR applies the trigger to all test inputs and is not conventional target-excluding ASR.

S9. Additional-seed evidence

The supplied archives contain a small set of genuine multi-seed experiments outside the balanced 500-cell canonical matrix. Seven legacy/scaffold MNIST families can be formed with seeds 1–3. Table S7 reports their mean, sample SD, minimum, and maximum accuracy. These values provide direct evidence that run-to-run variation can range from very small to several percentage points depending on the method and condition.

The crucial limitation is coverage: these families are sparse, MNIST-specific, and not uniformly aligned with the canonical benchmark. They are therefore not merged into the canonical SD or used to justify benchmark-wide confidence intervals. Figure S10 visualizes only these genuine three-seed families, with error bars representing seed-level SD within each family.

Table S7. Additional three-seed legacy/scaffold MNIST experiment families. SD is seed-level sample SD within the listed family only.

Family	Method	Condition	*n* seeds	Mean acc. (%)	SD (pp)	Min (%)	Max (%)
fedavg	FedAvg	Clean	3	98.42	0.08	98.36	98.51
fedpareto badnets	FedPARETO	BadNets	3	96.49	0.94	95.62	97.48
fedpareto	FedPARETO	Clean	3	96.08	0.84	95.36	97.01
fedpareto signflip	FedPARETO	Sign flipping	3	95.22	0.39	94.90	95.66
fltrust	FLTrust	Sign flipping	3	68.14	7.14	61.30	75.55
krum	Krum	Sign flipping	3	92.72	1.36	91.18	93.75
trimmed mean	Trimmed Mean	Gaussian	3	94.95	0.96	94.13	96.00

S10. FedPARETO ablation and anchor-size diagnostics

The FedPARETO ablation experiments are single-seed legacy/scaffold MNIST runs. Table S8 reports final accuracy, change from the 95.36% baseline, ECE, and worst-client accuracy. Removing calibration or the low-local-accuracy priority reduces accuracy modestly, removing the Pareto bonus changes little, and removing the robustness component increases final accuracy by 1.61 percentage points in this one run. The correct interpretation is therefore not that every component is beneficial, but that component effects remain uncertain under single-seed evidence.

Figure S11 visualizes the same single-seed ablation results with colours indicating whether the observed change is positive, moderately negative, or strongly negative relative to the baseline. The visualization is diagnostic only; it is not an estimate of causal component importance.

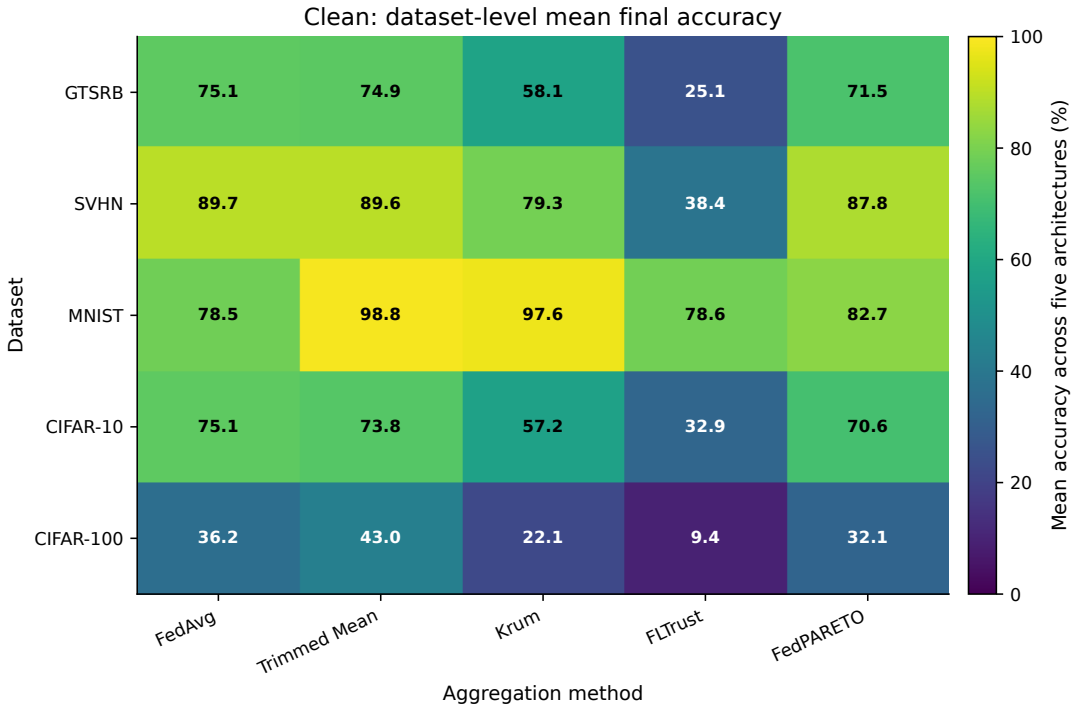

Figure S2. Dataset-by-method mean accuracy under clean training, averaged over the five architectures. Each cell is a descriptive mean over architectures; the heatmap does not represent seed-level uncertainty.

Table S8. Single-seed FedPARETO ablations relative to the 95.36% legacy/scaffold baseline. Positive differences indicate higher final accuracy than the baseline in that one run.

Variant	Final acc. (%)	Difference (pp)	ECE (%)	Worst-client acc. (%)
no_calibration	94.28	-1.08	1.74	84.25
no_fairness	94.64	-0.72	1.06	89.95
no_robustness	96.97	+1.61	1.15	93.18
no_pareto_bonus	95.27	-0.09	0.73	92.06

Anchor-size sensitivity is similarly non-monotonic. Table S9 and Fig. S12 show final accuracy at anchor sizes 16, 64, 128, and 512. Accuracy increases from 16 to 64, falls at 128, and returns near the baseline at 512. The supplied evidence therefore does not support a monotonic scaling law between anchor size and predictive accuracy.

S11. Provenance composition and cross-condition rank similarity

Figure S13 complements Table S1 by showing the numerical dominance of original successful logs within the canonical evidence base. The repaired/rerun subset is small but non-negligible, and the summary-only subset contains 10 cells. The figure is included because provenance strength is a property of the evidence, not an outcome metric, and should remain visible when interpreting reproducibility.

Figure S14 summarizes Spearman correlations between the five-method condition-level orderings. Clean and BadNets ordinary-accuracy rankings are strongly aligned, whereas Gaussian has a negative correlation with clean. Sign flipping lies between these regimes. Because each correlation is computed from only five aggregate method ranks, the values are treated as compact descriptive summaries rather than precise population estimates.

S12. Verified configuration scope and software environment

The repaired 36-run subset contains explicit configuration files and therefore provides the strongest evidence for detailed hyperparameters. Table S10 summarizes settings that are consistent across that subset. These values are not automatically

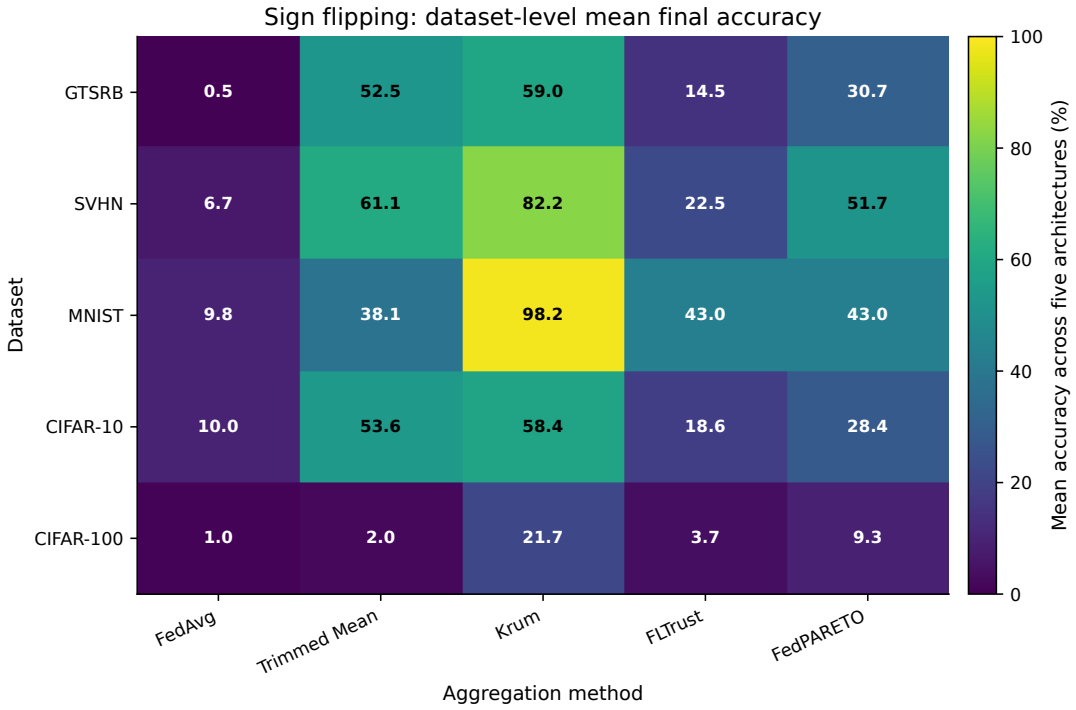

Figure S3. Dataset-by-method mean accuracy under sign-flipping poisoning, averaged over the five architectures. Krum remains high across several datasets while multiple alternative methods exhibit severe reductions.

Table S9. Single-seed anchor-size sensitivity relative to the 512-sample baseline.

Anchor size	Final acc. (%)	Difference (pp)	ECE (%)	Worst-client acc. (%)
16	92.98	-2.38	1.89	84.57
64	95.49	+0.13	1.74	90.24
128	92.19	-3.17	2.39	80.30
512	95.36	+0.00	1.06	92.06

generalized to the remaining canonical cells because complete configuration snapshots are not available for every historical run.

The software evidence also differs across repair attempts. The final repair bundle dated 22 July 2026 records Python 3.10.13, PyTorch 2.5.1, CUDA availability, and NVIDIA L40S hardware. An earlier repair attempt dated 21 July 2026 records Python 3.13.5 and PyTorch 2.8.0+cu128 on NVIDIA L40S. These differences do not invalidate the retained final accuracy records, but they make direct runtime comparisons across repair environments uncontrolled. The supplementary analysis therefore does not treat heterogeneous runtime values as evidence of method-level latency superiority.

S13. FedPARETO implementation audit

Table S11 distinguishes mechanisms present in the executable FedPARETO scaffold from mechanisms appearing only in the broader conceptual draft. This distinction is central to the paper's evidence discipline: empirical benchmark results can only be attributed to mechanisms that were actually implemented in the supplied executable path.

The audit shows that the executable method contains anchor accuracy, negative ECE, a low-local-accuracy priority proxy, root-direction similarity, temporal reliability, per-round min–max normalization, dominance-count ranking, scalar utility, exponential transformation, simplex projection, and uniform mixing. It does not implement the conceptual draft's weighted-Tchebycheff solver, explicit ℓ_1 trust region, reliability-dependent upper weight caps, benign-subspace SVD, or a formal constrained Pareto optimization problem.

A separate implementation observation concerns update–utility correspondence. In the audited sign-flip/Gaussian client path, the local model and clean delta are formed, the submitted delta can then be corrupted, and predictive summaries are still computed from the uncorrupted locally trained model. The server can therefore score predictive evidence corresponding to

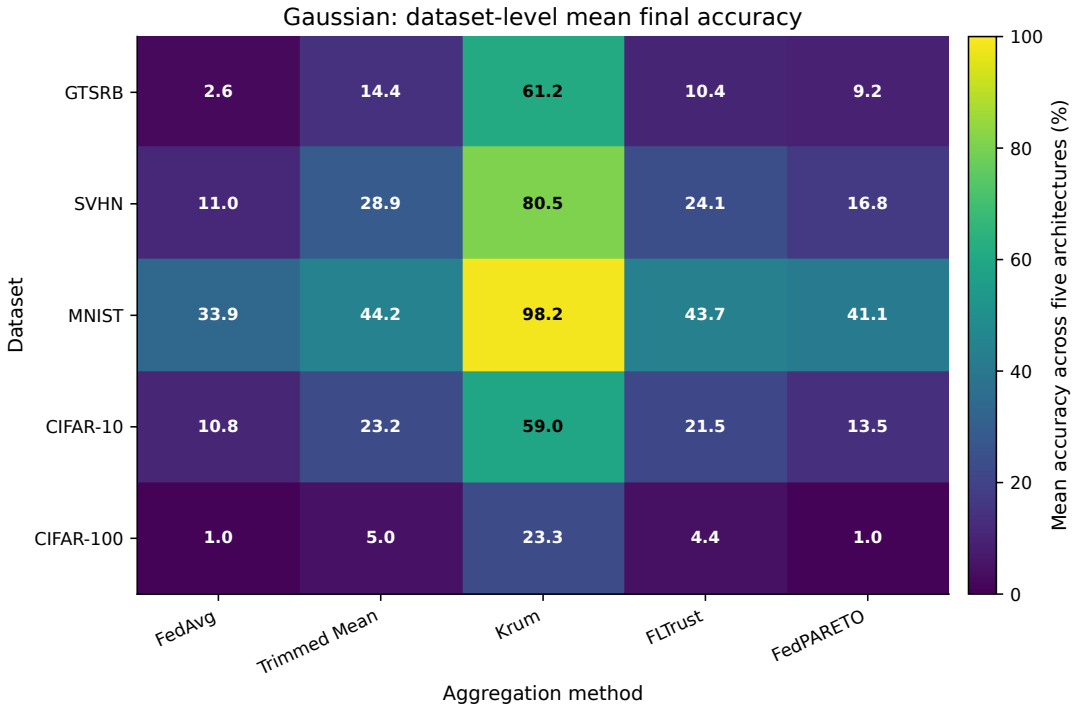

Figure S4. Dataset-by-method mean accuracy under Gaussian poisoning, averaged over the five architectures. The Krum-first pattern is visible across datasets rather than being produced by one isolated task family.

one model object while assigning influence to a different submitted update. This is a verified source-code observation for the supplied scaffold. It is not treated as causal proof for every canonical numerical cell because exact immutable code-to-result provenance is incomplete.

S14. Claim-to-evidence audit and reproducibility resources

The claim-to-evidence ledger provides a final guard against overstatement. Table S12 reproduces the major audit decisions in compact form. Claims supported directly by empirical records or source-code inspection are retained with restricted wording; unsupported generalizations—including universal Krum robustness, uniform multi-seed replication, guaranteed benefit from every FedPARETO component, and experimentally validated FedPARETO-UC improvement—are not made.

The complete machine-readable ledger is supplied as `audit/claim_evidence_matrix.csv`; the implementation audit is supplied as `audit/code_audit.md`; and unresolved provenance questions are retained in `audit/unresolved_questions.md`. The source-data directory contains the canonical results, condition summaries, pairwise comparisons, floor counts, BadNets metrics, leave-one-group-out tables, additional seeds, ablations, anchor-size results, condition-rank correlations, and provenance labels.

Together, these resources make the supplementary analysis reproducible from the packaged evidence without requiring the reader to infer missing run settings from source-code defaults. The principal unresolved limitations remain balanced multi-seed replication of the canonical matrix, complete immutable code/configuration provenance for every cell, and conventional target-excluding ASR for the retained BadNets runs.

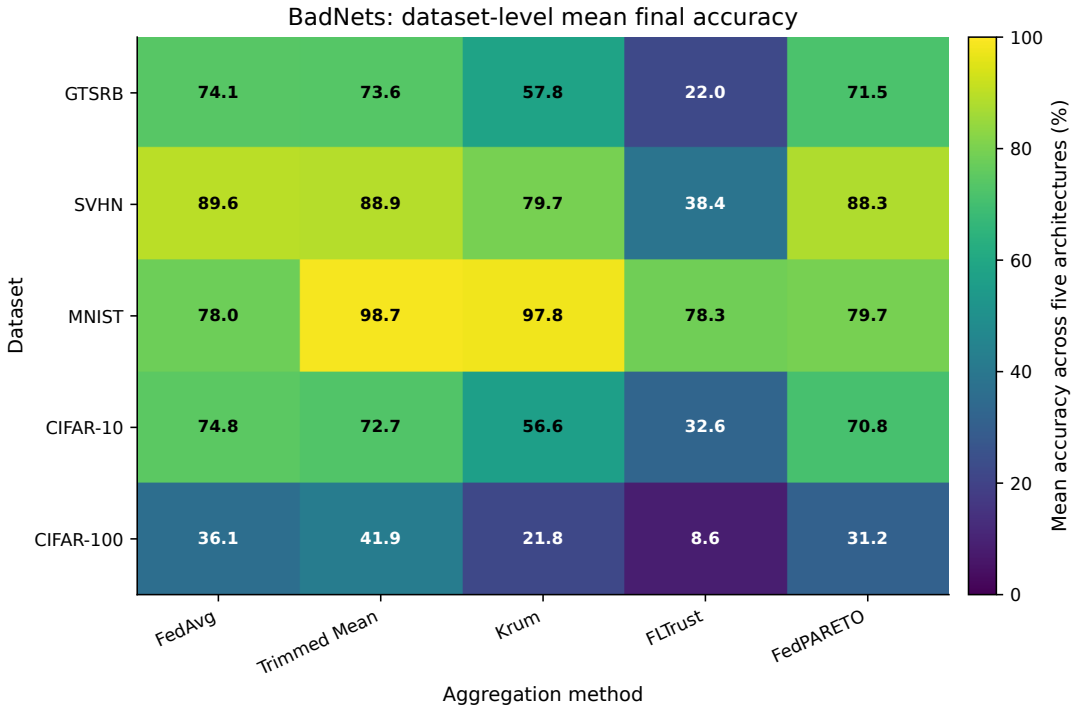

Figure S5. Dataset-by-method mean ordinary accuracy under BadNets, averaged over the five architectures. The clean-like ordinary-accuracy pattern does not establish targeted backdoor resistance; TTLR is analysed separately in Section S6.

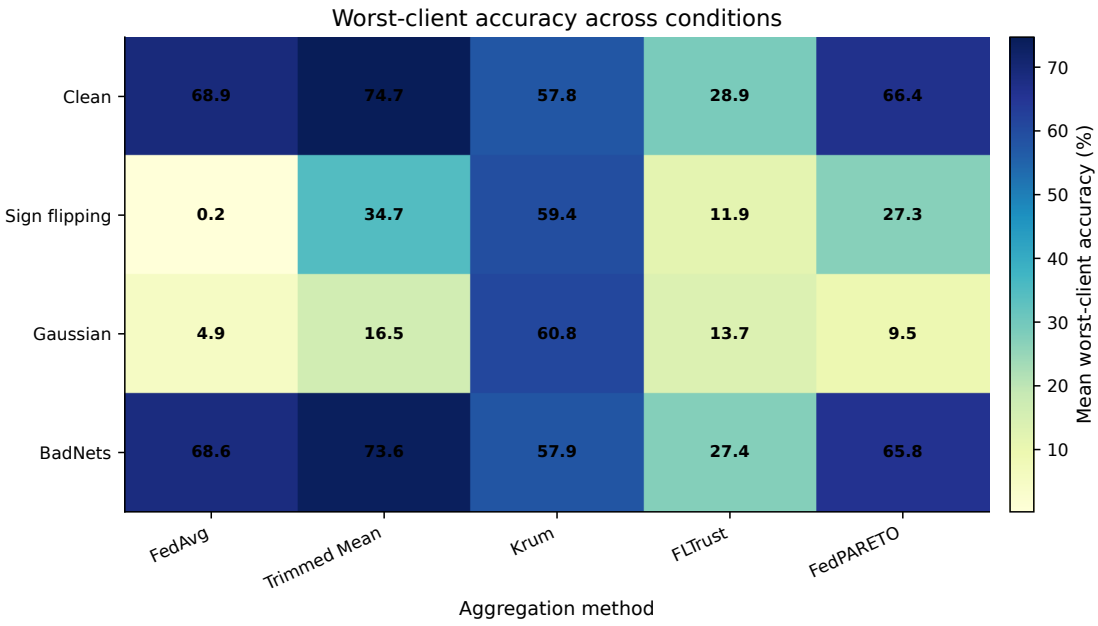

Figure S6. Mean final worst-client accuracy by method and condition across the 25 dataset–architecture configurations. This lower-tail outcome complements global accuracy and should not be interpreted as a complete fairness metric.

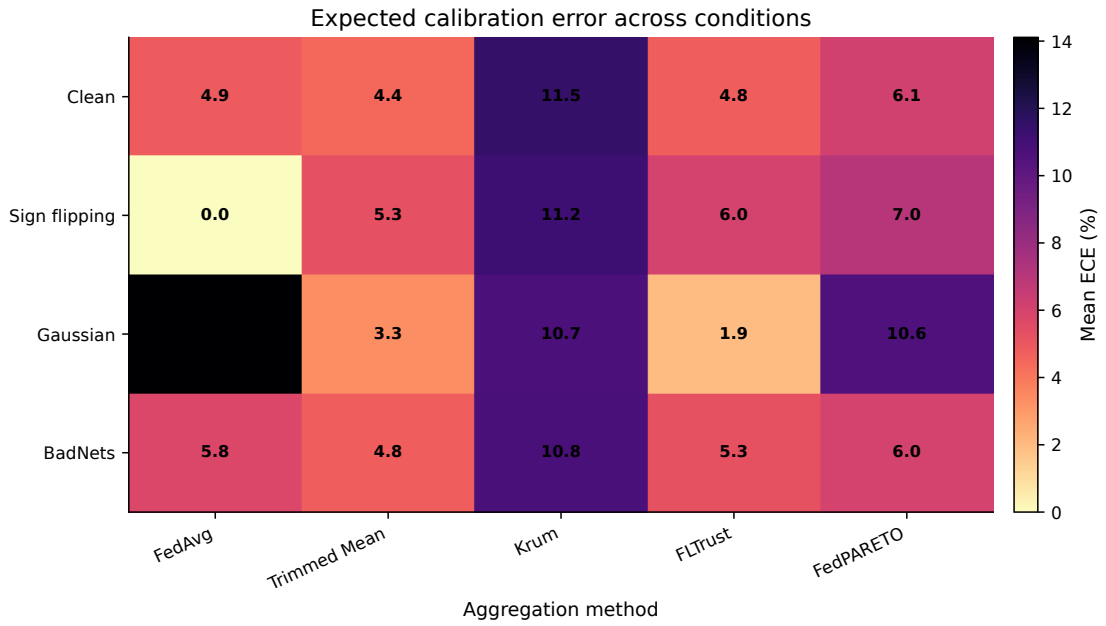

Figure S7. Mean final expected calibration error (ECE) by method and condition across the 25 configurations. ECE is reported descriptively; a low ECE value in a severely collapsed classifier is not automatically evidence of desirable performance.

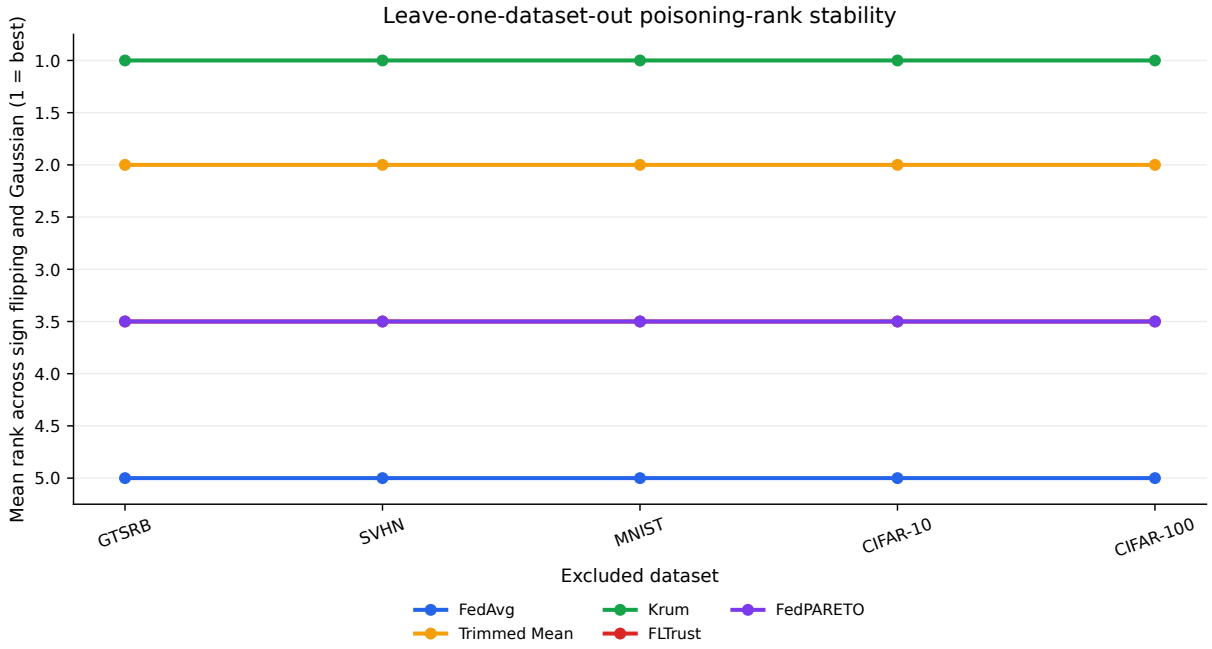

Figure S8. Leave-one-dataset-out poisoning-rank stability. Curves report the average rank across sign flipping and Gaussian poisoning after excluding one dataset. Lower rank is better. The analysis tests benchmark-composition sensitivity rather than inferential generalization.

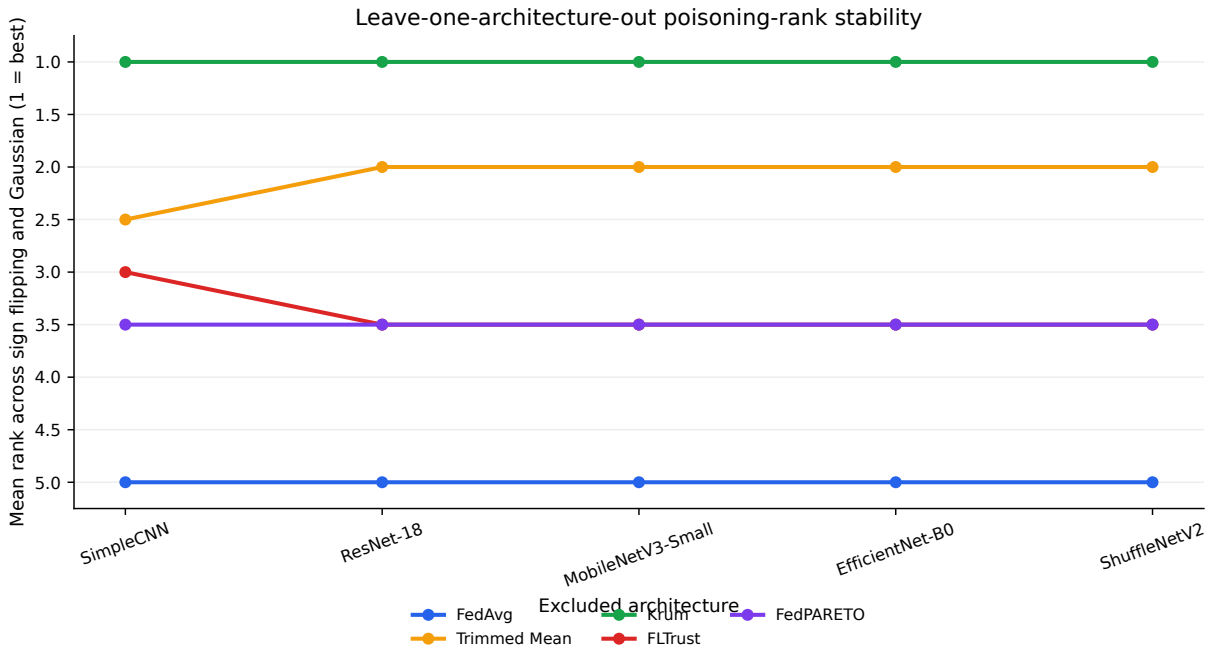

Figure S9. Leave-one-architecture-out poisoning-rank stability. Curves report the average rank across sign flipping and Gaussian poisoning after excluding one architecture. The Krum-first ordering remains visible across the tested exclusions.

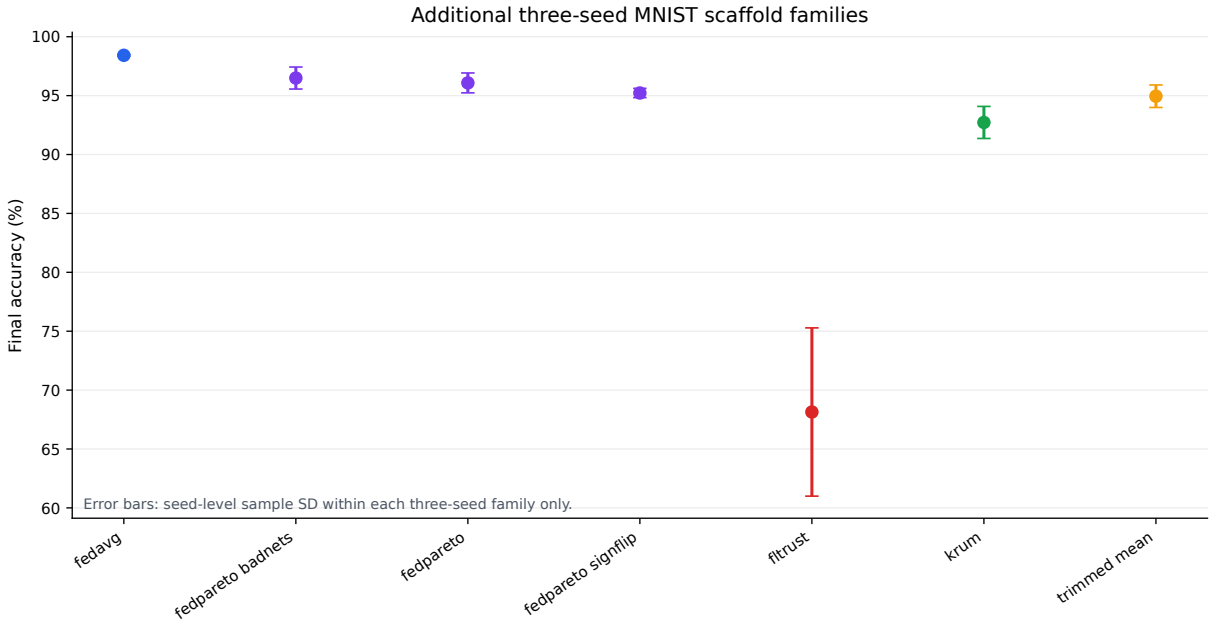

Figure S10. Additional-seed accuracy for the available legacy/scaffold MNIST families. Points show three-seed means and error bars show seed-level sample SD. These experiments are intentionally separated from the canonical matrix because replication is sparse and unbalanced.

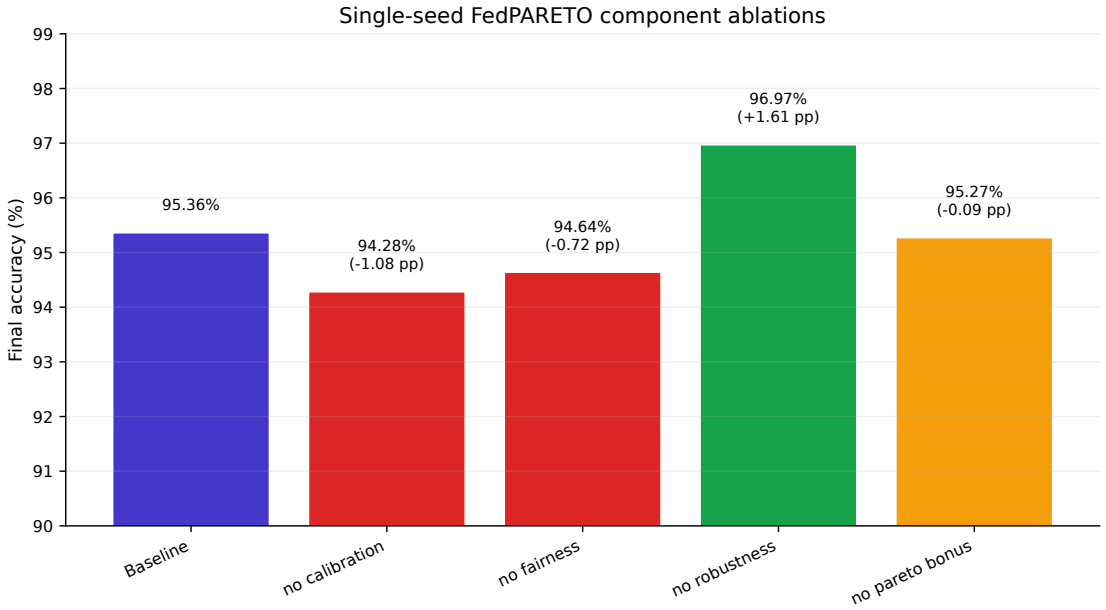

Figure S11. Single-seed FedPARETO component ablation. Labels show final accuracy and the difference from the corresponding baseline. No inferential uncertainty is available.

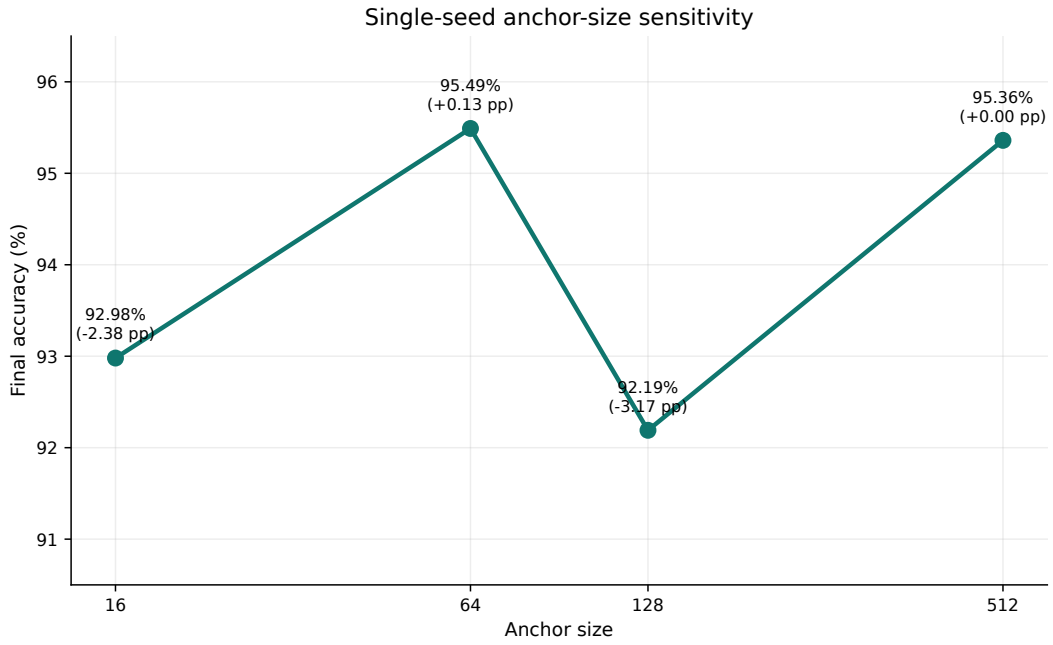

Figure S12. Single-seed anchor-size sensitivity. Annotated points report final accuracy and the difference from the 512-sample baseline. The pattern is non-monotonic and should not be interpreted as a fitted scaling relationship.

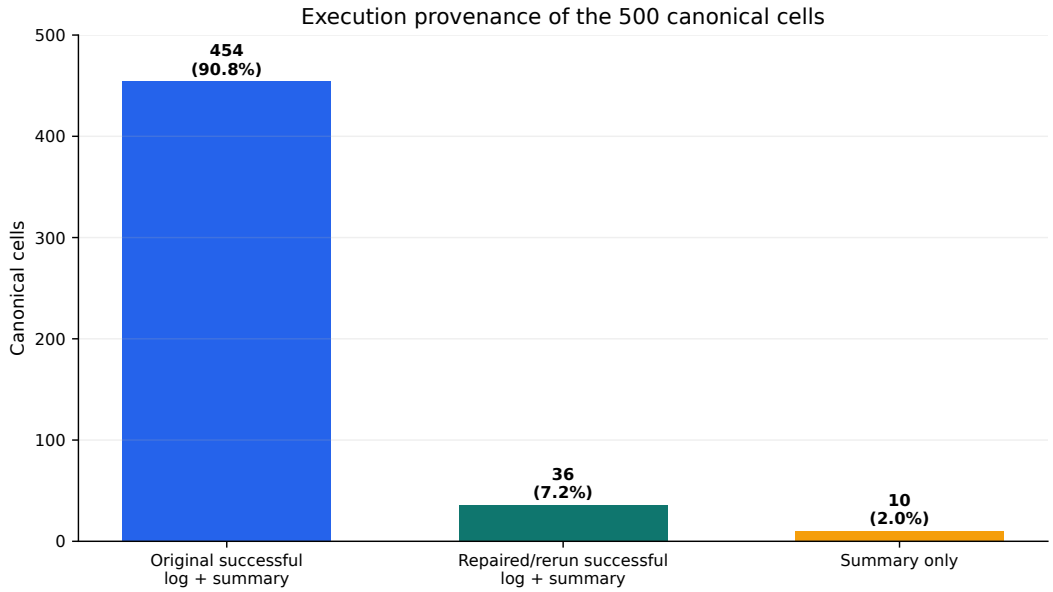

Figure S13. Execution provenance of the 500 canonical cells. Colours distinguish original successful logs, successful repaired/rerun logs, and summary-only evidence.

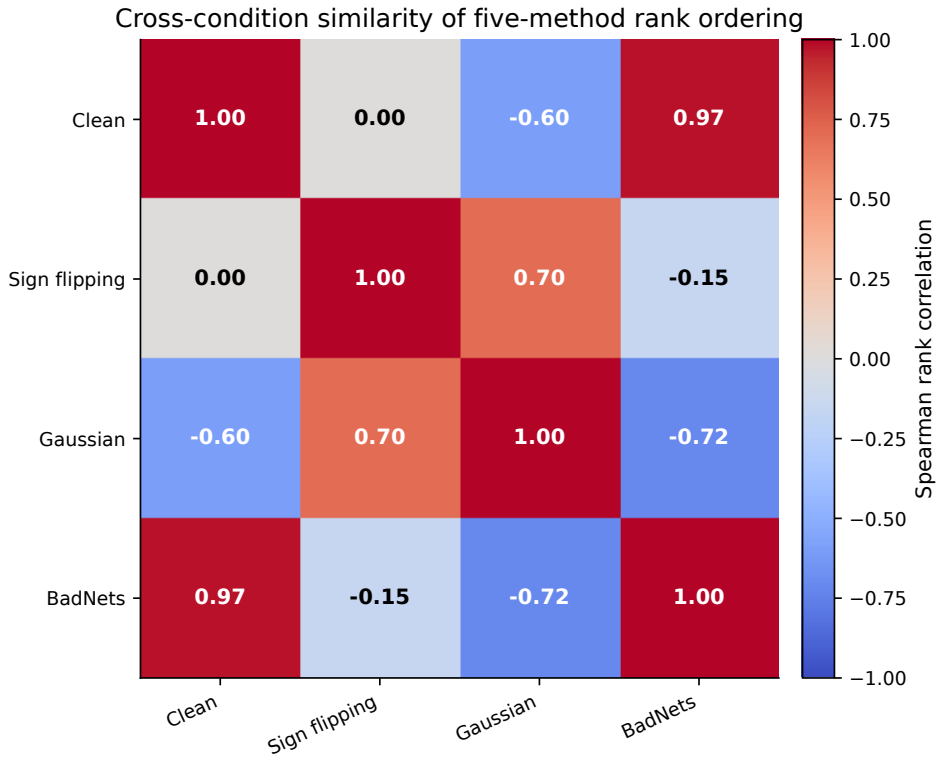

Figure S14. Cross-condition similarity of the five-method rank ordering. Values are Spearman rank correlations between condition-level method orderings and are descriptive because only five methods contribute to each correlation.

Table S10. Configuration parameters verified for the repaired 36-run subset. "Subset-verified" means directly supported for those repaired runs but not assumed benchmark-wide.

Parameter	Verified value	Scope
Number of clients	20	Repaired 36-run subset
Clients sampled per round	10	Repaired 36-run subset
Communication rounds	500	Repaired 36-run subset
Local epochs	1	Repaired 36-run subset
Batch size	64	Repaired 36-run subset
Optimizer settings	SGD; lr 0.003; momentum 0.9; weight decay 5×10^{-4}	Repaired 36-run subset
Dirichlet α	1.0	Repaired 36-run subset
Malicious-client fraction	0.25	Attacked repaired runs
Gaussian standard deviation	0.5	Gaussian repaired runs
BadNets poison fraction	0.3	BadNets repaired runs
BadNets target label	0	BadNets repaired runs
BadNets trigger size	4×4	BadNets repaired runs
Anchor size	4096	Repaired 36-run subset
ECE bins	15	Repaired 36-run subset

Table S11. Implemented-versus-conceptual FedPARETO mechanism audit. Green rows denote implemented mechanisms; red rows denote mechanisms not found in the supplied executable scaffold.

Mechanism	Status	Evidence-bounded interpretation
Anchor accuracy	Implemented	Absolute anchor accuracy of local model
Negative expected calibration error	Implemented	ECE of anchor predictions
Low-local-accuracy priority	Implemented	1 - submitting client local accuracy; not a measured fairness improvement
Root-direction similarity	Implemented	Non-negative cosine to server root update
Temporal reliability	Implemented	Momentum state updated from weights and robustness score
Per-round min-max normalization	Implemented	Each objective normalized among participating clients
Dominance-count ranking	Implemented	1 + number of clients that dominate candidate; not iterative Pareto fronts
Weighted scalar utility	Implemented	Configured objective weights plus rank bonus and trust penalty
Exponential transformation	Implemented	exp(temperature * scalar)
Simplex projection	Implemented	Euclidean projection utility
Uniform mixing	Implemented	entropy_reg mixes projected weights with uniform weights
Weighted Tchebycheff optimization	Not implemented	Appears in conceptual draft only
Explicit L1 trust region around sample-size weights	Not implemented	Conceptual draft only
Reliability-dependent weight caps	Not implemented	Conceptual draft only
Benign-subspace SVD	Not implemented	Conceptual draft only
Formal Pareto constrained optimization	Not implemented	Executable method is Pareto-inspired heuristic

Table S12. Claim-to-evidence audit. Classification A denotes direct empirical observation, B a derived descriptive statistic, C a verified source-code observation, and G an unsupported claim removed or corrected.

Class	Claim	Evidence-bounded verdict/correction
A	500/500 canonical final-accuracy records are present	Use as full benchmark matrix; distinguish 490 successful-log cells from 10 summary-only cells
A	490 canonical cells have a successful execution log	10 clean SVHN MobileNetV3-Small/ShuffleNetV2 method cells remain summary-only
B	Krum is strongest by accuracy under evaluated sign-flip and Gaussian conditions	Restrict claim to evaluated conditions; no universal robustness claim
B	BadNets ordinary accuracy remains close to clean values	Does not imply backdoor robustness
C	The recorded attack_success_rate is TTLR rather than conventional target-excluding ASR	Report as TTLR; conventional ASR not estimable
C	FedPARETO implementation uses anchor accuracy, negative ECE, low-local-accuracy priority, root-direction/temporal reliability, min-max normalization, dominance count, scalar score, exponential transform, simplex projection, uniform mixing	Describe executable mechanism only
C	FedPARETO scaffold has update-utility mismatch for post-training sign-flip/Gaussian corruption	Implementation observation; not causal proof for every workbook cell
G	FedPARETO-UC improves robustness	Remove performance claim; describe only as proposed future method
G	Cross-task SD is seed uncertainty	Call it cross-configuration dispersion only
G	Additional seeds establish uniform replication of canonical benchmark	Analyze separately, do not merge
G	Removing every FedPARETO component degrades accuracy	Replace with descriptive ablation values; no general causal conclusion