Poisoning with A Pill: Circumventing Detection in Federated Learning

AAAI-26 / IAAI-26 / EAAI-26
JANUARY 20–27, 2026 | SINGAPORE

Hanxi Guo[1], Hao Wang[2], Tao Song[3], Tianhang Zheng[4], Yang Hua[5], Haibing Guan[3], Xiangyu Zhang[1]

Purdue University[1], Stevens Institute of Technology[2], Shanghai Jiao Tong University[3], Zhejiang University[4], Queen's University Belfast[5]

1. Motivation

The Attacker's Insight

Model parameters have unequal impact. Altering redundant parameters reduces both attack stealthiness and effectiveness.

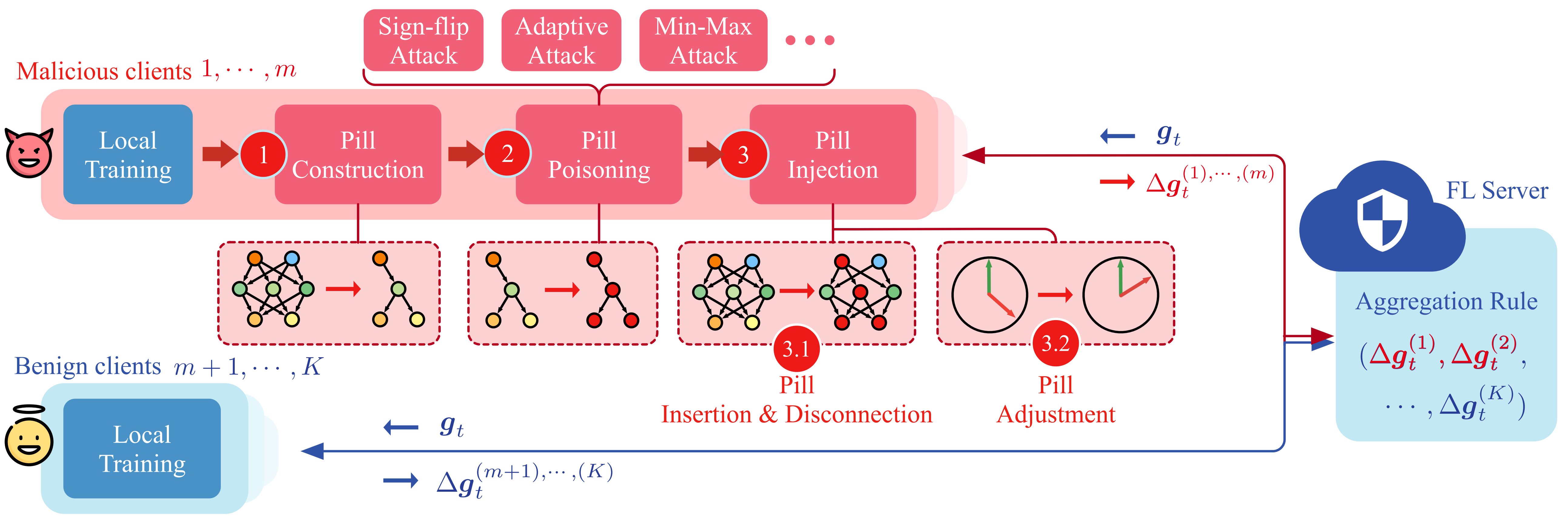
The Defender's Oversite

Existing methods focus on the overall statistics of client updates, ignoring attacks on specific parameters.

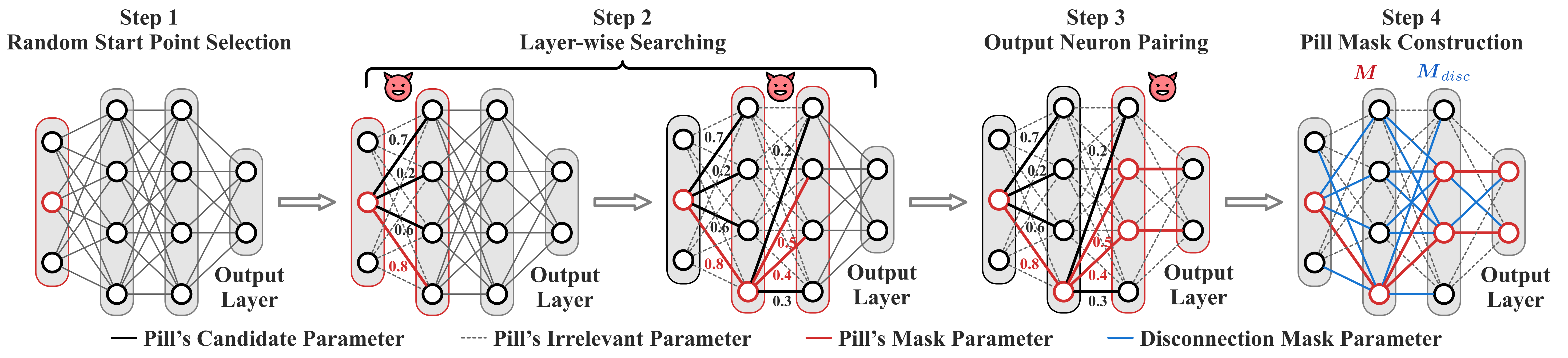
Our Poisoned Pill Approach

We propose an attack-agnostic augmentation that injects poison only into a critical subnet, bypassing current defenses and underscoring the urgent need for more robust and fine-grained detection mechanisms

2. Overview of Our Method


Sign-flip Attack
Adaptive Attack
Min-Max Attack
• • •
Malicious clients $1, \cdots, m$
Local Training
1
Pill Construction
2
Pill Poisoning
3
Pill Injection
$\boldsymbol{g}_t$
$\Delta \boldsymbol{g}_t^{(1),\cdots,(m)}$
FL Server
Aggregation Rule
$(\Delta \boldsymbol{g}_t^{(1)}, \Delta \boldsymbol{g}_t^{(2)}, \cdots, \Delta \boldsymbol{g}_t^{(K)})$
3.1
Pill Insertion & Disconnection
3.2
Pill Adjustment
Benign clients $m+1, \cdots, K$
Local Training
$\boldsymbol{g}_t$
$\Delta \boldsymbol{g}_t^{(m+1),\cdots,(K)}$

3. Pill Construction

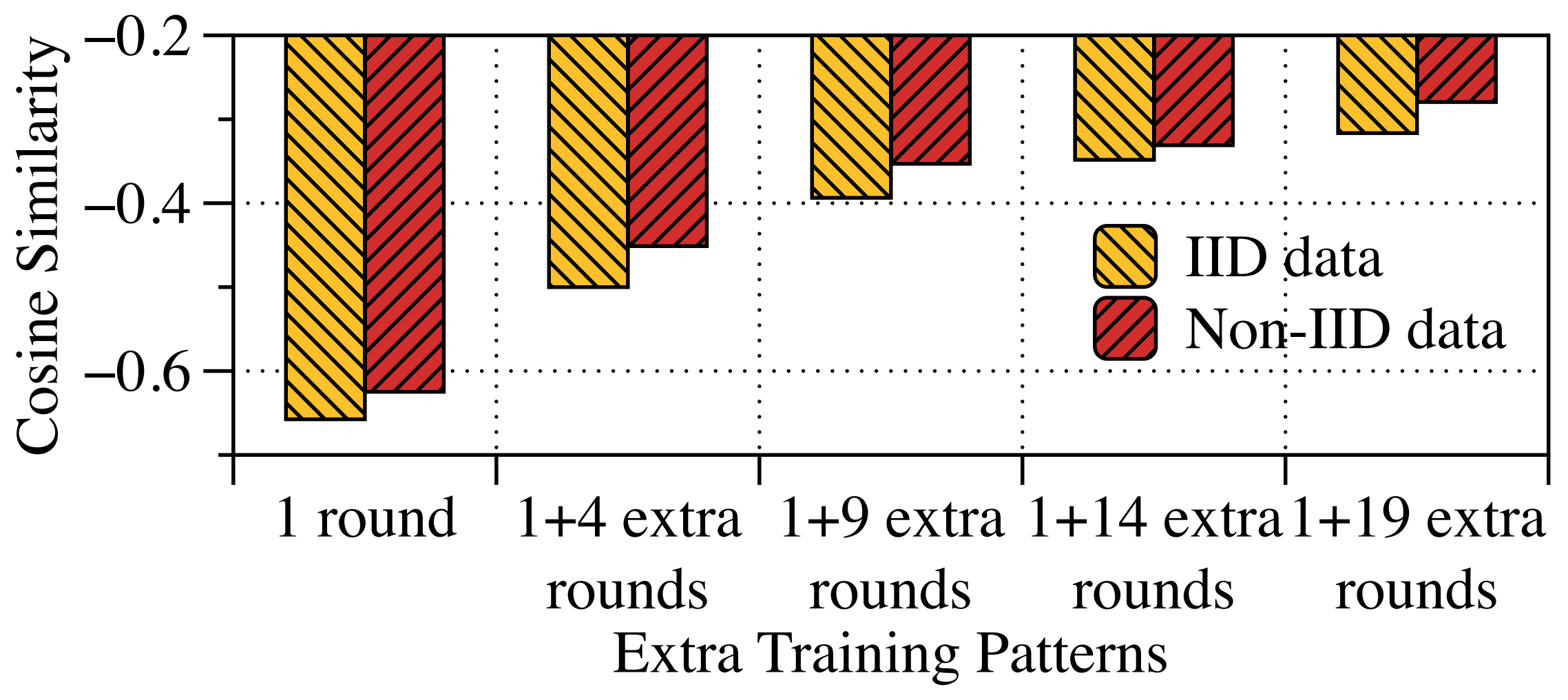


Pill Blueprint Design

⊕

Dynamic Pill Search

4. Pill Poisoning



We reuse existing attacks without any intrusive modification but replacing the reference model update with an extra-trained one.

5. Pill Injection

Algorithm 2: Similarity-based and distance-based adjustment functions in the Poison Pill Injection stage.

```
function SimAdjust ($param, \Delta\widetilde{g}_{t+1}, \Delta g^{(i)}_{t+1}$)
    $\{\Delta g'^{(1),\cdots,(m)}_{t+1}, M_{all}\} \leftarrow param$;
    $S_{max} \leftarrow max(0, max\{\mathtt{Sim}(\Delta\widetilde{g}_{t+1}, \Delta g^{(i)\prime}_{t+1}); i \in \{1,\cdots,m\}\})$;
    $iter \leftarrow 0$;
    while $\mathtt{Sim}(\Delta\widetilde{g}_{t+1}, \Delta g^{(i)}_{t+1}) < S_{max}$ && $iter < C_{iter}$ do
        if $iter\,\%2$ then
            $\Delta g^{(i)}_{t+1} \leftarrow (C_{\uparrow} \cdot (\mathbf{1} - M_{all}) + M_{all}) \odot \Delta g^{(i)}_{t+1}$;
            $iter \leftarrow iter + 1$;
        else
            $\Delta g^{(i)}_{t+1} \leftarrow ((\mathbf{1} - M_{all}) + C_{\downarrow} \cdot M_{all}) \odot \Delta g^{(i)}_{t+1}$;
            $iter \leftarrow iter + 1$;
    return $\Delta g^{(i)}_{t+1}$;

function DistAdjust ($param, \Delta\widetilde{g}_{t+1}, \Delta g^{(i)}_{t+1}$)
    $\{\Delta g'^{(1),\cdots,(m)}_{t+1}, M_{all}\} \leftarrow param$;
    $Dist_{max} \leftarrow max\{||\Delta g'^{(i)}_{t+1} - \Delta\widetilde{g}_{t+1}||; i \in \{1,\cdots,m\}\}$;
    $Dist \leftarrow ||\Delta g^{(i)}_{t+1} - \Delta\widetilde{g}_{t+1}||$;
    if $||C_{\downarrow} \cdot \Delta g^{(i)}_{t+1} - \Delta\widetilde{g}_{t+1}|| < ||C_{\uparrow} \cdot \Delta g^{(i)}_{t+1} - \Delta\widetilde{g}_{t+1}||$ then
        $C_{dist} \leftarrow C_{\downarrow}$;
    else
        $C_{dist} \leftarrow C_{\uparrow}$;
    while $Dist \geq Dist_{max}$ && $||C_{dist} \cdot \Delta g^{(i)}_{t+1} - \Delta\widetilde{g}_{t+1}|| \leq Dist$ do
        $\Delta g^{(i)}_{t+1} \leftarrow C_{dist} \cdot \Delta g^{(i)}_{t+1}$;
        $Dist \leftarrow ||\Delta g^{(i)}_{t+1} - ||$;
    return $\Delta g^{(i)}_{t+1}$;
```

6. Evaluation Results

Effectiveness, Compatibility and Generalizability

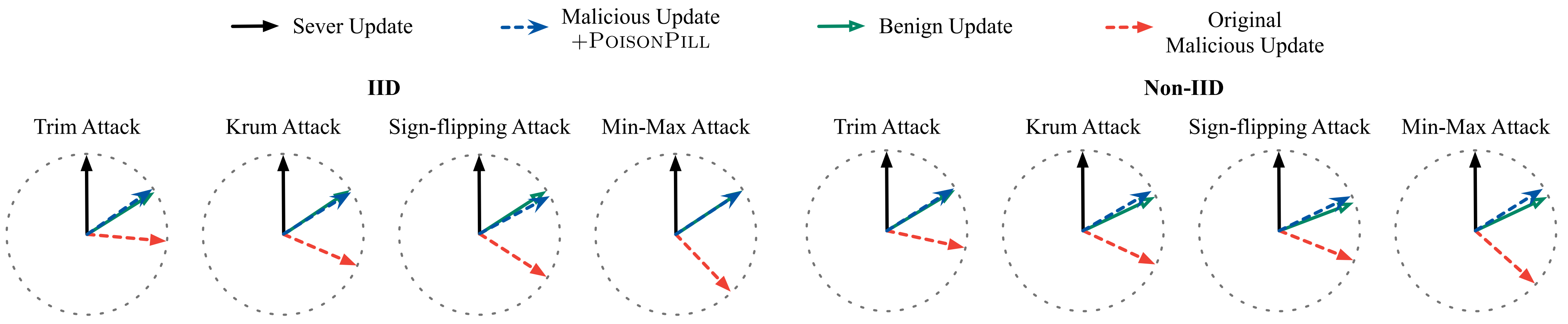
Fashion-MNIST, 50-client, 20% Malicious

Data Distribution	IID							Non-IID						
Attack	**FedAvg**	**FLTrust**	**MKrum**	**Bulyan**	**Median**	**Trim**	**FLD**	**FedAvg**	**FLTrust**	**MKrum**	**Bulyan**	**Median**	**Trim**	**FLD**
No Attack	0.109	0.107	0.105	0.105	0.123	0.106	0.115	0.113	0.115	0.115	0.112	0.142	0.115	0.122
Sign-Flipping	**0.943**	0.114	0.108	0.126	0.136	0.116	0.118	**0.917**	**0.126**	0.117	0.132	0.152	0.124	0.127
+ Poison Pill	0.667	**0.115**	**0.764**	**0.379**	**0.523**	**0.314**	**0.646**	0.543	0.122	**0.754**	**0.430**	**0.522**	**0.311**	**0.688**
Trim Attack	0.243	0.109	0.139	0.146	0.174	0.179	**0.116**	0.332	0.120	0.201	0.163	0.231	**0.238**	0.124
+ Poison Pill	**0.618**	**0.576**	**0.638**	**0.284**	**0.453**	**0.219**	0.115	**0.668**	**0.517**	**0.687**	**0.292**	**0.473**	0.223	**0.222**
Krum Attack	0.116	0.109	0.189	0.201	0.172	0.137	**0.786**	0.128	0.116	0.235	0.276	0.217	0.160	**0.947**
+ Poison Pill	**0.735**	**0.155**	**0.715**	**0.422**	**0.578**	**0.310**	0.637	**0.716**	**0.151**	**0.737**	**0.468**	**0.730**	**0.334**	0.690
Min-Max Attack	0.183	0.110	0.431	**0.330**	0.183	0.218	**0.825**	0.269	0.125	**0.619**	**0.434**	0.255	0.278	**0.831**
+ Poison Pill	**0.702**	**0.303**	**0.668**	0.327	**0.514**	**0.314**	0.778	**0.629**	**0.320**	0.612	0.406	**0.547**	**0.376**	0.822

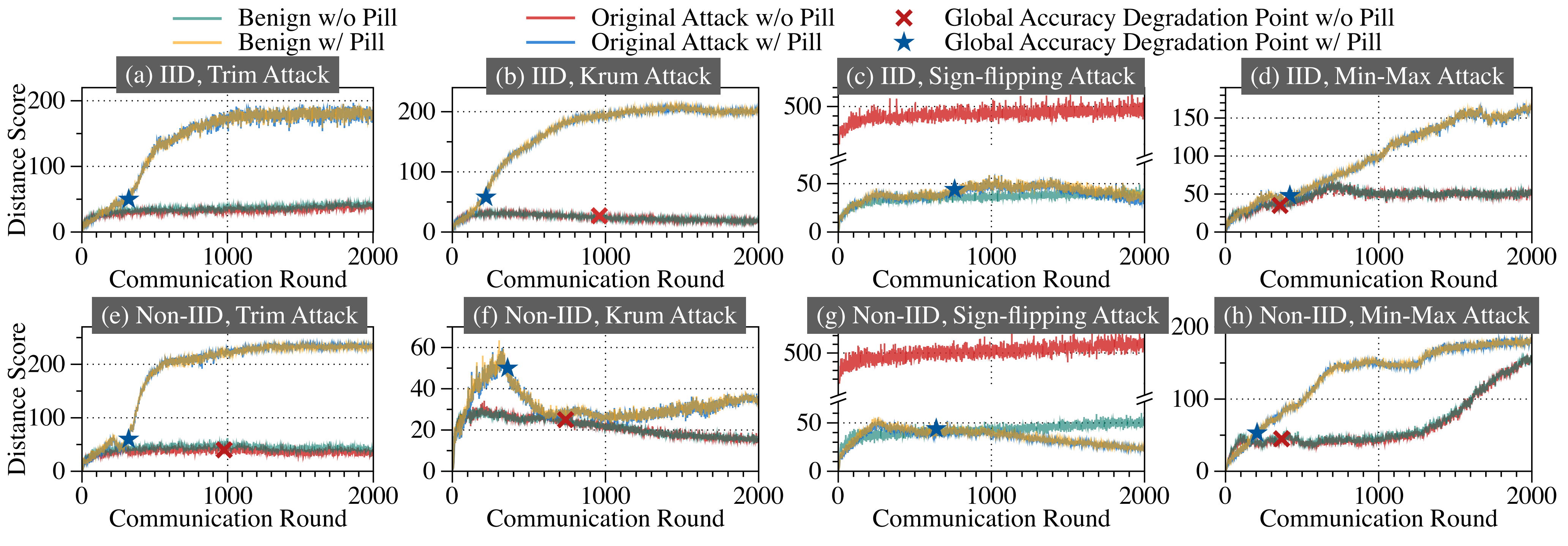
CIFAR-10, 30-client, 20% Malicious

Distribution	IID									Non-IID								
Attack	**FAvg**	**FLT**	**MKr**	**Bulyan**	**Med**	**Trim**	**DnC**	**FLD**	**Flame**	**FAvg**	**FLT**	**MKr**	**Bulyan**	**Med**	**Trim**	**DnC**	**FLD**	**Flame**
No Attack	0.48	0.48	0.50	0.46	0.55	0.45	0.44	0.49	0.49	0.48	0.47	0.49	0.49	0.58	0.52	0.46	0.50	0.53
Sign-Flipping	**0.89**	0.47	0.58	0.53	0.62	0.46	0.46	0.49	0.50	**0.90**	0.51	0.51	0.62	0.65	0.57	0.50	0.60	0.53
+ Poison Pill	0.73	**0.88**	**0.92**	**0.69**	**0.70**	**0.69**	**0.53**	**0.89**	**0.70**	0.87	**0.86**	**0.89**	**0.67**	**0.76**	**0.68**	**0.56**	**0.90**	**0.67**
Trim ATK	0.48	0.50	0.48	0.53	0.62	0.51	0.45	0.45	0.50	0.57	0.49	0.60	0.59	0.65	0.54	0.48	0.48	0.50
+ Poison Pill	**0.85**	**0.87**	**0.88**	**0.65**	**0.67**	**0.66**	**0.51**	**0.89**	**0.54**	**0.89**	**0.86**	**0.90**	**0.77**	**0.68**	**0.63**	**0.51**	**0.89**	**0.62**
Krum ATK	0.47	0.54	0.47	0.56	0.54	0.51	0.45	0.80	0.50	0.48	0.50	0.49	0.52	0.64	0.51	0.48	**0.89**	0.50
+ Poison Pill	**0.70**	**0.89**	**0.90**	**0.76**	**0.75**	**0.64**	**0.52**	**0.89**	**0.87**	**0.72**	**0.84**	**0.90**	**0.67**	**0.74**	**0.64**	**0.58**	0.88	**0.87**
Min-Max ATK	0.45	0.50	0.46	0.50	0.57	0.46	0.51	0.52	0.52	0.47	0.50	0.49	0.56	0.63	0.60	0.47	0.48	0.48
+ Poison Pill	**0.75**	**0.71**	**0.90**	**0.77**	**0.80**	**0.64**	**0.54**	**0.90**	**0.81**	**0.66**	**0.64**	**0.88**	**0.67**	**0.78**	**0.66**	**0.52**	**0.90**	**0.79**

Stealthiness - Cosine Similarity Score Comparison



Stealthiness - Distance Score Comparison


Benign w/o Pill
Benign w/ Pill
Original Attack w/o Pill
Original Attack w/ Pill
Global Accuracy Degradation Point w/o Pill
Global Accuracy Degradation Point w/ Pill
(a) IID, Trim Attack
(b) IID, Krum Attack
(c) IID, Sign-flipping Attack
(d) IID, Min-Max Attack
(e) Non-IID, Trim Attack
(f) Non-IID, Krum Attack
(g) Non-IID, Sign-flipping Attack
(h) Non-IID, Min-Max Attack
Distance Score
Communication Round