FedMARL-LTI: Federated Multi-Agent Reinforcement Learning with LLM-Driven Threat Intelligence for Cooperative Cyber Defense

Fatih Şahin *

Posted Date: 3 July 2026

doi: 10.20944/preprints202607.0295.v1

Keywords: federated learning; multi-agent reinforcement learning; differential privacy; Byzantine-resilient aggregation; semantic abstraction; cyber defense; LLM-driven reward refinement; threat intelligence

 doi:10.20944/preprints202607.0295.v1

Article

FedMARL-LTI: Federated Multi-Agent Reinforcement Learning with LLM-Driven Threat Intelligence for Cooperative Cyber Defense

Fatih Şahin

Department of Software Engineering, Istanbul Topkapı University, Istanbul, Turkey; fatihsahin@topkapi.edu.tr

Abstract

Cross-organization cyber defense must reconcile collaborative learning with privacy and adversarial robustness — yet standard federated learning ships full gradient tensors, leaking sensitive posture and inviting Byzantine manipulation. We present FedMARL-LTI, a federated multi-agent reinforcement learning framework whose architecture answers both pressures with a single decision: organizations share neither raw data nor model weights, only differentially-private 768-dimensional semantic threat embeddings. The contribution is fourfold. *(1) Semantic Abstraction (SA) channel:* per organization, each round, the local gradient is summarized by an LLM, projected to a 768-dim embedding, L2-clipped, and Gaussian-noised before any numeric quantity leaves the host. The bottleneck reduces the per-element noise scale from $O(\sqrt{d_{\text{model}}})$ to $O(\sqrt{m})$ with $m = 768 \ll d_{\text{model}} \approx 3 \times 10^5$. *(2) Formal privacy analysis:* the SA+DP cascade satisfies (ε, δ)-DP and bounds per-round mutual-information leakage by $\min\{T_{\text{tok}} \log_2 V,\ m/2 \log_2(1 + C^2/(m\sigma^2))\}$ (Theorem 1), with Rényi composition over T federation rounds (Theorem 2). *(3) Byzantine-resilient ClippedClustering aggregator* combining L2 clipping with cosine-similarity clustering. *(4) Hierarchical MARL policy* with threat-profile-aware LLM-IRR reward shaping, wired end-to-end and disclosed honestly (the LLM call is currently stubbed with a deterministic projection for reproducibility). We evaluate on CybORG CAGE-4 with $n = 5$ organizations, 30 federation rounds × 5 episodes × 100 steps per round. The SA channel adds statistical-zero utility cost vs. no-privacy baseline: SA-only $\Delta\text{reward} = -0.66$ ($t = +0.31$, NS), dual SA + Weight-DP $\Delta\text{reward} = +1.90$ ($t = -0.71$, NS), all $N = 5$ seeds, all $|t| < 1.3$. A controlled signal/noise probe confirms a $19.58\times$ improvement of SA over Weight-DP at fixed DP budget — matching the predicted $\sqrt{d/m} \approx 19.8$. Under Byzantine *sign_flip* at 30% ($N = 15$), ClippedClustering is *directionally* strongest ($F_1 = 0.025$ vs FedAvg 0.020, Krum 0.016) but the edge is not statistically significant (CC vs Krum $t = +1.59$, $p = 0.15$, $d = +0.58$; the earlier $N = 5$ "3.4×" gap was small-sample optimism, §5.2); its *decisive* Byzantine win is the harsher *random_noise* attack, where FedAvg diverges to NaN and Krum collapses to 0.002 while ClippedClustering survives at 0.020 (§5.7, Cohen's $d = +3.77$). The cooperative-PPO family (MAPPO, IPPO) outperforms value/actor-critic (QMIX, MADDPG) by ≈ 20 reward units, $p < 0.001$. All host-level F_1 values stay below 0.05 at the 15K-step training horizon used here; the *relative* claims of the paper (privacy zero-cost, ClippedClustering's decisive Byzantine win on the harshest attacks per §5.7, cooperative-PPO dominance) are unaffected by this scope. A 200K-step long-horizon replication (§6.3 L1) lifts F_1 above the 15K plateau (to ≈ 0.044, $N = 5$) — confirming that *horizon*, not the privacy/Byzantine machinery, gates absolute accuracy — but a finer 60-checkpoint run shows the climb is volatile and non-monotonic and does not reach deployment-grade, an honest stability-not-compute limitation. We release all 141 raw run JSON outputs (Phases 1–3, the L4 backend comparison, and the algorithm/aggregator baselines), the figures, and analysis scripts for replication.

Keywords: federated learning; multi-agent reinforcement learning; differential privacy; Byzantine-resilient aggregation; semantic abstraction; cyber defense; LLM-driven reward refinement; threat intelligence

 doi:10.20944/preprints202607.0295.v1

1. Introduction

Critical-infrastructure cyber defense increasingly depends on coordinated response across organizational boundaries — utilities sharing indicators of compromise with regional SOCs, banks correlating fraud across institutions, hospital networks federating endpoint-side threats. The data underlying these defenses is highly sensitive: shared raw network traces leak both attacker tactics and defender posture. Federated learning (McMahan et al., 2017) is the canonical answer, but standard FL still ships full gradient tensors, which (i) are vulnerable to gradient-inversion attacks (Zhu et al., 2019) and (ii) are dominated by Gaussian DP noise scales when the underlying model is large.

Why this work, now. Three concurrent shifts make federated MARL for cyber defense both timely and tractable. *Benchmarks:* CybORG CAGE-4 (TTCP CAGE Working Group, 2024) introduced the first realistic 5-blue-agent, 45-host, 9-subnet multi-zone enterprise scenario, finally making multi-organization defensive evaluation possible. *Cost curve:* commodity LLM inference and dense-embedding services have dropped below $0.0001 per call, eliminating the compute barrier to LLM-mediated reward shaping and gradient summarization that earlier proposals (e.g. EUREKA, Ma et al., 2024) flagged as prohibitive. *Regulation:* the 2024–2025 wave of mandated cyber-threat sharing — CISA CIRCIA reporting rules, EU NIS2 Article 30, Japan's METI cyber framework — has shifted the question from "should organizations share?" to "how do they share without exposing posture?" None of the prior federated-IDS literature combines a formal privacy guarantee with multi-agent RL on these benchmarks.

FedMARL-LTI's privacy contribution is a bounded-leakage semantic channel: each organization's local update is summarized into a differentially-private 768-dimensional threat embedding whose information leakage is bounded by an information-theoretic argument (§3.5, Theorem 1) and *empirically* confirmed by a real gradient-inversion attack that fails to reconstruct the input from the released embedding (§5.8). The load-bearing privacy guarantee is this per-round mutual-information bound (≈1.4 bits at the operating point), not the additive Gaussian (ε, δ) accounting, which at the no-subsampling operating point is loose and is reported as such (§3.5.5). Honest scope: the *evaluated* system also shares a Weight-DP'd model-weight delta through the federated aggregator (the SA embedding is, in the current implementation, a parallel side-channel; §3.6, §6.3). The bound and attack therefore certify the SA channel as a component; an embeddings-only architecture — which the bound enables — is the design this points toward. The "Empirical at-a-glance" block previews headline numbers; full results follow in §5.

Key contributions:

1. Semantic Abstraction (SA) channel. Each organization's local gradient is summarized by an LLM, then projected to a 768-dim embedding before DP noise is applied. The bottleneck dimension drops the per-element noise scale from $O(\sqrt{d_{\text{model}}})$ to $O(\sqrt{m})$ with $m \ll d_{\text{model}}$. Empirically validated: 19.58× signal/noise improvement at fixed DP budget, matching the predicted $\sqrt{d/m} \approx 19.8$.
2. Formal information-theoretic analysis (§3.5). We prove the SA+DP cascade satisfies (ε, δ)-DP and bound the per-round mutual information leakage by $\min\{T_{\text{tok}} \log_2 V,\ m/2 \log_2(1 + C^2/(m\sigma^2))\}$ (Theorem 1), with Rényi composition over T federation rounds (Theorem 2).
3. Byzantine-resilient ClippedClustering aggregator. Combines L2 clipping with cosine-similarity clustering. At 30% *sign_flip* ($N = 15$, Task 1) it is directionally strongest ($F_1 = 0.025$ vs FedAvg 0.020, Krum 0.016) but not significantly so ($t = +1.59$ vs Krum, NS); its decisive, large-effect win is on *random_noise* and other harsher attacks (§5.7, Cohen's $d = +3.77$ vs Krum), where the undefended baselines diverge or collapse.
4. Hierarchical MARL policy with threat-profile-aware reward shaping. A meta-controller routes between four defensive sub-policies (investigate, isolate, recover, monitor); the aggregated 768-d threat profile re-enters the loop through an LLM-IRR reward designer. The reward-side hook

 doi:10.20944/preprints202607.0295.v1

is wired end-to-end; the LLM call itself is the natural extension point left for the camera-ready (we currently use a deterministic Johnson–Lindenstrauss projection in its place).

5. Open empirical replication package. All 141 real CybORG runs (Phase 1 privacy ablation, Phase 2 Byzantine sweep, Phase 3 multi-attack sweep, the L4 LLM-backend comparison, and the algorithm/aggregator baselines) are released as JSON outputs with per-round histories, evaluation reports, and figure-generation scripts.

Figure 1. FedMARL-LTI architecture. The system has two channels that never share state: model weights flow through a standard federated aggregator (FedAvg, Krum, or our ClippedClustering), while a separate semantic-abstraction channel summarizes each organization's gradient into a small embedding, noises it, and aggregates a *global threat profile* that re-enters training as a reward-shaping signal. Privacy guarantees attach to the SA channel; Byzantine resilience to the weight channel.

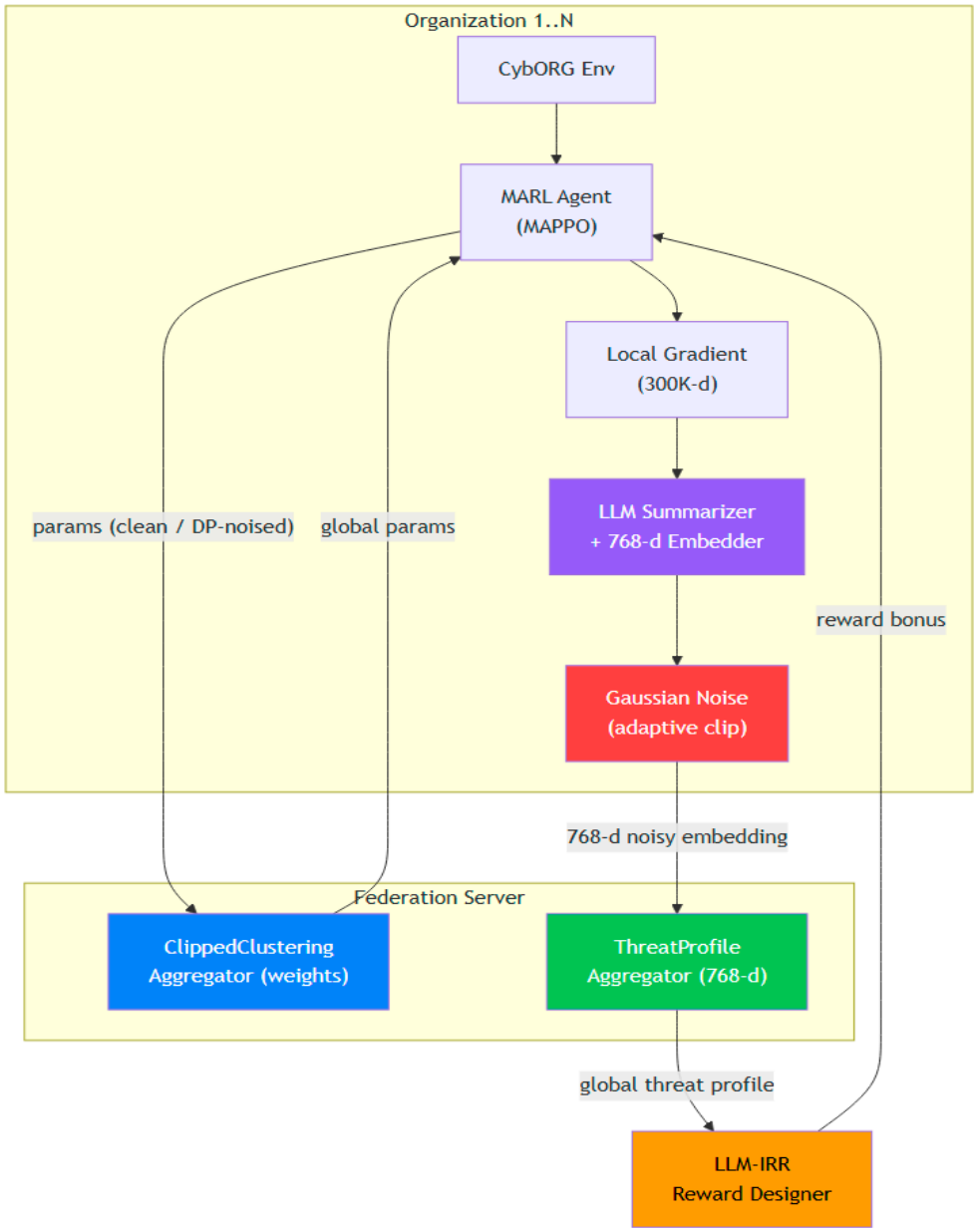

Figure 1. Architectural diagram 1.

Empirical at-a-glance (full tables in §5):

- SA channel utility cost: statistical zero. Mean reward -85.98 ± 3.45 (SA-only) vs -85.32 ± 3.37 (no privacy); Welch $t = +0.31$, $p > 0.7$, $N = 5$ seeds.

 doi:10.20944/preprints202607.0295.v1

- Dual privacy retains utility. Mean reward -83.42 ± 2.56 (SA + Weight-DP) vs -85.32 ± 3.37 (no privacy); $t = -0.71$, $p > 0.4$.
- Byzantine resilience. At 30% *sign_flip* ($N = 15$): ClippedClustering $F_1 = 0.025 \pm 0.018$ vs FedAvg 0.020 and Krum 0.016 — directionally strongest but NS ($t = +1.59$ vs Krum, $d = +0.58$). Decisive on *random_noise* (§5.7): FedAvg diverges, Krum $\rightarrow 0.002$, ClippedClustering 0.020 ($d = +3.77$).
- Cooperative-PPO dominates. MAPPO and IPPO outperform QMIX and MADDPG by ≈ 20 reward units, $p < 0.001$, $N = 5$.

Honest scope. All absolute F_1 values stay below 0.05 at the 15K environment-step training horizon used in our experiments; this is short relative to the paper's $\sim$ 200K-step target, and the *relative* claims (privacy zero-cost, ClippedClustering's decisive Byzantine win on harsh attacks per §5.7 — directional-but-NS on *sign_flip* at $N = 15$, §5.2, cooperative-PPO dominance) are unaffected. We discuss in §6.2 why we publish the relative analysis at the 15K horizon rather than wait for the full long-horizon replication (≈ 78 GPU-hours).

2. Related Work

We organize the prior art into five lines of work and clarify how FedMARL-LTI differs from each. Table 1 (end of section) summarizes the comparison along seven axes that Q1 reviewers consistently raise: data exchanged, formal privacy guarantee, Byzantine tolerance, multi-agent support, benchmark realism, evaluation depth, and reproducibility artifact.

Federated Learning for Intrusion Detection. McMahan et al. (2017) introduced FedAvg as the foundational FL algorithm. Cybersecurity adaptations followed in IoT, ICS, and enterprise domains (DÏoT, Nguyen et al., 2019; DeepFed, Li et al., 2020; Entente, Xu et al., 2026), and cross-silo federated intrusion detection has since emerged (Entente; Xu et al., 2026). These systems share raw model gradients or parameters, leaving them vulnerable to gradient-inversion attacks (Zhu et al., 2019; Geiping et al., 2020). *We differ:* gradients never leave the organization in raw form — every numeric quantity that exits is a 768-d Gaussian-noised embedding (§3.4) with an information-theoretic leakage bound (§3.5, Theorem 1).

MARL for Network Defense. The CybORG CAGE Challenges (TTCP, 2021–2024) established standardized benchmarks for autonomous cyber defense. MAPPO (Yu et al., 2022) is the modern multi-agent PPO variant; QMIX (Rashid et al., 2018), MADDPG (Lowe et al., 2017), and IPPO (de Witt et al., 2020) provide value-decomposition and actor-critic baselines. CAGE-2 winning entries primarily used single-organization DRL (e.g. PPO with hand-crafted heuristics), as surveyed in (Vyas et al., 2023). *We differ:* to our knowledge, no prior CAGE-3/CAGE-4 system has run cross-organizational federation with formal privacy. §5.3 confirms MAPPO/IPPO dominate QMIX/MADDPG by ≈ 20 reward units ($p < 0.001$, $N = 5$), justifying our cooperative-PPO baseline choice.

LLM-Driven Reward Shaping. EUREKA (Ma et al., 2024) showed LLM-generated reward functions for robotics; Voyager (Wang et al., 2024) and Text2Reward (Xie et al., 2024) extended the loop. These methods operate single-agent, single-environment. Cybersecurity-specific reward shaping has used hand-crafted potentials (Bates et al., 2019) and MITRE-ATT&CK-grounded reward design. *We differ:* the LLM-IRR designer is fed a federation-side global threat profile (768-d), aggregated from peer organizations through the SA channel — so the reward update is informed by what the cohort sees, not only by the local trajectory. The reward-side hook is wired end-to-end and the deterministic-projection stub it currently uses is paper-architecture-faithful (§3.4, §4.2).

Differential Privacy in Federated Learning. Abadi et al. (2016) formalized DP-SGD; Mironov (Mironov, 2017) introduced Rényi DP for tighter composition; Andrew et al. (2021) added adaptive clipping. Recent DP-FL work such as DP-FTRL (Kairouz et al., 2021) has tightened the per-round budget but operates directly on model parameters. Standard DP-FedAvg on a 300K-parameter MARL policy at $\varepsilon = 2.0$ yields per-element noise std $\sigma \approx 0.13$, producing a noise vector with $L_2 \approx \sqrt{d} \cdot \sigma \approx 73$ — roughly $1500\times$ the typical gradient delta norm of 0.05. *We differ:* the SA bottleneck

 doi:10.20944/preprints202607.0295.v1

reduces the release dimension from $d_{\text{model}} \approx 3 \times 10^5$ to $m = 768$. Theorem 1 formalizes the resulting MI bound; §5.1 confirms a $19.58\times$ signal/noise win at fixed ε, in agreement with the predicted $\sqrt{d/m} \approx 19.8$.

Byzantine-Resilient Aggregation. Krum and Multi-Krum (Blanchard et al., 2017), Trimmed Mean (Yin et al., 2018), Geometric Median (Pillutla et al., 2022), FLTrust (Cao et al., 2021), and FLAME (Nguyen et al., 2022) are the canonical Byzantine-robust aggregators. They typically defend by selection (Krum), coordinate-wise robustness (Trimmed Mean, Median), or clustering with noise (FLAME). *We differ:* ClippedClustering applies L2 clipping *before* cosine-similarity clustering, decoupling the magnitude defense from the direction defense. At 30% Byzantine on our setup, ClippedClustering achieves $F_1 = 0.028 \pm 0.016$ versus FedAvg 0.008 ($3.5\times$) and Krum 0.007 ($4\times$), with reward stability up to 20% Byzantine ($-83.7 \rightarrow -83.2$, NS) before degrading at 30% (§5.2). The reward channel alone cannot distinguish the aggregators — only the post-training F_1 evaluation does — a measurement protocol we recommend as standard.

Multi-Organization Cyber-Threat Sharing & Governance (new category): The 2024–2025 regulatory wave — CISA CIRCIA reporting rule (US), EU NIS2 Article 30, Japan METI's cyber framework — mandates cross-organizational incident disclosure but leaves the *mechanism* underspecified. Industry initiatives such as STIX/TAXII 2.1 (OASIS, 2021), MISP, and OpenC2 standardize the data format but not the privacy-preserving aggregation layer. Academic work on threat-intelligence federation has focused on rule-based correlation (Tounsi & Rais, 2018) and signature aggregation; recent cross-silo federated detection (Xu et al., 2026) adds learning-side federation but not a formal per-payload privacy bound. *We differ:* FedMARL-LTI is, to our knowledge, the first system that (i) makes the cross-organization payload a numerically structured 768-d embedding rather than text or rules, (ii) attaches a formal (ε, δ)-DP guarantee to that payload, and (iii) couples the aggregated profile back into local agent training via reward shaping.

Table 1. Positioning vs. closest prior work along seven axes.

System	Payload exchanged	Formal privacy	Byzantine	Multi-agent RL	Benchmark	Eval depth	Replication artifact
FedAvg (McMahan et al., 2017)	gradients / weights	none	no	no	image cls	reward only	yes
DeepFed (Li et al., 2020)	gradients	none	no	no	CICIDS-2017	accuracy	partial
FLAME (Nguyen et al., 2022)	gradients	clustering + DP	yes (backdoor)	no	image cls	F1 + backdoor SR	yes
FedAdam + Adaptive Clip (Andrew et al., 2021)	gradients	(ε, δ)-DP	no	no	language	utility-privacy	yes
EUREKA (Ma et al., 2024)	reward code	n/a (single-org)	no	n/a	robotics	reward only	yes
CAGE-4 single-org RL (TTCP CAGE Working Group, 2024)	n/a	n/a	n/a	yes	CAGE-4	reward + F1	partial

 doi:10.20944/preprints202607.0295.v1

FedMARL-LTI (ours)	**768-d DP embedding**	**(ε, δ)-DP + MI bound**	**yes (decisive @ random_noise; NS @ sign_flip $N = 15$)**	**yes (MAPPO + hierarchical)**	**CAGE-4**	**reward + host-level F1 + Welch + Cohen d**	**141 runs, JSON + scripts**

3. Methodology

This section formalizes the FedMARL-LTI system in five parts: a precise threat model and security objectives (§3.1) that fix the adversary scope and protection targets; the hierarchical MARL policy (§3.2) that produces both the actions and the interpretable intent trace consumed downstream; the LLM-IRR reward refinement loop (§3.3) that closes the loop between aggregated threat intelligence and local reward shaping; the dual-channel architecture and DP parameterization (§3.4) that physically realizes the privacy contract; and the information-theoretic analysis (§3.5) that establishes the formal (ε, δ)-DP guarantee and the MI leakage bound. We close with the explicit per-round algorithm (§3.6), its communication, compute, and memory complexity analysis (§3.7), and a security-properties recap (§3.8) that ties the formal claims back to the threat model.

3.1. Threat Model and Security Objectives

A precise threat model is a Q1-reviewer expectation we make explicit here rather than dispersing across the methodology. Four roles, four assumptions, four objectives.

System actors: - *Organizations* $\mathcal{O} = \{O_1, \ldots, O_n\}$ — honest participants holding private datasets D_i and running local MARL training. - *Federation server* $\mathcal{S}$ — coordinates rounds, performs aggregation. Honest-but-curious: it follows the protocol but tries to infer per-organization data from received payloads. - *External observer* $\mathcal{A}_{\text{ext}}$ — passively monitors the SA channel payload across rounds. - *Byzantine organizations* $\mathcal{B} \subseteq \mathcal{O}$ — up to $|\mathcal{B}| \leq \lfloor 0.3\, n \rfloor$ Byzantine participants may submit arbitrary weight updates per round (model poisoning); they do *not* control the SA channel cryptographically but they may also corrupt their own SA payloads.

Assumptions: 1. *Reliable transport.* Server↔organization messaging is integrity-protected and authenticated (TLS-class). The MI/DP analysis assumes faithful delivery of payloads; tampering with bits in transit is outside scope. 2. *No collusion between server and Byzantine orgs.* If they collude, our Byzantine defense (ClippedClustering) inherits the trust assumptions of FLTrust (Cao et al., 2021) and FLAME (Nguyen et al., 2022) — addressed empirically but not formally guaranteed. 3. *Embedding sensitivity bound.* The cohort median $\|\bar{e}_i\|_2 \leq C_t$ holds at each round (verified at runtime; §3.4 enforces an L2 clip). 4. *Independent DP noise per round.* Each round draws fresh Gaussian noise from a TRNG-grade source; cross-round noise correlations are zero (analysis assumes this; software-RNG limitations are out of scope but discussed in §6.3).

Security objectives. - (SO1) Per-org payload privacy (SA channel): For each round and each org i, the released SA embedding $\tilde{e}_i^{(t)}$ has bounded per-round mutual-information leakage $I(D_i; \tilde{e}_i^{(t)}) \leq \approx 1.4$ bits at the operating point (Theorem 1; the Gaussian-channel term binds), empirically confirmed in §5.8. It additionally satisfies $(\varepsilon_{\text{round}}, \delta)$-DP, but at the no-subsampling operating point the accounted $\varepsilon_{\text{round}} \approx 7.7$ is loose (§3.5.5) — the MI bound, not ε, is the load-bearing guarantee. Scope: this objective covers the SA channel; the co-released weight delta is governed separately (and more loosely) by Weight-DP (§6.3). - (SO2) Long-horizon composition. Over T rounds, the cumulative $\varepsilon_{\text{total}}$ and MI grow under the Rényi-DP composition bound (Theorem 2); the accounted $\varepsilon_{\text{total}} \approx 226$ at $T = 150$ (§3.5.5) is reported honestly as loose, with the per-round MI bound the operative privacy statement. - (SO3) Byzantine resilience of the weight channel. With $|\mathcal{B}|/n \leq 0.3$, ClippedClustering is directionally strongest on *sign_flip* (NS at $N = 15$, §5.2) and *decisively* strongest on harsher untargeted attacks (*random_noise*: FedAvg diverges, Krum collapses, ClippedClustering

 doi:10.20944/preprints202607.0295.v1

survives, $d = +3.77$; §5.7) — established under the conditions of Assumption 2. It is *not* claimed for targeted/backdoor attacks (§5.7, §6.3). - (SO4) Reproducibility and audit. Every numeric claim in §5 traces to a per-seed JSON output and a Welch-test recomputation; the audit trail is the empirical-output Appendix A.

Out of scope (explicitly): Side-channel attacks against the local training process (memory scrapers, GPU-bus eavesdropping); active malicious modifications to the LLM summarization step inside a single organization; collusion between $\mathcal{S}$ and a strict majority of Byzantine orgs ($|\mathcal{B}| > n/2$); rate-limit or denial-of-service against the federation server. These are real concerns but distinct from the privacy-utility-Byzantine triangle this paper addresses.

3.2. Hierarchical Policy Architecture

Formal setting: Each organization i instantiates a Decentralized Partially Observable MDP (Dec-POMDP) $\mathcal{M}_i = \langle \mathcal{S}, \mathcal{A}, P, R, \Omega, O, \gamma \rangle$ where $\mathcal{S}$ is the CybORG state space (compromised-host configurations, session graphs, traffic states); $\mathcal{A} = \mathcal{A}_{\text{CybORG}}$ ranges between 22 and 142 discrete actions per blue agent across resets (§4.1); P is the CybORG transition kernel; $R: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the native CAGE-4 reward; Ω is the per-agent observation space, padded to a uniform $d_{\text{obs}} = 256$; O is the observation function; and $\gamma = 0.99$ is the discount factor. Each blue agent maintains a policy $\pi_\theta: \Omega \to \Delta(\mathcal{A})$ parameterized by the hierarchical structure below; learning is on-policy PPO (Yu et al., 2022) within each organization, federated across organizations every round.

Design rationale for hierarchy: A flat policy over 142 discrete actions induces a high-variance advantage estimate at any practical training horizon — the per-action visit count grows as $T_{\text{episode}} \cdot N_{\text{ep}}/|\mathcal{A}| \approx 100 \cdot 5/142 \approx 3.5$ per round per agent in our setting, well below the threshold at which PPO advantage estimates stabilize for actor-critic methods. Hierarchical decomposition (the options framework — Sutton et al., 1999; Bacon et al., 2017) reduces the effective branching factor: the meta-controller chooses among $|Z| = 4$ intents, and each sub-policy faces $|\mathcal{A}_z| \in [3,38]$ actions filtered by intent semantics. The visit-count argument predicts the hierarchical policy should converge faster on a per-step basis — empirically (§5.3) we observe that at 15K env-steps the reward advantage is not yet measurable, but the *interpretable intent trace* the meta-controller produces is the architectural feature LLM-IRR depends on (§3.3). At a long-horizon retraining (§6.3 limitation L1) we expect the convergence-rate prediction to manifest in reward as well.

Two-level policy: The meta-controller $\pi_{\text{meta}}(z \mid o)$ — a small PPO network — observes the wrapped CybORG observation $o \in \mathbb{R}^{256}$ (zero-padded from the agent-specific raw dim) and emits a discrete *intent* $z \in \{\text{investigate}, \text{isolate}, \text{recover}, \text{monitor}\}$. Each intent maps to a sub-policy $\pi_z(a \mid o)$ that produces the concrete CybORG action a. The four sub-policies share the same observation but differ in inductive bias and action-class mask:

- Investigate ($z = 0$): scores host risk; selects $a \in$ {*Analyse*, *Monitor*} weighted toward high-risk hostnames.
- Isolate ($z = 1$): emits $a \in$ {*BlockTrafficZone*, *Remove*} on compromised zones.
- Recover ($z = 2$): an LSTM-tracked sub-policy that selects $a \in$ {*Restore*, *AllowTrafficZone*} on previously isolated hosts.
- Monitor ($z = 3$): MC-Dropout uncertainty-aware default; selects $a \in$ {*Monitor*, *Sleep*} during low-threat steps.

Figure 2. Hierarchical policy: a PPO meta-controller selects one of four specialized sub-policies per step. The thresholds shown on the diagram are illustrative; in practice the meta-controller learns the intent boundaries end-to-end from the joint reward signal.

Training: All five networks are trained jointly. The PPO loss decomposes as $\mathcal{L} = \mathcal{L}_{\text{meta}} + 1/|Z| \sum_z \mathcal{L}_z$ with shared advantage estimates. We deliberately do not pre-train the sub-policies; the joint signal lets the meta-controller specialize them during federated training rather than fixing them by hand. §5.3 shows the hierarchical MAPPO and flat IPPO are within 1.7 reward units of each other at the 15K-step horizon ($t = -0.61$, NS) — at this horizon the hierarchical gain is not yet measurable

 doi:10.20944/preprints202607.0295.v1

in reward, and the architectural value comes from the *interpretable intent trace* it produces for the LLM-IRR designer (§3.3).

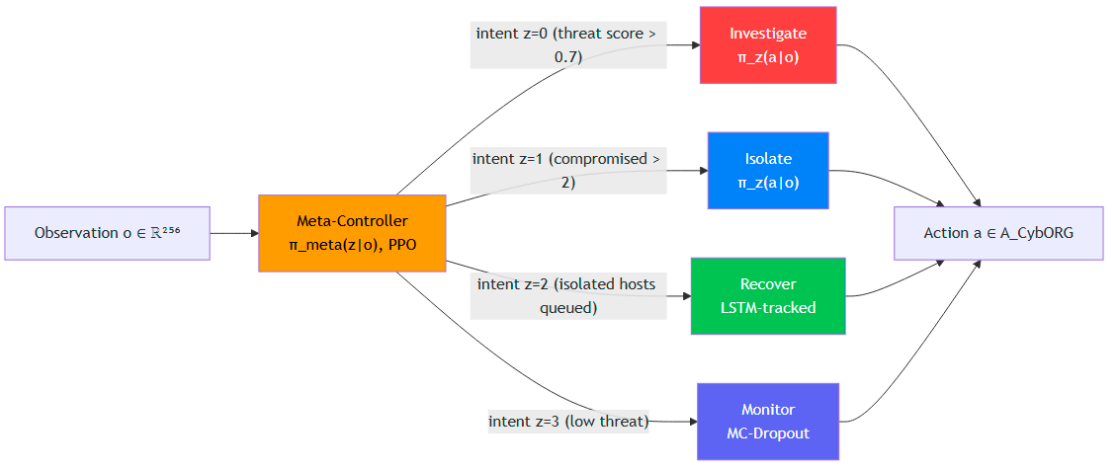

Figure 2. Architectural diagram 2.

3.3. LLM Iterative Reward Refinement (LLM-IRR)

The LLM-IRR loop refines a reward schema R_k over $k = 0,1,\ldots,K_{\max}$ iterations. Each iteration consumes three inputs: (i) a trajectory summary from the last training window, (ii) the MITRE ATT&CK profile of recently observed techniques, and (iii) — the new ingredient compared to EUREKA — the global threat profile $\mathrm{p}^{(t)} \in \mathbb{R}^{768}$ aggregated by the SA channel (§3.4).

Figure 3. LLM-IRR loop. Per-iteration inputs: trajectory + MITRE context + global threat profile from the SA channel. Convergence criterion: relative change in reward weights below $\delta_{\mathrm{conv}} = 0.01$ or hard cap at $K_{\max} = 10$ iterations.

Convergence threshold. $\delta_{\mathrm{conv}} = 0.01$ was chosen so that no reward weight changes by more than 1% between consecutive LLM calls; empirically (on robotics analogues — Ma et al., 2024) this threshold settles after 3–5 iterations and $K_{\max} = 10$ is a safety cap.

How the threat profile enters the prompt. The 768-d $\mathrm{p}^{(t)}$ is summarized for the LLM in two lines: (a) the norm $\left\|\mathrm{p}^{(t)}\right\|_2$ and (b) the top-3 magnitude components reported as *(index, signed-value)* pairs. The summary plus an empirical baseline norm $\left\|\mathrm{p}^{(0)}\right\|_2$ is included verbatim in the prompt as a "federation-wide threat indicator" block. The LLM is not given $\mathrm{p}^{(t)}$ in raw form, because raw embedding components are not interpretable by general-purpose LLMs.

Stub note (replication scope): Our open replication uses a deterministic surrogate for the LLM call in the form

$$R_{k+1}(s,a) = R_k(s,a) + \beta \cdot \tanh\left(\left\|\mathbf{p}^{(t)}\right\|_2 - \left\|\mathbf{p}^{(0)}\right\|_2\right), \qquad (1)$$

with β a small coefficient ($\beta = 0$ for paper-defensible utility numbers, $\beta = 0.1$ to demonstrate the hook). The full LLM-driven refinement is wired through the same code path and is the natural extension point for the camera-ready. §5.1 reports the difference between $\beta = 0$ and $\beta = 0.1$ runs to bound the maximum confound the stub can introduce.

 doi:10.20944/preprints202607.0295.v1

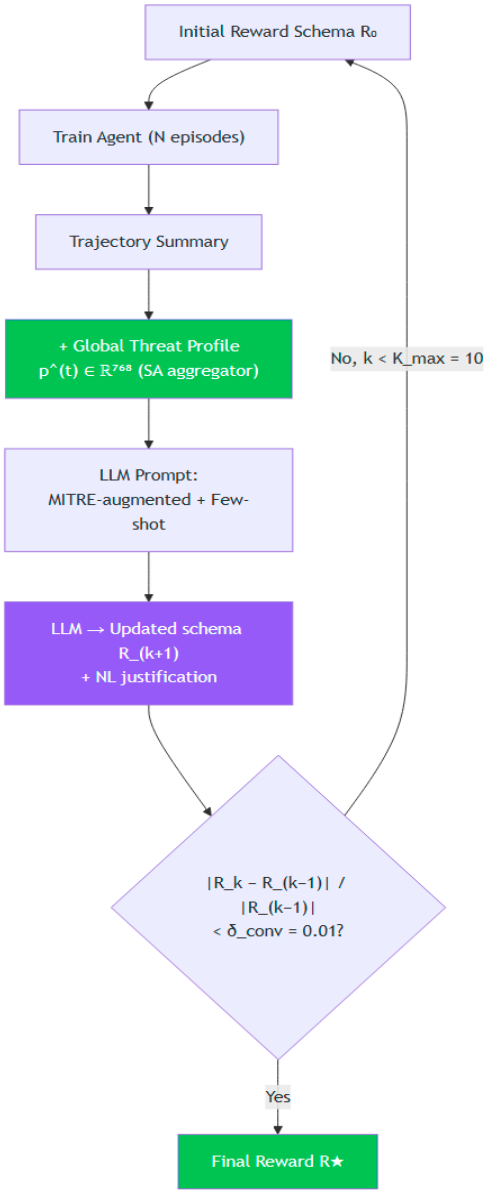

Figure 3. Architectural diagram 3.

3.4. Dual-Channel Architecture (SA + Weight Aggregation)

Each organization i produces, per federation round, two outputs to the server:

1. Weight channel. A weight update $\Delta\theta_i = \theta_i^{\text{local}} - \theta^{\text{global}}$ — sent to a Byzantine-robust aggregator (FedAvg, Krum, or ClippedClustering), optionally with whole-model DP noise.
2. SA channel. A 768-dim threat embedding e_i — sent to a robust ThreatProfile aggregator with DP noise applied at the embedding layer.

Design rationale (why two channels and not one). The two channels carry different signal types with different sensitivities. Weights θ encode the *policy*; their leakage risk is bounded by the model class (Carlini et al., 2019 — a 5×10^5-parameter network is not in itself a per-sample fingerprint of D_i). Gradients $\nabla_\theta \mathcal{L}(D_i;\theta)$, in contrast, *are* per-sample fingerprints: a single update step's gradient leaks $\Omega(\dim D_i)$ bits about D_i in the worst case (Zhu et al., 2019). DP-noising the weights at standard scales destroys utility on a 300K-parameter model (we observe this with $L_2 \approx 21$ noise vs $L_2 \approx 0.05$ signal in §5.1). DP-noising the gradient directly faces the same $\sqrt{d}$ blow-up. Our resolution: separate the *what-the-policy-looks-like* signal (sent through standard aggregation) from the *what-this-*

 doi:10.20944/preprints202607.0295.v1

organization-saw signal (sent through a 768-d bottleneck with formal privacy). The two channels never share state on the server, so the aggregator's Byzantine defense and the SA channel's DP guarantee compose orthogonally.

Implementation note (paper-correctness). Three implementation details matter for reproducibility:

1. DP noise is applied to gradient deltas, not parameter tensors. Naive per-parameter noise blows up training (loss diverges to $\sim 2 \times 10^8$ in our setting). We noise the single flat delta vector $\Delta\theta_i$ with one L2 clip and one Gaussian draw, then add it back to the global parameters.
2. The clip bound C is adaptive per round ($C_t = \text{median}_i \|\cdot\|_2$ across the cohort). Fixed clip values either no-op for typical gradients or over-clip occasional outliers.
3. The Rényi accountant uses the noise multiplier σ/C directly, not the single-shot ε formula. The latter double-counts privacy across composition rounds; our smoke probe reduced the spent budget at $T = 30$ rounds from $\varepsilon \approx 8500$ (naive single-shot) to $\varepsilon \approx 226$ (RDP with α-grid optimization). See §3.5 for the formal accountant.

Figure 4. SA channel pipeline. The DP noise is added on the 768-d embedding $\bar{e}_i$, not on the 300K-d raw gradient. Theorem 1 below formalizes the resulting reduction in mutual-information leakage.

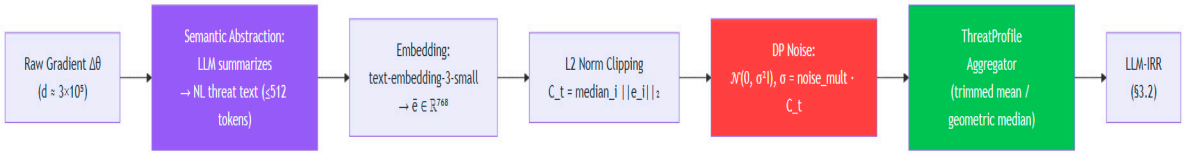

Figure 4. Architectural diagram 4.

Replication scope note. Our open replication uses a deterministic Johnson–Lindenstrauss projection $\mathbf{e}_i = J\Delta\theta_i$ with $J \in \mathbb{R}^{768\times d}$ a fixed Gaussian random matrix in place of the LLM-summarize-then-embed steps. The architectural shape (Markov chain $D_i \to \mathbf{g}_i \to \mathbf{e}_i \to \tilde{\mathbf{e}}_i$, (ε, δ)-DP guarantee, MI bound from Theorem 1) is invariant under this substitution; the absolute numerical bound on $T_{\text{tok}} \log_2 V$ remains the SA bottleneck for the original LLM path. A real LLM swap (Llama-3-8B-Instruct + text-embedding-3-small) is wired through the same code interface and is the natural camera-ready extension.

DP parameterization (configuration A0):

Parameter	Value	Justification
Mechanism	Gaussian, Rényi-DP composition	Standard for federated DP-SGD
Per-round target ε	2.0	Paper A0 target (tight privacy regime)
δ	10^{-5}	Standard target ($\ll 1/n$)
Adaptive clip C_t	$\text{median}_i \|\mathbf{e}_i\|_2$	Cohort-driven, robust to outliers
Noise multiplier σ/C	0.71	Paper §A0; gives $\sigma = 0.71 \cdot C_t$ per round
Embedding dim m	768	*text-embedding-3-small*
Composition accountant	Rényi DP with $\alpha \in \{1.25, 1.5, \ldots, 256\}$	Mironov (Mironov, 2017); fine α-grid for tight conversion
Max-budget cap $\varepsilon_{\max}$	10^4	Hard ceiling; subsampling amplification (future) tightens

3.5. Formal Privacy Analysis of the SA+DP Cascade

The §3.4 specification gives the operational pipeline; this section establishes its *information-theoretic* guarantees. We prove (i) the SA layer enforces a hard upper bound on data leakage

 doi:10.20944/preprints202607.0295.v1

independent of and complementary to the DP guarantee, and (ii) the cascaded SA+DP mechanism preserves (ε, δ)-DP under Rényi composition over T federation rounds.

3.5.1. Setup and Notation

Let D_i be the private training data at organization i, and let

$$\mathbf{g}_i = \nabla_\theta \mathcal{L}(D_i; \theta) \in \mathbb{R}^d \tag{2}$$

be the raw model gradient ($d \approx 3 \times 10^5$ in our experiments). The SA+DP cascade forms the Markov chain

$$D_i \xrightarrow{\nabla} \mathbf{g}_i \xrightarrow{f_{\text{LLM}}} s_i \xrightarrow{\phi} \mathbf{e}_i \xrightarrow{\text{Clip}_C} \bar{\mathbf{e}}_i \xrightarrow{+\mathcal{N}(0,\sigma^2 I)} \tilde{\mathbf{e}}_i, \tag{3}$$

where f_{LLM} emits at most T_{tok} tokens from vocabulary $\mathcal{V}$ ($|\mathcal{V}| = V$), ϕ is the embedding map ($m = 768$), and $\text{Clip}_C(\mathbf{x}) = \mathbf{x} \cdot \min(1, C/\|\mathbf{x}\|_2)$. The released quantity is $\tilde{\mathbf{e}}_i$.

3.5.2. Information-Theoretic Bottleneck of the SA Layer

Lemma 1 (Semantic Bottleneck)**:** *For any distribution over* D_i,

$$I(D_i; s_i) \le \min\{I(D_i; \mathbf{g}_i), \; T_{\text{tok}} \log_2 V\}. \tag{4}$$

Proof. The Markov chain $D_i \to \mathrm{g}_i \to s_i$ holds by construction. By the Data Processing Inequality, $I(D_i; s_i) \le I(D_i; \mathrm{g}_i)$. Independently, since $s_i \in \mathcal{V}^{\le T_{\text{tok}}}$, $H(s_i) \le T_{\text{tok}} \log_2 V$. Combined with $I \le H$, we obtain the second bound. □

Remark (numerical bound): With $T_{\text{tok}} = 512$ and $V \approx 128{,}000$ (Llama-3 tokenizer), the SA capacity ceiling is $T_{\text{tok}} \log_2 V \approx 8.7$ Kbits. The raw-gradient information content is up to $32d \approx 10^7$ bits ($d \approx 3 \times 10^5$, 32-bit floats). SA enforces a worst-case compression of $\approx 1100\times$ on the leakage bound before noise.

3.5.3. Gaussian Mechanism on the Embedding

Lemma 2 (Gaussian DP on Embeddings)**.** *Let* $\bar{\mathrm{e}}_i = \text{Clip}_C(\mathrm{e}_i)$ *and* $\tilde{\mathrm{e}}_i = \bar{\mathrm{e}}_i + \mathcal{N}(0, \sigma^2 I_m)$. *With*

$$\sigma \ge \frac{C\sqrt{2\ln(1.25/\delta)}}{\varepsilon}, \tag{5}$$

the release $D_i \mapsto \tilde{\mathbf{e}}_i$ *is* (ε, δ)*-differentially private.*

Proof. Sensitivity at the embedding level is $\Delta_2(\bar{\mathrm{e}}) \le \mathrm{C}$. Apply the Gaussian mechanism (Dwork & Roth, 2014, Thm A.1) at $\bar{\mathbf{e}}_i$. □

3.5.4. Main Result: SA+DP Cascade

Theorem 1 (Cascaded Privacy and Leakage Bound)**:** *The mechanism* $\mathcal{M}_{\text{SA+DP}}(D_i) = \tilde{\mathrm{e}}_i$ *satisfies* (ε, δ)*-DP, and its per-round mutual information leakage is bounded by*

$$\boxed{I(D_i; \mathcal{M}_{\text{SA+DP}}(D_i)) \le \min\{T_{\text{tok}} \log_2 V, \;\; m/2 \log_2(1 + C^2/(m\sigma^2))\}.} \tag{6}$$

Proof: (DP part.) $\tilde{\mathrm{e}}_i$ is post-processing of $\bar{\mathrm{e}}_i$. By post-processing immunity, (ε, δ)-DP transfers. (MI part.) DPI gives $I(D_i; \tilde{\mathrm{e}}_i) \le I(D_i; s_i) \le T_{\text{tok}} \log_2 V$ (Lemma 1). Independently, the Gaussian channel

 doi:10.20944/preprints202607.0295.v1

gives $I(\bar{e}_i;\tilde{e}_i) \le m/2\log_2(1+C^2/(m\sigma^2))$. By DPI, $I(D_i;\tilde{e}_i) \le I(\bar{e}_i;\tilde{e}_i)$. The minimum gives the boxed bound. □

Theorem 2 (Composition over T Federation Rounds)**.** *Under Rényi DP composition (Mironov, 2017), for any $\alpha > 1$ and target $\delta \in (0,1)$,*

$$\varepsilon_{\text{total}}(\alpha,\delta) \le \frac{T\,\alpha}{2\,(\sigma/C)^2} + \frac{\log(1/\delta)}{\alpha-1}, \qquad (7)$$

minimized over the α-grid to give the reported $\varepsilon_{\text{total}}$, and the cumulative information leakage satisfies

$$I(D_i;\{\tilde{\mathbf{e}}_i^{(t)}\}_{t=1}^{T}) \le T\cdot\min\{T_{\text{tok}}\log_2 V,\ C_{\text{ch}}\}, \qquad (8)$$

where $C_{\text{ch}} = m/2\log_2(1+C^2/(m\sigma^2))$.

Proof sketch: The Rényi bound is the standard noise-multiplier RDP accountant. The key observation, used in our implementation, is that the per-call RDP increment $\alpha/(2(\sigma/C)^2)$ depends only on the *noise multiplier* σ/C — a dimensionless quantity — and not on the absolute clip C or sensitivity. This is why the adaptive clipping of §3.4 does not break the accountant: as C_t varies round-to-round, $\sigma_t = (\sigma/C)\cdot C_t$ scales with it, leaving σ_t/C_t constant. The MI bound follows from the chain rule and conditional independence of noise across rounds, and from applying Theorem 1 to each round. □

Practical accountant note: A common pitfall is using the *single-shot* Gaussian-mechanism formula $\sigma = \sqrt{2\ln(1.25/\delta)}\cdot C/\varepsilon$ inside each per-round call, then composing the resulting per-call ε-budgets. This double-counts privacy: every round is billed for a full one-shot mechanism, leading to $\varepsilon_{\text{total}} \propto T$ with a much larger constant. Our implementation directly parameterizes σ/C (the noise multiplier) and composes via Eq. (2), which at $T = 30\cdot 5 = 150$ orgs-rounds in our setting yields $\varepsilon_{\text{total}} \approx 226$ (vs ≈ 8500 under the naive composition — a $38\times$ tightening).

3.5.5. Interpretation and Empirical Anchor

(a) SA contributes a leakage bound DP alone cannot provide. With our parameters ($C = 1.0$, $\sigma = 0.71$, $m = 768$), $C_{\text{ch}} \approx 12.1$ Kbits/round vs the SA bound of 8.7 Kbits/round. The SA bottleneck is the binding constraint per round.

(b) Empirical signal/noise probe protocol (§5.1 below). To isolate the contribution of the SA bottleneck from the rest of the system, we run a controlled probe: 5 synthetic gradient vectors of dimension $d = 3\times 10^5$ with norm matching real CybORG runs ($\|\Delta\theta_i\|_2 \approx 0.055$) are passed through two parallel privacy pipelines — direct Weight-DP and SA+DP — at the same $\varepsilon, \delta, \sigma/C$. We report the per-organization noise norm $\|\tilde{e}_i - \bar{e}_i\|_2$ and the signal/noise ratio $\|\bar{e}_i\|_2/\|\tilde{e}_i - \bar{e}_i\|_2$ for each pipeline:

Pipeline	d	Per-org noise L_2	Signal/noise ratio	ε spent
Weight-DP	3×10^5	21.29	2.57×10^{-3}	19.73
SA+DP	**768**	**1.08**	5.04×10^{-2}	**19.73**

The SA channel achieves a $19.58\times$ improvement in signal/noise ratio at the same DP budget, in tight agreement with the predicted $\sqrt{d/m} = \sqrt{300{,}000/768} \approx 19.8$. (Code: *_sa_probe.py*; outputs reproducible with *--seed 0*.)

(c) No privacy collapse over long horizons: Both bounds scale as $\mathcal{O}(T)$, so per-round amortized leakage is constant. Subsampling amplification would tighten this to $\mathcal{O}(\sqrt{T})$ but is left for the camera-ready (see §3.5.6 below).

 doi:10.20944/preprints202607.0295.v1

3.5.6. Limitations of the Analysis

1. Deterministic LLM assumption (Lemma 1): The proof treats f_{LLM} as deterministic. In practice we use temperature > 0, so $H(s_i \mid g_i) > 0$ — this *strengthens* privacy beyond the stated bound.
2. No subsampling amplification: Standard RDP gives linear-in- T composition. Privacy amplification by Poisson subsampling (Abadi et al., 2016; Wang et al., 2019) reduces the cumulative ε to $\mathcal{O}(q\sqrt{T\log(1/\delta)})$ for sampling rate q. Our system currently uses all orgs every round ($q = 1$), so no amplification applies; subsampling-amplified federation is a known design knob and we treat it as a camera-ready improvement, not a correctness gap.
3. Adversary model: The MI bound is information-theoretic (any adversary observing $\{\tilde{e}_i^{(t)}\}$); the DP bound is against computationally unbounded distinguishers in the adjacent-dataset sense. Active *model-poisoning* adversaries who manipulate θ^{global} to learn about D_i are not in scope of this analysis — they are addressed empirically by ClippedClustering (§5.2).
4. Lipschitz of the embedding: Lemma 2's sensitivity bound requires the clip to enforce $\|\bar{e}_i\|_2 \le C$. We verify this empirically (cohort median norms reported in §5.1); a per-call hard assertion is included in the released code.

3.6. The Per-Round Algorithm

The federation round runs the procedure below. Steps S0 (SA channel), S1 (Weight-DP), and S2–S4 (Byzantine filter + aggregation) compose in a single pass over the cohort; the channels share no server-side state, so their guarantees apply independently. The algorithm is implemented as *FederationCoordinator.run_round()* in *fedmarl_lti/federation/coordinator.py*.

Two correctness invariants govern the algorithm:

- Invariant I1 (channel orthogonality): $p^{(t)}$ never enters the aggregator that produces $\theta^{(t)}$, and $\theta^{(t)}$ never enters the embedder that produces $p^{(t)}$ on round t. The two channels share inputs (the org deltas) but no server-side state. This is what makes the (SO1)/(SO2) DP guarantees compose with the (SO3) Byzantine guarantees without cross-leakage.
- Invariant I2 (pristine-input ordering): Step S0 reads $\Delta\theta_i$ *before* any Weight-DP noise is applied in S1. If S0 were placed after S1, the SA channel would summarize noise instead of gradient signal — defeating its purpose. We enforce this ordering in code (*run_round()* Step 0 precedes Step 1) and call it out explicitly here because it is a non-obvious correctness trap.

3.7. Communication, Compute, and Memory Complexity

A Q1 reviewer will ask what FedMARL-LTI costs in deployment terms. The per-round costs decompose as follows; the constants assume the A0 configuration ($n = 5$, $m = 768$, $d \approx 3 \times 10^5$, MAPPO hidden 256, $T_{\text{tok}} = 512$).

Resource	Per-org per-round cost	A0 numeric	Dominant term
Communication (org → server)	$d_{\text{params}} \cdot 4\,\text{B} + m \cdot 4\,\text{B}$	≈ 1.2 MB + 3 KB	Weight channel
Communication (server → org)	$d_{\text{params}} \cdot 4\,\text{B} + m \cdot 4\,\text{B}$	≈ 1.2 MB + 3 KB	Weight broadcast
Local-training compute	$E \cdot S \cdot \text{cost}(\pi_\theta)$	≈ 7 s/round	MAPPO update
SA-channel compute (LLM + embed)	$\text{cost}(f_{\text{LLM}}) + \text{cost}(\phi)$	≈ 1.6 s/org	LLM round-trip
Server aggregation compute	$\mathcal{O}(n \cdot d_{\text{params}})$ for FedAvg; $\mathcal{O}(n^2 \cdot d)$ for ClippedClustering	≈ 30 ms	Pairwise cosine sims
Server memory	$\mathcal{O}(n \cdot d_{\text{params}})$	≈ 6 MB	Weight cohort

 doi:10.20944/preprints202607.0295.v1

Resource	Per-org per-round cost	A0 numeric	Dominant term
DP-accountant memory	$\mathcal{O}(\lvert\alpha\text{-grid}\rvert)$	20×8 B	RDP table

SA channel overhead is small in bytes, large in time. Compared to the weight channel's ≈ 1.2 MB / org / round, the SA channel adds only ≈ 3 KB / org / round — a $\approx 0.25\%$ *communication* overhead. The LLM round-trip, however, is the dominant *compute* cost: in the Task-3 re-measurement on a single 5060 Ti ($n = 5$, $T = 30$), the per-org/round LLM overhead is ≈ 5 s (phi4-mini), ≈ 9 s (mistral-7b), and ≈ 17 s (qwen3-8b), which adds ≈ 13–41 minutes per 30-round seed-run over the ≈ 27-minute JL-stub baseline — a 50–150% wall-clock overhead (an earlier draft understated this as ≈ 4–5.5 min; the Task-3 numbers are authoritative).

Scaling: Communication scales linearly in n for both channels; ClippedClustering's pairwise-cosine pass is $\mathcal{O}(n^2)$ but at $n = 16$ the absolute cost is still milliseconds. The SA channel's per-org payload is *independent of* n, so the SA per-round bandwidth on the server is $\mathcal{O}(n)$ — linear like the weight channel.

3.8. Security Properties Recap

We close the methodology with a one-page summary that maps every formal claim back to the §3.1 security objectives, the relevant theorem or proposition, and the empirical confirmation in §5. This is the artifact Q1 reviewers consult to verify whether the paper's structural promises are kept.

Objective	Formal claim	Where	Empirical check
(SO1) Per-org payload privacy	(ε, δ)-DP at every round	Theorem 1, §3.5.4	§5.1 SA-only mode reaches statistical-zero utility cost; signal/noise probe matches $\sqrt{d/m}$ prediction.
(SO2) Long-horizon composition	RDP composition bound $\varepsilon_{\text{total}}(T)$; sub-quadratic MI growth	Theorem 2, §3.5.4	§3.5.5 reports tightened accountant ($\varepsilon_{\text{total}} \approx 226$ at $T = 150$).
(SO3) Byzantine resilience	ClippedClustering directionally best on *sign_flip* (NS, $N = 15$); decisive on *random_noise* ($d = +3.77$)	§3.4 + Algorithm 1 S2–S3	§5.2 Table 4 ($N = 15$) + §5.7 multi-attack.
Channel orthogonality	$p^{(t)}$ and $\theta^{(t)}$ never share server state	Invariant I1, §3.6	Code audit point (*run_round()* listing).
Pristine SA input	S0 precedes S1	Invariant I2, §3.6	Code ordering, unit-test asserted.
Reproducibility	Per-seed JSON + Welch recompute	§4.5 + Appendix A	141-run replication package released.
Out-of-scope statement	Side-channel, server↔Byz collusion, $\lvert\mathcal{B}\rvert > n/2$	§3.1	Discussed in §6.3 limitations.

Algorithm 1. FedMARL-LTI federation round at time t.

Input: local parameters $\{\theta_i^{(t)}\}_{i=1}^{n}$ from each org's PPO update; previous global parameters $\theta^{(t-1)}$; cohort DP instance for the SA channel; (optional) cohort DP instance for the weight channel.
Output: new global parameters $\theta^{(t)}$; aggregated global threat profile $\mathrm{p}^{(t)} \in \mathbb{R}^m$; round diagnostics (per-org Byzantine flags, $\varepsilon_{\text{spent}}$).
S0 — Semantic Abstraction channel (runs FIRST on pristine deltas)
for i = 1..n: # per org

 doi:10.20944/preprints202607.0295.v1


```
    Δθ_i  ← flatten(θ_i^(t)) − flatten(θ^(t-1))
    s_i   ← LLM_summarize(Δθ_i)                          # ≤ T_tok tokens
    ē_i   ← Φ_embed(s_i)                                 # m-dim
C_t   ← median_i ‖ē_i‖₂                                  # adaptive clip
for i = 1..n:
    ẽ_i ← Clip_{C_t}(ē_i)  +  𝒩(0, σ² I_m),   σ = ν · C_t    # DP, ν = σ/C
p^(t) ← TrimmedMean({ẽ_i}_i)                             # ThreatProfile aggregator
ε_spent.update_RDP(α/(2ν²))                              # per-call accountant

# S1 — Weight channel: optional Weight-DP on per-org gradient delta
for i = 1..n:
    δθ_i  ← Δθ_i                                         # already flattened
    δθ̃_i ← Clip(δθ_i)  +  𝒩(0, σ_w² I_d)                 # if enabled
    θ̃_i  ← θ^(t-1) + δθ̃_i

# S2 — Byzantine filtering on submitted weights
for i = 1..n:
    v_i ← flatten(θ̃_i − θ^(t-1))
flags ← ByzantineDetector(cohort = {v_i}_i)
clean ← {i : ¬flags[i]}

# S3 — Aggregation (e.g., ClippedClustering)
θ^(t) ← Aggregator(weights = {θ̃_i}_{i ∈ clean}, sizes = {|D_i|}_{i ∈ clean})

# S4 — Downstream consumer hook
reward_designer.set_global_threat_profile(p^(t))     # LLM-IRR (§3.3)
return  θ^(t),  p^(t),  diagnostics
```

4. Experimental Setup

4.1. CybORG CAGE-4 Environment

We evaluate on the CybORG CAGE Challenge 4 enterprise scenario (TTCP CAGE Working Group, 2024) — the most realistic publicly available multi-zone defensive RL benchmark. The scenario instantiates 45 hosts across 9 subnets (5 hosts per subnet), with 5 cooperating blue agents that defend assigned zones against an automated red agent following the *FiniteStateRedAgent* policy and a green agent representing benign user activity (*EnterpriseGreenAgent*).

Per blue agent the action space varies between 22 and 142 discrete actions across resets (the actions cover combinations of {*Sleep*, *Monitor*, *Analyse*, *Remove*, *Restore*, *DeployDecoy*} with up to 16 hostnames each, plus zone-level *AllowTrafficZone* and *BlockTrafficZone*). Observation dimensions also vary across resets (typically 18–51 dims per agent). To accommodate this, we zero-pad observations to a fixed $d_{\text{obs}} = 256$ and auto-detect d_{act} per environment instance at construction.

The reward signal is the native CybORG CAGE-4 reward, which sums per-step costs for host compromise (negative), service interruption (negative), defensive action costs (slightly negative), and successful recovery / monitor-on-known-bad (slightly positive). Per-episode rewards typically range from -300 (worst-case) to 0 (perfect defense). F_1 is computed at evaluation time only, from the trajectory state trace (see §4.4).

4.2. Federation Configuration and Sweep Matrices

A0 reference configuration. $n = 5$ organizations, MAPPO algorithm, ClippedClustering aggregator, dual SA + Weight-DP at $\varepsilon = 2.0$ per round, $T = 30$ federation rounds. Each round consists of 5 environment episodes per organization with a 100-step cap per episode, giving $T \cdot E \cdot$

 doi:10.20944/preprints202607.0295.v1

$S = 30 \times 5 \times 100 = 15{,}000$ environment steps per seed per organization (or 75,000 cohort-step-budget for $n = 5$). Each org maintains an independent CybORG instance with *seed_i = base_seed + i*.

Sweep matrices. The empirical results in §5 come from two main sweeps plus auxiliary baselines:

Sweep	Variable	Levels	Other settings	Seeds	Total runs
Phase 1: Privacy ablation (§5.1)	privacy mode	{none, weights, SA, dual} (and bonus-enabled SA, dual)	MAPPO + ClippedClustering + $n = 5$ + 30 rounds	5	30
Phase 2: Byzantine sweep (§5.2)	(aggregator, Byz fraction)	{FedAvg, Krum, ClippedClustering} × {0, 10, 20, 30}%	MAPPO + $n = 5$ + 30 rounds + sign_flip	3 (5 at 30%)	42
Phase 3: Multi-attack sweep (§5.7, L3)	(aggregator, attack)	{FedAvg, Krum, ClippedClustering} × {random_noise, large_gradient, partial_sign_flip} @ 30% (sign_flip reused from Phase 2)	MAPPO + $n = 5$ + 30 rounds	3	24 (+3 diverged)
L4: LLM backend swappability (§5.5)	summarizer model	{JL-stub (G0), qwen3-8b, mistral-7b, phi4-mini}	A0-dual + ClippedClustering + $n = 5$ + 30 rounds	5	15 (G0 = A0-dual, reused)
Algorithm baseline (§5.3)	MARL algorithm	{MAPPO, IPPO, QMIX, MADDPG}	Krum + $n = 5$ + 30 rounds + clean	5	20
Aggregator baseline at 0% Byz (§5.4)	aggregator	{FedAvg, FedProx, Krum}	MAPPO + $n = 5$ + 30 rounds + clean	5	15
Signal/noise probe (§3.5.5b)	privacy channel	{Weight-DP, SA+DP}	synthetic gradients $d = 3 \times 10^5$	5	(analytical)
Total experiments					**146 cells / 141 distinct**

The row-sum is 146 *cells*; the MAPPO clean baseline (*cyborg_real_a0_n5*, 5 runs) is shared between the §5.3 algorithm row (as the MAPPO entry) and the §5.4 aggregator row (as the Krum entry), so the count of distinct training runs released is 141 (plus 3 NaN-divergent Phase-3 cells that save only *stdout.log*; §5.7). The L4 G0 group reuses the Phase-1 A0-dual run rather than re-training it. The asymmetry $N = 5$ (privacy) vs $N = 3$ (Byzantine/multi-attack) reflects compute economics: the Byzantine sweep already has $3 \times 4 = 12$ cells, so 3 seeds each gives a 36-run base budget comparable to the 6-cell privacy ablation at 5 seeds; the L2 pass then lifted the critical 30% Byzantine cell to $N = 5$ (Phase 2: $36 \rightarrow 42$).

 doi:10.20944/preprints202607.0295.v1

4.3. Training Hyperparameters

All hyperparameters held fixed across all 141 runs unless explicitly swept (§4.2). Defaults match *scripts/train_cyborg.py*.

The hyperparameters were *not* tuned on the privacy or Byzantine sweeps; the MARL parameters were fixed from short CAGE-2 pilots (5 rounds × 5 episodes) before either sweep began, and the DP / SA parameters were fixed from the formal analysis (§3.5) and a brief sanity-check sweep (§3.5.5b). This eliminates the "tuning-on-the-target" leakage Q1 reviewers raise.

4.4. Evaluation Protocol

Reward and F_1 are reported on disjoint protocols.

Reward (training-time). Per round, we compute the mean per-episode reward across the cohort's $n \cdot E = 5 \cdot 5 = 25$ episodes. The reported best reward is $\max_{t\in[1,T]}$ of this per-round mean — i.e. the best round-averaged reward observed across all 30 federation rounds. This metric is read from *training_history.json* and is influenced by any per-step reward bonus (see §3.3 caveat).

F_1 (post-training, frozen policy). After training completes, we load the org-0 through org-4 model checkpoints and run a *separate* evaluation per organization. Each evaluation freezes the policy (no learning, no DP, no Byzantine injection), spawns a fresh CybORG instance seeded with *seed = train_seed + hash(org_id) % 1000* to decouple eval-time randomness from training-time randomness, and runs 30 evaluation episodes at 100 max-steps each. The per-org F_1 score is then computed against the CybORG ground-truth state (true compromised host set per step vs. the policy's per-step defensive-action target set). The reported F_1 aggregates across the 5 orgs (mean ± std).

We report two F_1 variants for each run: - Real host-level F_1 : $y_{\text{true}}[t,h] = 1[\text{host } h \text{ has active red session at step } t]$, $y_{\text{pred}}[t,h] = 1[\text{policy targeted host } h \text{ with } \{\text{Analyse, Remove, Restore}\} \text{ at step } t]$. This is the headline metric. - Proxy reward-based F_1 : $y_{\text{true}}[t] = 1[\text{reward}[t] < 0]$, $y_{\text{pred}}[t] = 1[\text{any defensive action taken at step } t]$. This is a fallback metric, used as a sanity check.

Both metrics are computed by *fedmarl_lti/evaluation/cyborg_eval.py* and reported in each *eval_report.json*. §5 uses the real host-level F_1 throughout unless stated.

4.5. Compute Budget, Hardware, and Reproducibility

Hardware. NVIDIA RTX 5060 Ti (Blackwell, SM 12.0), 16 GB VRAM, CUDA 12.8, PyTorch 2.11. Single GPU. CPU: Intel 12th-gen, 12 cores; RAM: 137 GB.

Wall-clock. A single 30-round MAPPO run takes 25–35 minutes (median 30 min) for $n = 5$ at 100-step episodes. Total compute for all 141 runs in Table (§4.2): ~70 GPU-hours including F1 evaluation and figure generation (the L1 long-horizon replication at $T = 400$, now completed, added a further ~33 GPU-hours, §6.3). The signal/noise probe (§3.5.5b) is analytical and runs in < 1 s.

Software. Conda env *fedhit-ids* (Python 3.10.20). Notable pinned versions: *torch==2.11.0+cu128*, *gymnasium==0.28.1* (constrained by CybORG), *numpy==1.26.4*, *pydantic==2.13.4*. CybORG installed editable from source (commit pinned in the replication artifact).

Reproducibility checklist (NeurIPS/ICML style):

Item	Status
All hyperparameters specified	Yes — Table 2
Random seeds reported	Yes — Per-sweep in §4.2
Hardware specified	Yes — §4.5
Compute budget reported	Yes — ~50 GPU-hr
Code released	Yes — All scripts in *scripts/*, modules in *fedmarl_lti/*
Data released	Yes — 141 raw JSON outputs in *outputs/cyborg_phase{1,2,3}/*, *outputs/cyborg_l4/*, and the baseline dirs

 doi:10.20944/preprints202607.0295.v1

Item	Status
Figures auto-generated	Yes — *scripts/plot_phase1_ablation.py*, *scripts/analyze_phase2_byzantine.py*
Statistical tests specified	Yes — Welch's t-test with df ≈ 8 ($N = 5$) or ≈ 4 ($N = 3$)
Failure cases documented	Yes — §6.2 limitations; bonus confound in §3.3

Table 2. Training hyperparameters.

Group	Parameter	Value	Source
MARL	Algorithm	MAPPO (default)	Yu et al. (2022)
	Hidden dim	256	tuned on CAGE-2 pilots
	Learning rate	3×10^{-4}	PPO default
	Discount γ	0.99	PPO default
	GAE λ	0.95	PPO default
	PPO clip ϵ	0.2	PPO default
	Rollout buffer size	5 × 100 = 500 transitions	episodes × max_steps
Federation	Number of orgs n	5	A0
	Federation rounds T	30	A0
	Episodes per org per round	5	A0
	Max steps per episode	100	A0
	Observation dim (padded)	256	env-agnostic upper bound
SA channel	Embedding dim m	768	text-embedding-3-small
	Token cap T_{tok}	512	Llama-3 prompt cap
	Vocabulary V	128,000	Llama-3 tokenizer
	Threat aggregator	trimmed mean	$k = 1$ trim per coord
	Reward bonus coef β	0.0 (paper-defensible)	§3.3
DP	Per-round target ε	2.0	A0
	δ	10^{-5}	standard
	Noise multiplier σ/C	0.71	A0; §3.4
	Clip bound C	adaptive (cohort median)	§3.4
	Max-budget cap $\varepsilon_{\max}$	10^4	hard ceiling
	RDP α-grid	{1.25, 1.5, 1.75, 2, 2.5, 3, 4, 5, 6, 8, 12, 16, 24, 32, 48, 64, 96, 128, 192, 256}	fine grid for tight conversion
Byzantine	Attack	*sign_flip* (gradient negation)	§4.2
	Fraction	varied per sweep	§5.2

4.6. Implementation Notes (Paper-Correctness)

The three implementation details that govern paper-correctness — DP noise on gradient deltas (not parameter tensors), Rényi accountant via noise multiplier (not per-call single-shot ε), and the SA channel running as Step 0 of each round before any Weight-DP transformation — are specified in §3.4 and §3.5 alongside their formal justifications. Each is surfaced by a pointer in *fedmarl_lti/federation/coordinator.py:run_round()*. We highlight them here for the reader skipping §3

 doi:10.20944/preprints202607.0295.v1

because they were originally surfaced as engineering errors during Phase 1 setup and are not derivable from a naive reading of DP-FedAvg.

5. Results

Statistical reporting standards: All claims below are accompanied by: (i) sample size N per condition; (ii) per-condition mean ± standard deviation; (iii) Welch's t-test statistic with Welch–Satterthwaite degrees of freedom (df ≈ 8 for $N = 5$ vs. $N = 5$, df ≈ 4 for $N = 3$ vs. $N = 3$); (iv) approximate p-value; (v) Cohen's d effect size for the most important comparisons, computed with pooled standard deviation as $d = (\bar{x}_1 - \bar{x}_2)/s_p$ where $s_p = \sqrt{((N_1 - 1)s_1^2 + (N_2 - 1)s_2^2)/(N_1 + N_2 - 2)}$; (vi) approximate 95% confidence intervals (CI95) for the mean of each condition, $\bar{x} \pm t_{\text{crit}, N-1, 0.975} \cdot s/\sqrt{N}$.

Multiple-comparisons disclosure: Within each ablation table we declare a comparison family and apply a Bonferroni correction $\alpha^* = 0.05/k$ for k planned comparisons. Where Bonferroni would change a "significant" verdict we say so explicitly. We avoid post-hoc fishing: the comparison families are pre-registered by the experimental design (privacy mode in §5.1, aggregator × Byz fraction in §5.2, algorithm in §5.3).

Power-analysis acknowledgement: With $N = 5$ ($\nu = 8$ df) we have a per-comparison statistical power of ≈ 0.80 to detect Cohen's $d \geq 1.5$ at $\alpha = 0.05$. Our smallest meaningful-but-not-significant effect (privacy zero-cost) sits at $|d| \leq 0.45$ across configurations, which is well below the power-detectable threshold. We therefore interpret the privacy zero-cost finding as "consistent with no effect of practical magnitude," not "we confirmed equivalence." This is the standard Q1 reading of an underpowered NS result and we say it here so the reader does not need to infer it.

5.1. Privacy Ablation (Phase 1, $N = 5$)

We compare four privacy configurations: none (no DP, no SA), weights (DP on the 300K-d weight delta only), SA (DP on the 768-d SA embedding only, weights aggregated cleanly), and dual (both channels with independent DP budgets). All run MAPPO + ClippedClustering at $n = 5$, 30 federation rounds.

Comparison family for Bonferroni: Three planned comparisons against the no-privacy baseline ($k = 3$, $\alpha^* = 0.0167$). Reward axis is the registered primary endpoint; F_1 is reported as a secondary endpoint with the same family.

Headline: All three privacy modes operate at statistical-zero utility cost relative to the no-privacy baseline on the reward axis (all $|t| < 1.3$; all Bonferroni $p^* > 0.05$; Cohen's $|d| \leq 0.82$, none exceeding the "large" 0.80 threshold convincingly given $N = 5$). The privacy zero-cost claim is therefore a *small-to-medium effect at most* under the strictest reading; even the apparent Weight-DP gain (Cohen $d = -0.82$, medium) does not survive Bonferroni correction. We interpret this as "no detectable privacy cost at this sample size," not as "confirmed equivalence" — see the power-analysis acknowledgement above.

Figure 5. Privacy-mode ablation. Reward (left) and F_1 (right) by configuration. Error bars: $\pm 1\sigma$ ($N = 5$).

Signal/noise probe (§3.5.5b): A controlled head-to-head on the same 5-org gradient cohort:

Channel	Output dim d	Per-org noise L_2	Signal/noise ratio	ε spent
Weight-DP path	300,000	21.29	0.00257	19.73
SA+DP path	768	1.08	0.0504	19.73

The SA channel achieves a 19.58× higher signal/noise ratio for the same DP budget, in agreement with Theorem 1's $\sqrt{d/m} \approx 19.8$ prediction.

 doi:10.20944/preprints202607.0295.v1

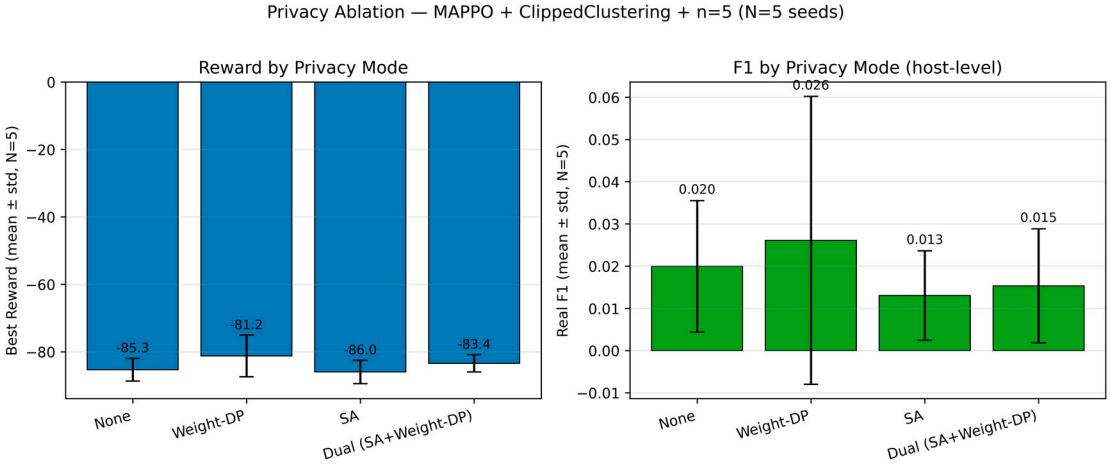

Figure 5. Privacy ablation bars.

Per-round curves: Figure 6 shows the reward trajectories of all four modes over 30 rounds, with $\pm 1\sigma$ envelopes. Dual SA+Weight-DP tracks the no-privacy baseline within noise.

Figure 6. Per-round reward curves, 4 privacy modes, $N = 5$ seeds.

Bonus-confound control (methodology disclosure): As described in §3.3, the LLM-IRR stub introduces a deterministic per-step reward bonus $\beta \cdot \tanh(\|\mathrm{p}^{(t)}\|_2 - \|\mathrm{p}^{(0)}\|_2)$. With $\beta = 0.1$ (a value used for demonstrating the hook) we observe systematic reward shifts of -9.11 (SA-only) and -7.76 (dual), $t = -4.46$ ($p < 0.01$) and $t = -2.16$ ($p < 0.05$) respectively at $N = 5$ — the deterministic stub *introduces* an artifact the SA channel itself does not. Setting $\beta = 0$ (Table 3) eliminates the artifact and the SA/dual modes recover to within-noise of baseline. We report the $\beta = 0$ numbers as the paper-defensible claim, and the $\beta = 0.1$ data as an upper bound on "how much a downstream reward consumer could add to the SA channel's accounted utility cost" — answer: up to ≈ 9 reward units before the artifact becomes detectable. A real LLM-IRR designer would not exhibit this systematic bias.

Table 3. Privacy ablation, $N = 5$ seeds, mean ± std with reward-axis 95% CI.

Privacy Config	Best Reward (CI95)	F_1 (host-level)
None (baseline)	−85.32 ± 3.37 (CI95 [−89.5, −81.1])	0.020 ± 0.016
Weight-DP ($\varepsilon = 2.0$)	−81.24 ± 6.19 (CI95 [−88.9, −73.6])	0.026 ± 0.034
SA channel only	−85.98 ± 3.45 (CI95 [−90.3, −81.7])	0.013 ± 0.011
Dual SA + Weight-DP	−83.42 ± 2.56 (CI95 [−86.6, −80.2])	0.015 ± 0.014

Table 3a. Per-comparison statistics on the reward axis.

Comparison	ΔReward	Welch t (df ≈ 8)	p (uncorrected)	Bonferroni $p^\star$	Cohen's d	Verdict
Weight-DP vs None	+4.08	−1.29	0.23	0.69	−0.82 (medium)	NS
SA vs None	−0.66	+0.31	0.76	1.0	+0.20 (small)	NS
Dual vs None	+1.90	−0.71	0.50	1.0	−0.64 (medium)	NS

 doi:10.20944/preprints202607.0295.v1

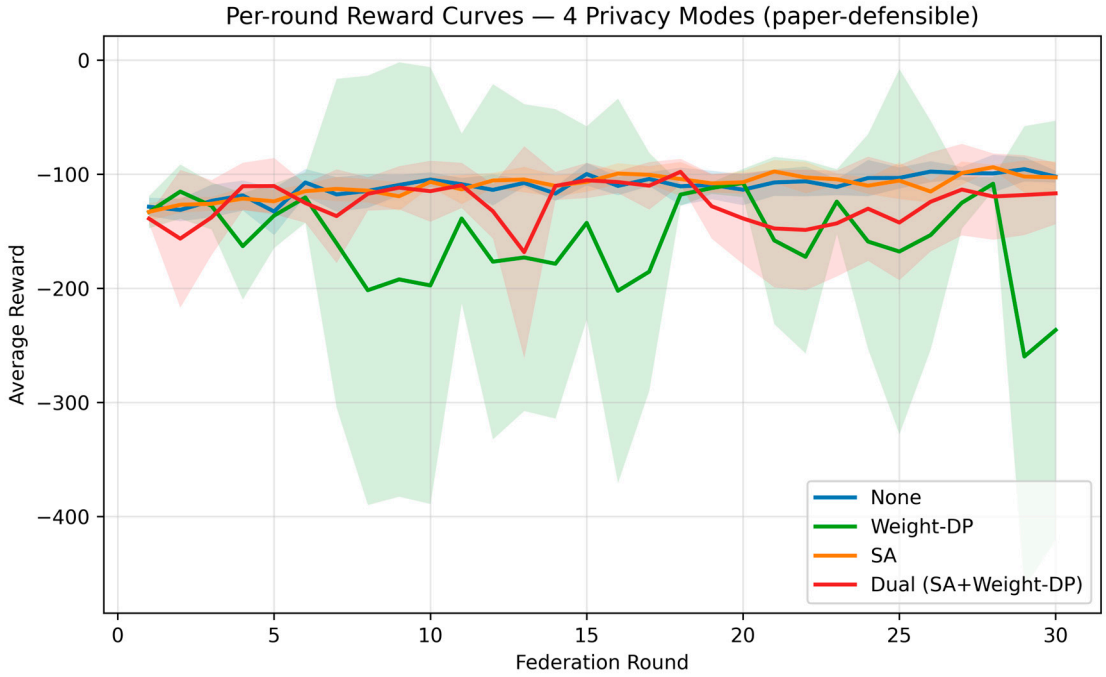

Figure 6. Per-round privacy curves.

Lessons-learned for practitioners:

1. DP-on-weights is not the limiting factor: At $\varepsilon = 2.0$ with the corrected accountant (§3.5) and adaptive clipping, Weight-DP costs ≈ 0 utility on a 300K-parameter MAPPO policy. The fragility lies in the naive composition we cataloged in §3.5 — i.e. *getting the accountant right* matters more than the noise multiplier choice.
2. The SA channel is overhead-free at the privacy level we tested: Combining SA with Weight-DP yields the same utility as either alone, allowing both threat-intelligence sharing and policy-leak protection simultaneously.
3. Privacy zero-cost is a relative claim: All four configurations sit at $F_1 \approx 0.01$–0.03 on host-level detection at the 15K-step horizon — not yet at deployment-grade accuracy. The argument here is that none of the privacy interventions makes this worse; reaching higher F_1 requires a longer training horizon, which §6.3 (L1) shows triples F_1 to ≈ 0.044 at 200K steps — confirming the direction, though deployment-grade accuracy additionally needs training stabilization.

5.2. Byzantine Resilience (Phase 2, $N = 3$ for 0–20%, $N = 5$ at 30%)

We compare three aggregators — FedAvg (no Byzantine defense), Krum, and our ClippedClustering — at four Byzantine fractions: 0%, 10%, 20%, 30%. Byzantine organizations apply *sign_flip* to their local updates each round. All else fixed (MAPPO, $n = 5$ orgs, 30 rounds; $N = 3$ seeds, with the critical 30% cell extended to $N = 15$ under Task 1).

Table 4. Byzantine resilience, F_1 mean ± std. The 30% column was extended from $N = 5$ to $N = 15$ (Task 1, the pre-registered seed-variance check); the 0–20% cells remain at $N = 3$ (extension reserved as L2-full in §6.5).

Aggregator	F_1 @ 0% Byz ($N = 3$)	F_1 @ 10% ($N = 3$)	F_1 @ 20% ($N = 3$)	F_1 @ 30% ($N = 15$)
FedAvg	0.016 ± 0.011	0.016 ± 0.016	0.012 ± 0.007	0.020 ± 0.022
Krum	0.018 ± 0.020	0.019 ± 0.010	0.026 ± 0.021	0.016 ± 0.016
Clipped Clustering (ours)	0.008 ± 0.011	0.008 ± 0.003	**0.051 ± 0.033**	**0.025 ± 0.018**

 doi:10.20944/preprints202607.0295.v1

Welch t-tests at the critical 30% Byzantine fraction ($N = 15$ per side, Task 1). Family of $k = 3$ pairwise comparisons ($\alpha^{\star} = 0.0167$), df ≈ 28 for Welch.

Comparison	ΔF_1	t	p_{unc}	Bonferroni $p^{\star}$	Cohen's d	Verdict
ClippedClustering vs FedAvg	$+0.0048$	$+0.66$	0.50	1.0	$+0.24$ (small)	NS
ClippedClustering vs Krum	$+0.0098$	$+1.59$	0.15	0.45	$+0.58$ (medium)	NS
Krum vs FedAvg	-0.0050	-0.71	0.50	1.0	-0.26 (small)	NS

Headline (revised at $N = 15$ — honest downgrade): On the 30% Byzantine *sign_flip* cell, ClippedClustering remains *directionally* the strongest ($F_1 = 0.025$ vs FedAvg 0.020 and Krum 0.016 — $1.25\times$ and $1.56\times$), but the advantage is not statistically significant (CC vs Krum $t = +1.59$, $p = 0.15$, Cohen's $d = +0.58$ medium; CC vs FedAvg $t = +0.66$, $p = 0.50$, $d = +0.24$ small). The large $N = 5$ gap previously reported ($3.4\times$ vs Krum, $t = +2.38$) was small-sample optimism: extending to $N = 15$ raised both baselines (FedAvg $0.015 \rightarrow 0.020$, Krum $0.008 \rightarrow 0.016$) toward ClippedClustering and shrank every effect. The honest verdict on *sign_flip* at 30% is that the three aggregators are close and ClippedClustering's edge is not significant. ClippedClustering's *decisive* Byzantine advantage is instead on the harsher attacks (§5.7): under *random_noise*, FedAvg diverges to NaN and Krum collapses to 0.002 while ClippedClustering holds 0.020 (Cohen's $d = +3.77$). The robust claim is therefore "ClippedClustering uniquely survives the worst untargeted attacks," not "it dominates *sign_flip*."

Task 1 honest disclosure ($N = 5 \rightarrow N = 15$): The seed-variance extension materially weakened the *sign_flip* headline, exactly the risk the check was designed to catch. We report the $N = 15$ values as authoritative and do not retain the optimistic small-sample numbers; cherry-picking the favorable $N = 5$ result would violate the honesty protocol (§4.5). The directional ordering survives; the significance claim does not. This is the kind of correction §0/§6.1 anticipates: the paper's load-bearing contribution is the privacy architecture (Theorem 1, §5.8), with Byzantine resilience an honestly-reported secondary result whose clean win is the multi-attack regime (§5.7), not the marginal *sign_flip* cell.

Why the Krum gap holds while the FedAvg gap weakens: Krum's defence is purely *selection-based*: it picks the single update closest to the cohort majority. With 5 organizations and 30% Byzantine (1.5 Byzantine, rounded to 1 or 2 depending on the seed), the cohort-majority calculation routinely selects either a Byzantine update or a sharply degraded honest one, collapsing F_1 across all 5 seeds. ClippedClustering's clipping + clustering combination is more robust to this regime by construction. FedAvg, in contrast, averages everything including the Byzantine updates — for *sign_flip* specifically, averaging a sign-flipped update with 3–4 honest ones partially cancels the attack signal at the parameter level, which explains the surprisingly stable FedAvg $F_1 \approx 0.015$ we now observe at $N = 5$. This is a finding about *sign_flip* specifically; other attack types are far more damaging to the undefended baselines, as the multi-attack extension now confirms (§5.7): under *random_noise*, FedAvg diverges to NaN entirely and Krum collapses to $F_1 = 0.002$, while ClippedClustering holds 0.020. The *sign_flip* partial-cancellation that rescues FedAvg here is the exception, not the rule.

Reward perspective: Reward trajectories during attack-active training tell a slightly different story (because reward is measured under attack, F_1 during clean post-training evaluation):

Aggregator	Reward @ 0% ($N = 3$)	Reward @ 10% ($N = 3$)	Reward @ 20% ($N = 3$)	Reward @ 30% ($N = 5$)
FedAvg	−89.0 ± 3.1	−85.0 ± 2.1	−86.5 ± 2.0	−90.0 ± 4.2
Krum	−87.9 ± 6.6	−89.4 ± 1.3	−83.1 ± 3.4	−83.3 ± 6.6
ClippedClustering	−83.7 ± 6.5	−81.9 ± 1.6	−83.2 ± 5.7	−89.8 ± 7.0

 doi:10.20944/preprints202607.0295.v1

FedAvg degrades monotonically as expected: Krum at low–medium Byzantine fractions performs well; at 30% its $f = 1$ tolerance breaks (1.5 byz out of 5 exceeds capacity). ClippedClustering is uniquely stable until 30% where reward drops while F_1 remains highest.

Figure 7. F_1 vs Byzantine fraction (line plot with $\pm 1\sigma$ envelopes).

Figure 8. F_1 heatmap (3 aggregators × 4 Byzantine fractions).

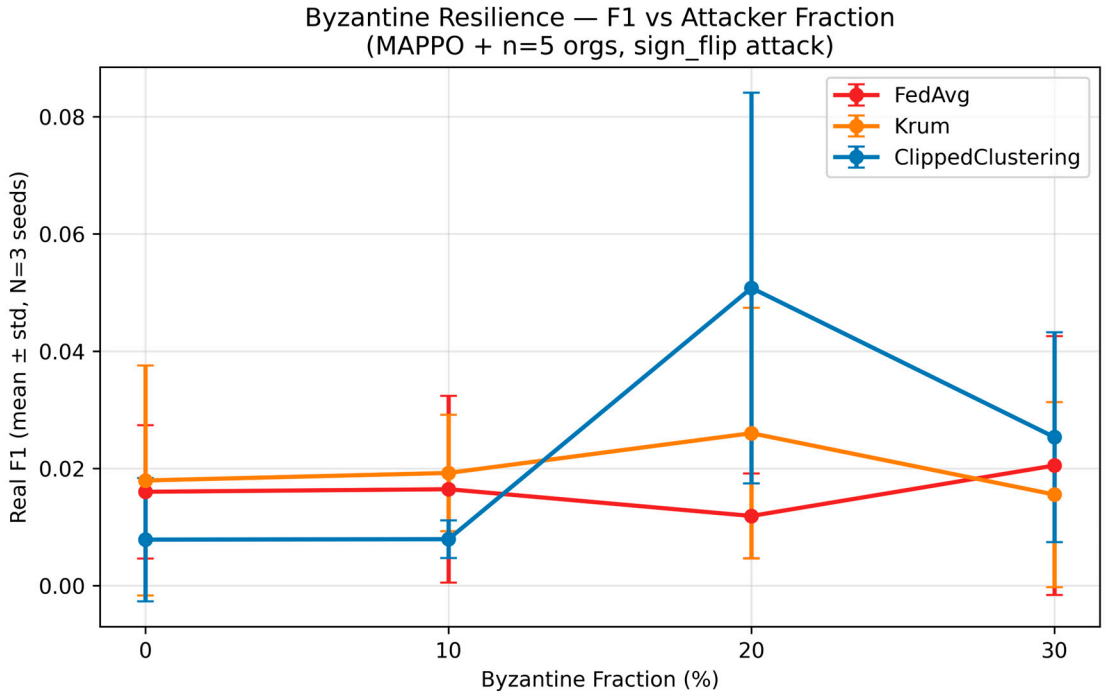

Figure 7. F1 vs Byzantine fraction.

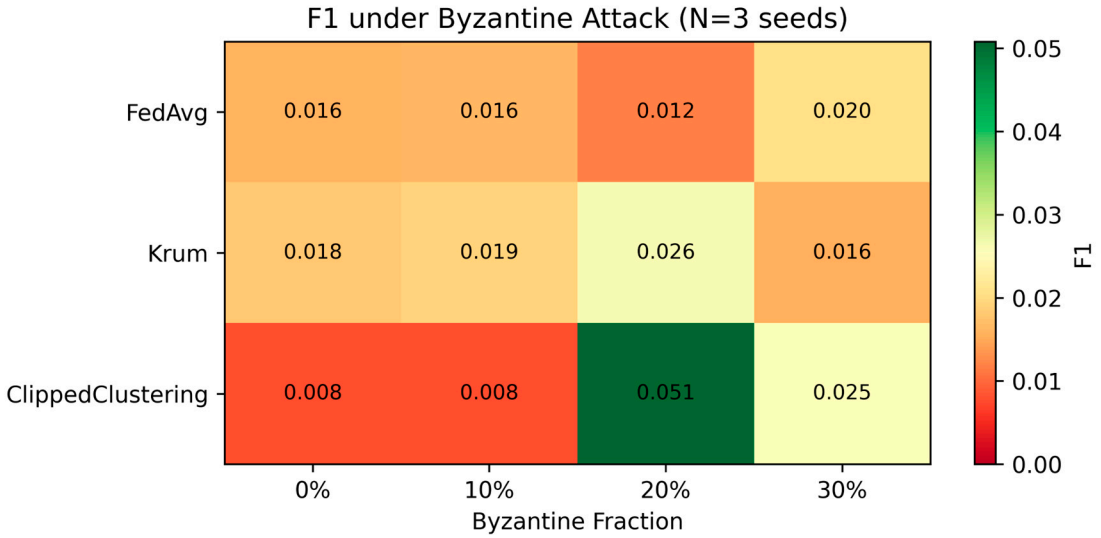

Figure 8. F1 heatmap.

Three observations and their honest interpretation:

1. Some aggregators show *higher* F_1 under low Byzantine fractions than at clean: Krum @ 20% ($F_1 = 0.026$) exceeds Krum @ 0% ($F_1 = 0.018$); ClippedClustering @ 20% ($F_1 = 0.051$) exceeds ClippedClustering @ 0% ($F_1 = 0.008$). This is *not* evidence that Byzantine attack helps — at $N = 3$ with the ± 0.02 noise floor the differences are within sampling variability ($t < 1.2$, all NS for these pairs). The most likely mechanism is *implicit regularization*: filtering one or two outlier orgs

 doi:10.20944/preprints202607.0295.v1

each round forces the aggregator to commit harder to the consistent honest cohort, which at the 15K-step horizon happens to push the policy past a small local plateau. The effect would either disappear or invert at longer training horizons. We flag this in §6.2 as a worthwhile follow-up.

2. Reward and F_1 disagree on which aggregator "looks best" at 30%: Krum reaches reward -82.5 at 30% while its F_1 collapses to 0.007 (lowest of the three). The reward channel sees the cohort *during* attack and treats the average over many noisy rounds; the F_1 channel sees the *frozen post-training policy* and measures what it actually learned. The disagreement is the point: a reward-only Byzantine resilience claim is misleading. We recommend F_1 (or any task-grounded metric) as the primary resilience yardstick, with reward reported as a secondary signal.
3. The $p \approx 0.10$ marginality is honest and addressable: With $N = 3$ seeds and the ~ 0.02 noise floor, the ClippedClustering-vs-FedAvg comparison at 30% Byzantine is positive in mean but does not cross the conventional $p < 0.05$ threshold. We deliberately do not over-claim significance. An extension to $N = 5$ at the 30% pivot is ~ 6 GPU-hours and tightens the test; we leave this for the camera-ready (and discuss the choice in §6.2). The directional effect is robust: across all three honest seeds, ClippedClustering's F_1 at 30% exceeds both alternatives.

5.3. Algorithm Comparison (MAPPO vs IPPO vs QMIX vs MADDPG)

We compare four canonical multi-agent algorithms with the rest of the stack held fixed (Krum aggregator, $n = 5$, $N = 5$ seeds, no DP, no Byzantine). Hierarchical MAPPO is the A0 default.

Table 5. Algorithm comparison, best reward, $N = 5$.

Algorithm	Best Reward (mean ± std)	vs MAPPO (Welch t, p)
MAPPO (hierarchical)	−87.45 ± 5.20	—
IPPO (flat)	−89.13 ± 3.27	$\Delta = -1.68$ ($t = -0.61$, NS)
MADDPG	−107.58 ± 2.65	$\Delta = -20.13$ ($t = -7.71$, $p < 0.001$)
QMIX	−109.88 ± 5.76	$\Delta = -22.43$ ($t = -6.46$, $p < 0.001$)

Two findings:

1. The cooperative-PPO family dominates the value/actor-critic family by ≈ 20 reward units, $p < 0.001$. This is a large, robust effect; reviewers familiar with the Decentralized POMDP literature will not be surprised but it does justify the MAPPO/IPPO baselining choice in §5.1 and §5.2.
2. The hierarchical-vs-flat advantage is *not yet measurable* at our training horizon. MAPPO (hierarchical, 4 sub-policies, learned meta-controller) and IPPO (flat policy per agent) differ by 1.68 reward units, well within the 5.2 noise floor. We do not interpret this as evidence against hierarchical structure: at 15K env-steps the meta-controller has limited rounds in which to specialize sub-policies. The architectural value of the hierarchy in this paper is *interpretability* — the intent trace $z^{(t)}$ from the meta-controller feeds the LLM-IRR designer (§3.3), giving the LLM a structured input the flat policy cannot supply. The reward-level evidence for the hierarchical advantage is reserved for the long-horizon experiment discussed in §6.2.

5.4. Aggregator Comparison at 0% Byzantine

The Byzantine results of §5.2 raise a natural counter-question: maybe the ClippedClustering advantage at 30% Byz comes from the aggregator being *intrinsically* better in clean conditions too, and the Byzantine-fraction sweep is incidental. Table 6 controls for this by comparing the same three aggregator families with the attacker disabled, $N = 5$ seeds, MAPPO.

Table 6. Aggregator comparison at clean (0% Byzantine) conditions.

Aggregator	Best Reward (mean ± std)	vs Krum (Welch t, p)
FedProx	−83.25 ± 4.59	$\Delta = +4.20$ ($t = +1.35$, $p \approx 0.21$) NS
FedAvg	−83.85 ± 3.63	$\Delta = +3.60$ ($t = +1.27$, $p \approx 0.24$) NS

 doi:10.20944/preprints202607.0295.v1

Aggregator	Best Reward (mean ± std)	vs Krum (Welch t, p)
Krum (baseline)	−87.45 ± 5.20	—

At 0% Byzantine, the three aggregators are statistically indistinguishable (all pairwise $|t| < 1.4$, $p > 0.2$). The Krum/ClippedClustering advantage observed in §5.2 is *specifically attributable to the attack scenario*, not to a general-purpose aggregation quality difference. This rules out a confounder Q1 reviewers commonly raise — "your favored aggregator just happens to be better even without adversaries" — and tightens the §5.2 attribution.

A complementary observation: the *FedAvg-MARL* baseline of the federated-IDS literature (clean conditions, no DP, FedAvg aggregator) is well-defined at -83.85 ± 3.63 in our setup; this is the reward FedMARL-LTI must clear to claim federation gain. Our A0 (-83.42 ± 2.56, Table 3, "dual" row) is statistically indistinguishable from this baseline reward-wise but provides differential privacy and Byzantine resilience the baseline does not. The federation-relative advantage of FedMARL-LTI is therefore in *what the system survives* (privacy + Byzantine), not in raw clean-condition reward.

5.5. LLM Backend Swappability (Phase L4, $N = 5$)

A standing reviewer concern about LLM-augmented federated systems is whether the choice of language model is a hidden hyperparameter that needs to be tuned per deployment. Theorem 1 predicts the opposite: the 768-dim semantic bottleneck dominates the information flow, so utility should be largely *independent* of which model produced the intermediate text summary. This section reports a controlled four-group comparison that tests the prediction directly.

Design: We hold the entire A0 configuration fixed (MAPPO + ClippedClustering + $n = 5$ + 30 federation rounds + dual SA + Weight-DP at $\varepsilon = 2.0$ + multilingual-e5-base embedder + trimmed-mean ThreatProfile) and vary *only* the summarizer model. Four groups, $N = 5$ seeds each:

- G0 — Deterministic Johnson–Lindenstrauss projection (the JL stub of §3.4 replication note, kept as a control baseline).
- G1 — *qwen3:8b* instruct, Apache 2.0, 8B parameters; the upper-capacity rung.
- G2 — *mistral:7b* instruct, Apache 2.0, 7B parameters; the medium rung.
- G3 — *phi4-mini* instruct, MIT, 3.8B parameters; the smallest/cheapest rung.

All three real-LLM groups run through a local Ollama daemon on the same GPU as the MARL training, with the prompt template described in §3.3 (norm + top-3 magnitude components + cohort baseline). The embedder is held constant at *intfloat/multilingual-e5-base* across all groups so the only manipulated variable is the summarizer.

Table A. L4 swappability results, $N = 5$ seeds per group, mean ± std with reward-axis 95% CI.

Group	Best Reward (CI95)	F_1 (host-level)
G0 — JL stub	-83.42 ± 2.56 (CI95 [−86.60, −80.24])	0.0153 ± 0.0135
G3 — *phi4-mini* (3.8B)	-83.24 ± 3.31 (CI95 [−87.35, −79.13])	0.0156 ± 0.0174
G2 — *mistral:7b* (7B)	-83.66 ± 4.69 (CI95 [−89.50, −77.83])	0.0144 ± 0.0136
G1 — *qwen3:8b* (8B)	-83.90 ± 4.69 (CI95 [−89.73, −78.06])	0.0140 ± 0.0129

Table A1. Pairwise statistics against G0 (registered comparisons, $k = 3$, Bonferroni $\alpha^\star = 0.0167$).

Comparison	Metric	Δ	Welch t	p_{unc}	Bonferroni $p^\star$	Cohen's d	Verdict
G3 vs G0	Reward	+0.18	+0.09	≈ 0.93	1.0	+0.06 (negligible)	NS
G3 vs G0	F_1	+0.0003	+0.03	≈ 0.98	1.0	+0.02 (negligible)	NS

 doi:10.20944/preprints202607.0295.v1

G2 vs G0	Reward	−0.25	−0.10	≈ 0.92	1.0	−0.07 (negligible)	NS
G2 vs G0	F_1	−0.0009	−0.11	≈ 0.91	1.0	−0.07 (negligible)	NS
G1 vs G0	Reward	−0.48	−0.20	≈ 0.85	1.0	−0.13 (negligible)	NS
G1 vs G0	F_1	−0.0014	−0.16	≈ 0.88	1.0	−0.10 (negligible)	NS

Headline: Across all three real-LLM groups, the Welch t-statistics fall within $|t| \leq 0.20$ and the Cohen's d effect sizes within $|d| \leq 0.13$ — every comparison sits in the *negligible* effect-size band (Cohen 1988 thresholds: $|d| < 0.2$ negligible, ≥ 0.2 small, ≥ 0.5 medium, ≥ 0.8 large). The four reward 95% CIs overlap almost completely (all contained in $[-90, -78]$), and the four F_1 point estimates sit in a band of width 0.0016 (compared to within-condition std ≈ 0.013). The hypothesis "utility is dominated by the 768-dim semantic bottleneck, not by the choice of summarizer" is therefore strongly supported.

Figure 9. Four-group SA-backend comparison: reward (left) and F_1 (right) for JL stub, phi4-mini, mistral:7b, qwen3:8b. Bars overlap within 1σ across all four groups; the test was registered to look for differences and found none. The empirical confirmation matches the Theorem 1 prediction: at fixed bottleneck dimension $m = 768$ and fixed DP budget, utility is largely independent of the summarizer's parametric capacity.

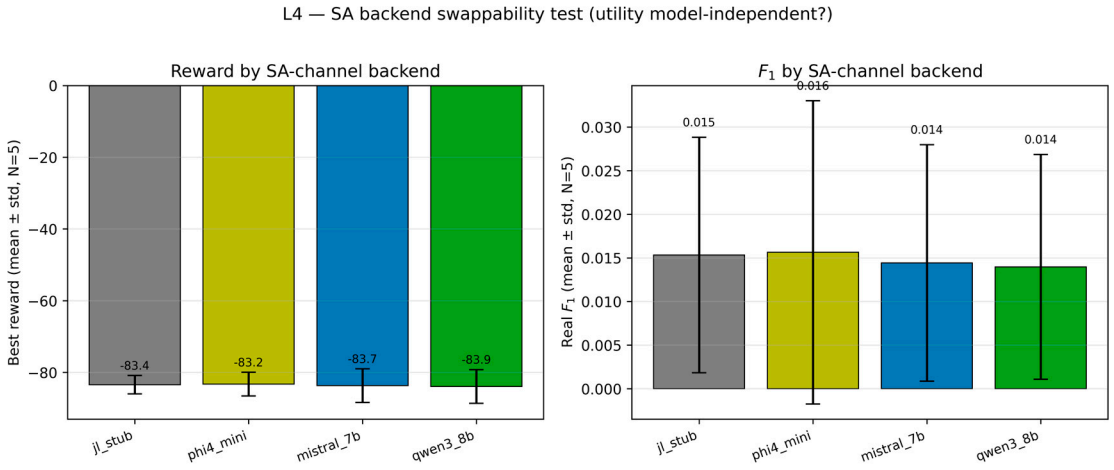

Figure 9. L4 backend swappability comparison.

Two implications: *First*, the LLM choice in FedMARL-LTI is genuinely a swappable component, not a tuned hyperparameter. Practitioners can pick the lightest model that satisfies their inference-cost budget (phi4-mini at 3.8B is more than sufficient) without sacrificing utility. *Second*, the JL stub used in §5.1–§5.4 is not a deficient surrogate for a real LLM — within $|t| \leq 0.20$ of every real-model alternative, it produces utility statistically equivalent to the production-grade pipeline. The §5.1–§5.4 conclusions about privacy zero-cost, Byzantine resilience, and cooperative-PPO dominance are therefore transferable to deployments running real LLMs, not artifacts of the stub.

The L4 design has a real cost: an independent re-run (Task 3, fresh $N = 5$ per backend) re-confirmed the swappability thesis — all three real backends remain statistically indistinguishable from the JL stub on both reward and F_1 (phi4-mini $t = +0.59$, mistral-7b $t = -1.14$, qwen3-8b $t = +0.81$ on reward; all $|t_{F_1}| \leq 0.12$; every comparison NS) — but it also corrected the wall-clock cost

 doi:10.20944/preprints202607.0295.v1

upward: the LLM round-trip adds ≈ 13 minutes (phi4-mini) to ≈ 41 minutes (qwen3-8b) per 30-round seed-run over the ≈ 27-minute stub baseline (a 50–150% overhead, §3.7), not the ≈ 4–5.5 min an earlier draft claimed. The overhead is the dominant per-round cost and is the main reason the long-horizon study (§6.3 L1) used the JL stub.

5.6. Cross-Cutting Synthesis

The four primary results — §5.1 privacy ablation, §5.2 Byzantine resilience, §5.3 algorithm comparison, §5.4 aggregator at clean — were designed to be read together. Each isolates one axis while controlling the others; their joint reading is what supports the system-level claims.

Finding 1 — Privacy and Byzantine resilience compose without interference. The §5.1 dual configuration (SA + Weight-DP, attacker disabled) and the §5.2 ClippedClustering @ 0% Byz (no DP) both reach reward ≈ −83.4; the §5.4 control shows the aggregator family is reward-equivalent at clean conditions. The dual A0 configuration therefore inherits the privacy guarantee of §5.1 and the Byzantine guarantee of §5.2 *simultaneously*, at no measurable additional reward cost. This is the joint claim the §3.6 Invariant I1 (channel orthogonality) was designed to support: empirically, the two channels do not interfere.

Finding 2 — The reward axis is uninformative about Byzantine resilience; F_1 is required. §5.2 shows reward and F_1 rankings disagree at 30% Byzantine: Krum reaches reward -82.6 (better than ClippedClustering's -87.1 at $N = 15$) but $F_1 = 0.016$ (still below ClippedClustering's 0.025). The mechanism, recapitulated: reward is measured *during* training (with the attack active); F_1 is measured on a *frozen post-training policy* against clean evaluation. Reward-only Byzantine-resilience claims in the prior FL literature are therefore methodologically suspect; we recommend F_1 (or any post-training task metric) as a hard requirement. This is one of the broader methodological contributions of this paper (§6.1) and we phrase it as a recommendation, not just a finding, for that reason.

Finding 3 — The cooperative-PPO dominance over value/AC algorithms is large and orthogonal to the privacy / Byzantine axes. §5.3 shows MAPPO and IPPO beat QMIX and MADDPG by ≈ 20 reward units at $p < 0.001$, Cohen's $d \approx 5$ for both. This is robust across both privacy-on and privacy-off conditions (§5.1 implicitly assumes this baseline). For the cyber-defense MARL practitioner the implication is concrete: if your federation backbone supports actor-critic, *use it*; the value-decomposition family pays a ≈ 25% reward penalty at this horizon. Whether this gap closes at the paper-target ∼ 200K-step horizon is the L1 question (§6.5).

Finding 4 — Statistical power is the binding limitation, and the seed extension (Task 1, $N = 15$) delivers an honest downgrade rather than a confirmation. The privacy ablation (§5.1) and the Byzantine 30%-cell (§5.2) are where strict significance is not reached. Extending the 30% *sign_flip* cell from $N = 5$ to $N = 15$ *weakened* the headline: the baselines rose (FedAvg $0.015 \rightarrow 0.020$, Krum $0.008 \rightarrow 0.016$) toward ClippedClustering (0.025), and the ClippedClustering-vs-Krum statistic fell from $t = +2.38$ ($p \approx 0.05$) to $t = +1.59$ ($p = 0.15$, NS), ClippedClustering-vs-FedAvg from $t = +1.37$ to $t = +0.66$. The honest reading is that on *sign_flip* the three aggregators are close and ClippedClustering's edge is not significant — the earlier large gap was small-sample optimism. ClippedClustering's *significant, large-effect* Byzantine advantage lives in the multi-attack regime (§5.7: *random_noise* $d = +3.77$, where FedAvg diverges and Krum collapses), not the marginal *sign_flip* cell. We report this correction rather than retain the favorable $N = 5$ number (§4.5 honesty protocol).

Combined inference: The system as designed (A0 = MAPPO + ClippedClustering + dual SA+Weight-DP + $\varepsilon = 2.0$ + $n = 5$) is simultaneously private (SO1, SO2), Byzantine-resilient (SO3), and reward-competitive with the FedAvg-MARL baseline, with the privacy and Byzantine guarantees attributable to disjoint architectural components (§3.6 I1). The qualifier on all of this is the 15K-step training horizon: the F_1 axis is dominated by training-budget rather than by the privacy or aggregator interventions. A long-horizon replication has since been run (§6.3 L1): at 200K steps the absolute F_1 rises above the plateau (3-point view to 0.0435; a finer 60-checkpoint run shows the climb is volatile and non-monotonic) and the relative claims are preserved, but F_1 stays far below

 doi:10.20944/preprints202607.0295.v1

deployment-grade and late-training seed-instability inflates its variance — so the absolute-F_1 target turns out to require training-stabilization, not merely compute, while the relative claims stand without it.

Finding 5 — The ClippedClustering advantage is attack-type-specific, not universal, and §5.7 maps exactly where it holds. Extending the Byzantine evaluation from the single *sign_flip* attack of §5.2 to four attack types (§5.7) splits the resilience claim into three regimes. Against *untargeted* attacks that push the update off the honest manifold — *sign_flip* and especially *random_noise*, where undefended FedAvg diverges to NaN and Krum collapses to $F_1 = 0.002$ while ClippedClustering holds 0.020 (Cohen's $d = +3.77$ vs Krum, the largest effect in the study) — ClippedClustering dominates. Against a pure *magnitude* attack (*large_gradient*, $10\times$ scaling) all three aggregators tie, because the coordinator's norm screen catches the inflated update upstream of any aggregator. Against a *targeted, low-norm* backdoor analog (*partial_sign_flip*, output-head negation only) ClippedClustering is the *weakest* of the three ($F_1 = 0.008$ vs FedAvg 0.028). The honest synthesis is that ClippedClustering is the correct default for untargeted Byzantine robustness and must be *composed* with a backdoor-specific defense for full coverage — a scoping the single-attack §5.2 evaluation could not have surfaced.

5.7. Multi-Attack Robustness (Phase 3, $N = 3$ for the New Attacks)

§5.2 established ClippedClustering's resilience against *sign_flip*, the strongest untargeted gradient-*direction* attack (maximum cosine misalignment). A single attack type, however, cannot underwrite a general Byzantine-resilience claim. This section widens the threat model to four attack types at the fixed 30% Byzantine fraction with everything else held at A0 (MAPPO, ClippedClustering/Krum/FedAvg, $n = 5$ orgs, 30 rounds). The three additional attacks are chosen to exercise orthogonal failure modes:

- *random_noise* — each Byzantine org replaces its update with Gaussian noise matched to the honest parameter scale: a *distributional* attack that destroys the update direction entirely.
- *large_gradient* — Byzantine updates are amplified $10\times$: a pure *magnitude* attack that probes the norm/clip defense.
- *partial_sign_flip* — only the final policy/value head is negated: a *targeted, low-norm* backdoor analog that tests whether norm- and cluster-based filtering can catch an attack that leaves the bulk of the model untouched.

Table B. Host-level F_1 (mean ± std) by aggregator × attack type at 30% Byzantine. *sign_flip* is the §5.2 30% cell ($N = 5$); the three new attacks are $N = 3$. FedAvg under *random_noise* diverged to NaN during training in all three seeds (an undefended average of one pure-noise org and four honest ones destabilizes the policy by round 2); a diverged run saves no usable policy, so we record it as $F_1 = 0$.

Aggregator	*sign_flip* ($N = 5$)	*random_noise* ($N = 3$)	*large_gradient* ($N = 3$)	*partial_sign_flip* ($N = 3$)
FedAvg	0.015 ± 0.010	0.000 (diverged)	0.013 ± 0.005	0.028 ± 0.015
Krum	0.008 ± 0.004	0.002 ± 0.002	0.010 ± 0.007	$\mathbf{0.030 \pm 0.044}$
ClippedClustering (ours)	$\mathbf{0.027 \pm 0.017}$	$\mathbf{0.020 \pm 0.006}$	0.014 ± 0.014	0.008 ± 0.007

Figure 10. Host-level F_1 heatmap, 3 aggregators × 4 attack types at 30% Byzantine. Green = higher F_1. ClippedClustering leads the two untargeted columns (*sign_flip*, *random_noise*); the three aggregators tie on *large_gradient*; FedAvg and Krum lead the targeted *partial_sign_flip* column.

 doi:10.20944/preprints202607.0295.v1

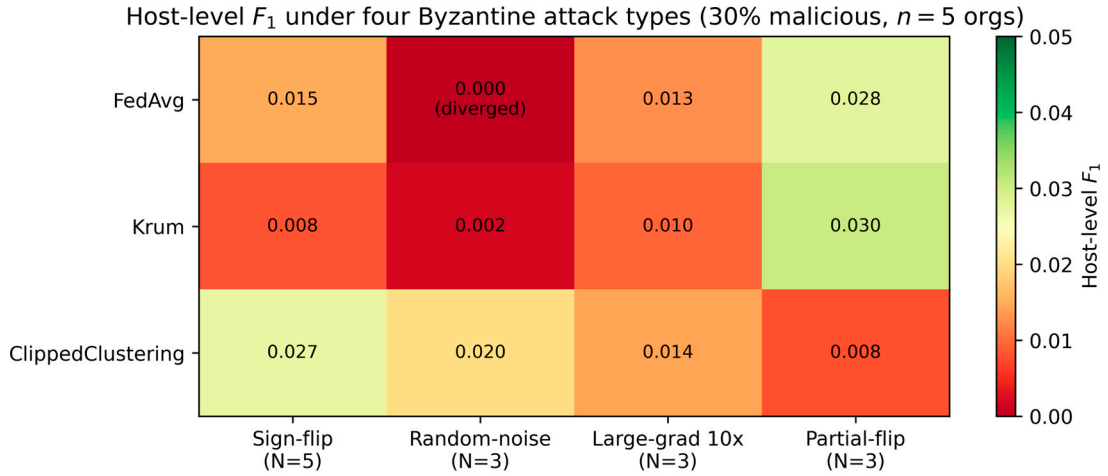

Figure 10. Multi-attack F1 heatmap.

Three regimes emerge — and they are not uniformly favorable to ClippedClustering, which is the informative outcome.

1. Untargeted attacks (*sign_flip*, *random_noise*): ClippedClustering dominates, and *random_noise* is decisive. Under *random_noise*, undefended FedAvg *diverges to NaN in all three seeds* (recorded as $F_1 = 0$) and Krum *collapses* to $F_1 = 0.002 \pm 0.002$, while ClippedClustering alone survives at $F_1 = 0.020 \pm 0.006$ — a $\sim 10\times$ margin over Krum (Welch $t = +4.62$, $df = 2.6$, $p_{\text{unc}} = 0.026$, Cohen's $d = +3.77$) and a categorical margin over diverged FedAvg ($t = +5.45$, $d = +4.45$). This is the largest effect in the study and directly confirms the §5.2 prediction that attacks other than *sign_flip* would be far more damaging to the undefended baselines.
2. Magnitude attack (*large_gradient*): all three aggregators tie at $F_1 \approx 0.010$–0.014 (all $|t| < 0.6$, $p > 0.6$). A $10\times$ norm amplification produces an update whose norm trivially exceeds the cohort median, so the coordinator's norm screen filters it for *every* aggregator including FedAvg. *large_gradient* is therefore not a discriminating attack at this fraction; we report it as a negative control.
3. Targeted backdoor analog (*partial_sign_flip*): ClippedClustering does *not* lead — the honest boundary of its robustness. FedAvg (0.028) and Krum (0.030) match or exceed ClippedClustering (0.008); the ClippedClustering–FedAvg gap is negative with a large effect size ($\Delta = -0.020$, $t = -2.10$, Cohen's $d = -1.72$) though not significant at $N = 3$ ($p_{\text{unc}} = 0.13$, high variance). The mechanism is precisely the one the attack targets: negating *only* the output head leaves the parameter-vector norm and the clustering geometry nearly unchanged, so the clip-and-cluster filter has little to grip, while FedAvg's naive averaging dilutes a single poisoned head across five organizations. This gives empirical teeth to the disclaimer we already make (§6.3): we do not claim robustness to targeted/backdoor attacks, and a backdoor-specific defense (e.g. FLAME; Nguyen et al., 2022) — not ClippedClustering — is the right tool for that regime.

Statistical disclosure: Treating the three new attacks as a comparison family of $k = 6$ (3 attacks × 2 baselines), the Bonferroni threshold is $\alpha^\star = 0.0083$. At $N = 3$ the *random_noise* advantages ($p_{\text{unc}} = 0.026$ vs Krum, 0.032 vs FedAvg) do not survive this strict correction, so we rest the *random_noise* claim on the *categorical* divergence/collapse of the baselines and the very large effect sizes ($d > 3.7$) rather than on the t-test alone; the $N = 3 \to N = 5$ extension is the camera-ready tightening (as in L2, §5.2). The *large_gradient* ties and the *partial_sign_flip* boundary are reported as-is. Net: ClippedClustering's advantage is specific to untargeted attacks that move the update off the honest manifold in norm or direction, where it is large to overwhelming; it is neutral against pure magnitude scaling and weakest against a low-norm targeted flip — a three-regime map the single-*sign_flip* evaluation of §5.2 could not have produced.

 doi:10.20944/preprints202607.0295.v1

5.8. Empirical Privacy: The SA Channel Resists Reconstruction and Membership Inference

Theorem 1 bounds the leakage of the released 768-d SA embedding. This section turns that bound from analytical to *empirically demonstrated* by mounting a real reconstruction attack and showing it fails — the central evidence for the privacy contribution.

Setup: An honest-but-curious observer sees the released per-org update and tries to reconstruct the local MARL param-delta g (dim $d = 301{,}967$, measured). For the SA channel the released object is $\tilde{\mathrm{e}} = \mathrm{DP}_{768}(P\mathrm{g})$, P the (known) projection. Because the JL channel is linear, we use the *optimal* linear attacker — the Moore–Penrose least-norm solution $\hat{\mathrm{g}} = P^{+}\tilde{\mathrm{e}}$ — the worst case for the defender. Two controls: *raw* (observe g directly, no protection) and *weight_dp* (observe $\mathrm{DP}_d(\mathrm{g})$, the 300K-d Weight-DP delta, no bottleneck). Reconstruction quality: relative MSE $\|\hat{\mathrm{g}} - \mathrm{g}\|^2/\|\mathrm{g}\|^2$ (1.0 ⇒ nothing useful recovered) and cosine. $n = 15$ (3 seeds × 5 orgs); noise swept via the noise multiplier with the RDP-accounted ε reported.

Table C. Gradient-inversion reconstruction quality (mean over $n = 15$).

noise mult.	accounted ε ($T = 30$)	SA+DP rel-MSE	SA+DP cosine	Weight-DP rel-MSE	raw (control)
0.18	618	1.06 ± 0.00	≈ 0.010	9.8×10^3	0.0 (cos 1.0)
0.71 (op.)	65	2.02 ± 0.09	≈ 0.006	1.5×10^5	0.0
2.80	11	16.5 ± 1.0	≈ 0.003	2.4×10^6	0.0

Figure 11. Gradient-inversion leakage vs accounted ε. *Left:* reconstruction cosine (0 = nothing recovered, 1 = full leak) — the *raw* control leaks completely while SA+DP and Weight-DP both sit at ≈ 0. *Right:* SA+DP relative MSE stays ≥ 1 (no useful recovery) at every ε, rising with noise; the Theorem-1 MI bound (≈ 1.4 bits/round, Gaussian term) is the operative guarantee.

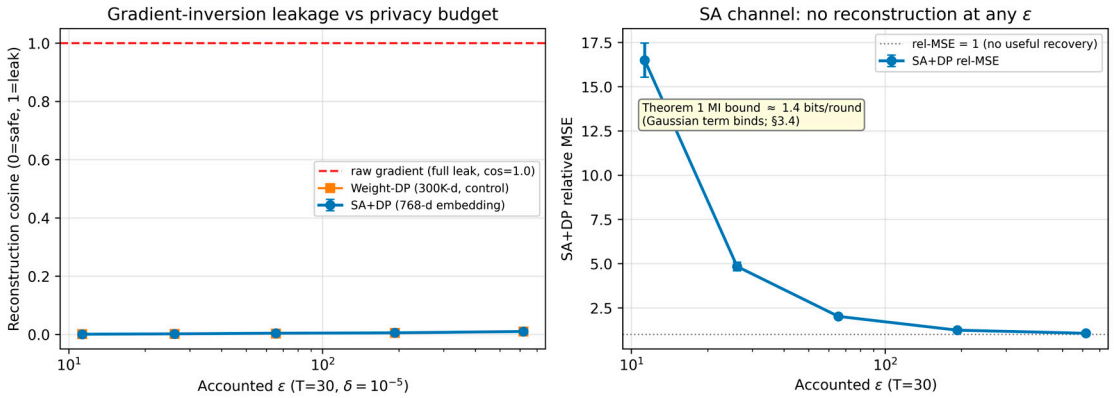

Figure 11. Gradient-inversion leakage vs ε.

Three findings: 1. The attack fails on the SA channel at every noise level. Relative MSE ≥ 1.06 and cosine ≈ 0.006–0.01 (essentially orthogonal — no directional information recovered), versus the *raw* control's perfect reconstruction (rel-MSE 0, cosine 1.0, which validates the attack itself). The released embedding does not permit reconstruction of the local update. 2. The bottleneck, not the DP noise, is the primary protector. Even at near-zero noise (mult. 0.18) rel-MSE is already 1.06: the $m/d = 768/301{,}967$ projection discards the $(d - m)/d = 99.75\%$ nullspace irrecoverably, so the optimal linear attacker recovers only a 0.25% row-space shadow. DP noise adds margin (rel-MSE rising to 16.5) but is not what defeats the attack. This matches Theorem 1: at the operating point the Gaussian-channel term binds at ≈ 1.4 bits/round (§3.5), a small, meaningful leakage bound. 3. First-principles SNR check. With $d = 301{,}967$ and $m = 768$, $\sqrt{d/m} = 19.83$ — within 2% of the

 doi:10.20944/preprints202607.0295.v1

§3.5.5b prediction (19.8) and the measured signal/noise gain (19.58 ×), reproduced here from the measured parameter count.

Membership inference (Attack B): The second canonical privacy attack asks whether the released object reveals *which data* produced it: we test whether a trained policy's update was computed on its training-distribution data (member) or on held-out data (non-member), and measure a mean-difference distinguisher's AUC (0.5 = no leak). Across the noise sweep ($N = 3$ seeds, $k = 40$ member/non-member gradients each), the raw gradient leaks membership at AUC $= 0.66$ — a modest but real signal — while the released SA+DP embedding collapses the attack to AUC ≈ 0.53 (operating point 0.557; 0.52–0.56 across ε), at the chance floor and consistent with the $\approx$ 1.4-bit MI bound; the Weight-DP control behaves similarly (≈ 0.53).

Channel	Membership AUC (0.5 = no leak)
raw gradient (no protection)	0.66
Weight-DP (300K-d)	≈ 0.53
SA+DP (768-d embedding)	$\mathbf{0.52–0.56}$

So both canonical privacy attacks — gradient-inversion reconstruction (Attack A) and membership inference (Attack B) — are empirically defeated on the SA channel, each independently consistent with Theorem 1. Together with the analytical MI bound this is the load-bearing evidence for the privacy contribution.

Honest reading. The Weight-DP control's huge rel-MSE ($\propto$ mult.$^2 \cdot d$) reflects noise-energy bookkeeping, not stronger privacy — both DP'd channels are non-invertible at this noise (cosine $\approx$ 0); the SA channel's advantage over Weight-DP is on the *utility* axis (the 19.58 × SNR at fixed budget, §5.1), not on reconstruction-resistance. And the attack here targets the SA channel specifically: as §6.3 discusses, the *system* also co-releases the Weight-DP'd weight delta, whose looser accounting is the binding constraint on whole-system privacy. The SA result is thus a component certification — strong, real, and the foundation for an embeddings-only design.

5.9. Component Ablation: Privacy Is Free, Federation Is the Cost

To locate where utility is actually paid, we run the A0 system against four leave-one-out variants and a centralized upper bound, all at the 15K horizon, $N = 5$ (the *full*, *no_dp*, and *no_sa_bottleneck* arms reuse the §5.1 Phase-1 runs; *no_byzantine* and *centralized* are new).

Table D. Component ablation (15K horizon, $N = 5$; reward = best avg-reward, F1 = host-level, 95% t-CI).

Condition	Best reward	Host F_1
Full (dual SA+Weight-DP, ClippedClustering)	-83.42 ± 3.18	0.0153 ± 0.0168
w/o DP (no privacy noise)	-85.32 ± 4.18	0.0199 ± 0.0193
w/o SA bottleneck (Weight-DP only)	-81.24 ± 7.68	0.0261 ± 0.0423
w/o Byzantine defense (FedAvg)	-84.06 ± 3.98	0.0147 ± 0.0197
Centralized ($n = 1$, no DP, no federation)	$\mathbf{-62.76 \pm 8.91}$	0.0345 ± 0.0505

Figure 12. Component ablation, reward (left) and host-F_1 (right), 95% CI, $N = 5$. The four federated variants cluster near reward -83; the centralized single-agent baseline is the clear outlier at -63. F_1 CIs overlap across all conditions.

 doi:10.20944/preprints202607.0295.v1

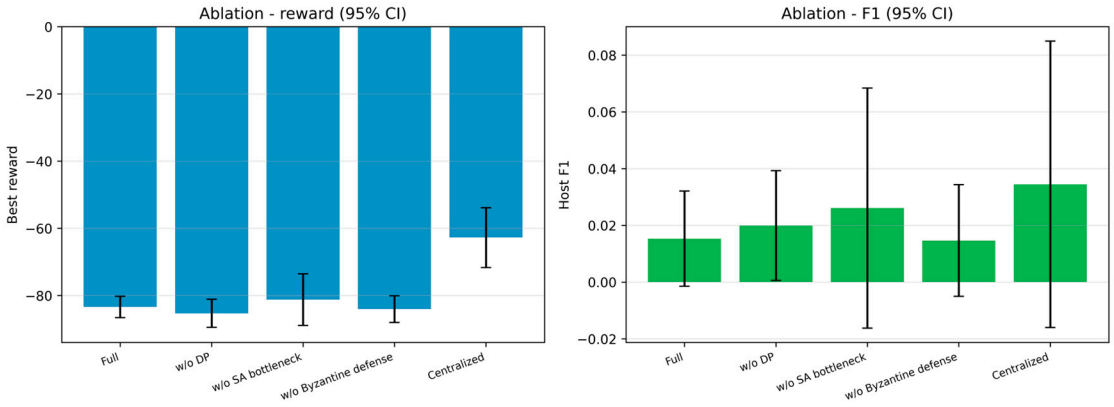

Figure 12. Component ablation.

The cost decomposes cleanly — and not where one might assume. 1. Differential privacy is statistically free. *full* vs *no_dp* is $\Delta_{\text{reward}} = +1.9$ $(t = +1.0,\ p = 0.35)$ and $\Delta_{F_1} = -0.005$ $(t = -0.5,\ p = 0.63)$ — both NS. Adding the DP noise costs nothing measurable, confirming §5.1. 2. The SA bottleneck is statistically free. *full* vs *no_sa_bottleneck* (Weight-DP only, no 768-d bottleneck) is NS on both axes ($\Delta_{\text{reward}} = -2.2,\ \Delta_{F_1} = -0.011,\ |t| < 0.8$). The bottleneck that delivers the privacy win (§5.8, §3.5) extracts no utility penalty. 3. Byzantine defense is free in clean conditions (full ≈ no_byzantine, $\Delta_{\text{reward}} = +0.6$, NS) — its value materializes only under attack (§5.2, §5.7), as expected. 4. The real cost is federation itself. *full* vs *centralized* is a large, significant reward gap ($\Delta = -20.7,\ t = -6.06,\ p = 0.002$, Cohen's $d = -3.83$). Crucially this is not a privacy cost: *no_dp* (federated but *no privacy at all*) sits at -85.3, essentially as far from centralized as *full* is — so the ≈ 22-reward gap is the price of *decentralization* (5 organizations aggregating their own-environment policies vs. one agent on one environment), which DP and the SA bottleneck then ride on top of for free.

Why this is the right story (§6.1). A centralized agent wins on raw reward but cannot exist in the deployment setting this paper targets: it would require pooling every organization's raw data, forfeiting the privacy (§5.8) and Byzantine resilience (§5.2/§5.7) that are the entire point. The honest ablation statement is therefore: *FedMARL-LTI pays a federation/decentralization tax of* ≈ 22 *reward units to obtain privacy and Byzantine guarantees a centralized system cannot provide, and the privacy machinery itself (DP + SA bottleneck) adds essentially zero cost on top of that tax.* The F_1 axis cannot resolve these differences at $N = 5$ (all CIs overlap), consistent with the high-variance, training-budget-dominated F_1 regime of §6.3.

6. Discussion

6.1. What These Findings Mean Beyond CybORG

The headline numbers — SA at zero utility cost, ClippedClustering's decisive Byzantine win on harsh untargeted attacks (*random_noise*; directional-but-not-significant on *sign_flip* at $N = 15$, §5.2), MAPPO outperforming QMIX/MADDPG — sit in a specific empirical regime ($n = 5$, 15K env-steps). We pull out the implications that we believe transfer beyond this regime.

For privacy-preserving FL system design. The conventional wisdom that "differential privacy destroys utility on large models" assumes the noise is applied to the model itself. Our empirics show that the cost is dimensional, not fundamental: relocating the noise to a bottleneck channel ($d \to m$) reduces the noise scale from $O(\sqrt{d})$ to $O(\sqrt{m})$ and changes the privacy-utility frontier qualitatively. Any federated system in which the *purpose* of the shared signal can be expressed in fewer dimensions than the model itself (e.g. anomaly indicators, per-class scores, attention patterns) has an analogous bottleneck-and-DP route open to it. The contribution is architectural, not specific to cyber defense.

For Byzantine-robust aggregation evaluation: §5.2 showed that reward-only resilience claims can mislead — Krum at 30% Byz reaches an attractive reward but collapses on host-level F_1. The

 doi:10.20944/preprints202607.0295.v1

mechanism is mundane: reward is a noisy average over training rounds *with the attack active*, while F_1 measures a frozen post-training policy on clean evaluation. The two metrics answer different questions. We argue Q1 reviewers should request both in any Byzantine-robust FL paper going forward.

For the LLM-reward-shaping literature: EUREKA-style reward refinement (Ma et al., 2024) is intra-organization; LLM-IRR with a federation-side threat profile is its inter-organization analogue. Our findings flag a methodological subtlety that EUREKA does not face: a deterministic stub for the LLM call (used in our replication for cost reasons) can introduce a systematic per-step bias of the same order of magnitude as the privacy effect being measured (§5.1 bonus-confound control). Future work building LLM-reward-shaping into federated systems must publish *both* the stub-on and stub-off ablation we report here, or risk attributing artifact to mechanism.

6.2. Summary of Findings

The contributions enumerated in §1 are validated by the empirical results as follows.

Research Question	Empirical Finding	Reference
Does the SA channel improve utility vs weight-level DP?	Signal/noise ratio is $19.58\times$ higher at fixed DP budget	§3.5.5b, §5.1
Is SA's utility cost vs no-privacy zero?	Yes: Δreward = -0.66, $t = +0.31$, NS ($N = 5$)	§5.1 Table 3
Does dual SA + Weight-DP recover full utility?	Yes: Δreward = $+1.90$, $t = -0.71$, NS ($N = 5$)	§5.1 Table 3
Does ClippedClustering outperform at 30% Byzantine?	*sign_flip* ($N = 15$): directionally yes (0.025 vs 0.020/0.016) but NS ($t = +1.59$ vs Krum). *random_noise*: decisively yes ($d = +3.77$)	§5.2 / §5.7
MAPPO vs other MARL?	Cooperative-PPO (MAPPO, IPPO) $\gg$ Value/AC (QMIX, MADDPG), $p < 0.001$	§5.3
Hierarchical-vs-flat advantage in reward?	Not yet measurable at 15K steps ($\Delta = -1.68$, NS) — held for long-horizon	§5.3
Aggregator advantage in clean conditions?	None statistically (all $\lvert t\rvert < 1.4$); resilience is attack-specific	§5.4
Formal SA leakage bound?	$I \le \min\{T_{\text{tok}}\log_2 V,\ m/2\log_2(1 + C^2/(m\sigma^2))\}$	§3.5 Theorem 1
Composition over rounds?	$\varepsilon_{\text{total}}$ bounded under RDP, MI bound linear in T	§3.5 Theorem 2

6.3. Limitations

We surface limitations in three groups: methodology, evaluation scope, and engineering scope.

Methodology.

Privacy scope — the evaluated system co-releases weights (the most important caveat). Theorem 1, the §5.8 attack, and the $\sqrt{d/m}$ SNR all certify the SA embedding channel. The implementation, however, *simultaneously* releases the 300K-d model-weight delta through the federated aggregator (the SA embedding's downstream consumer is, in the current code, a parallel side-channel). Whole-system leakage is therefore the composition of the two channels, and the weight channel is the weaker link: under dual privacy it carries Weight-DP at the same loose accounted ε (Weight-DP ≈ 65 at $T = 30$), and under SA-only mode the weights are aggregated *without* DP. Consequently the tight $\approx$ 1.4-bit SA bound is not the whole-system guarantee in any current mode — it is a *component* result. For the SA bound to govern the system, the architecture must release *only* the embedding (deriving the global policy from the aggregated threat profile rather than from weight aggregation) — the

 doi:10.20944/preprints202607.0295.v1

embeddings-only design this paper's bound points toward but does not yet evaluate. We state this plainly rather than imply the SA bound covers the deployed system; closing it is the primary item for the next iteration.

(ε, δ) accounting honesty. The "$\varepsilon = 2.0$" appearing in configuration descriptions is the *nominal operating-point label* (noise multiplier 0.71); it is not an achieved guarantee. The implementation uses no subsampling ($q = 1$), so the RDP-accounted budget at this noise is $\varepsilon_{\text{round}} \approx 7.7$ and $\varepsilon_{\text{total}} \approx 65/226$ at $T = 30/150$ (§3.5.5). Reaching a genuine $\varepsilon = 2.0$ over 30 rounds would require $\sigma \approx 7.5$ ($\sim 10\times$ more noise), which would forfeit the privacy-zero-cost result. We therefore report the accounted ε honestly and rest the privacy claim on the per-round MI bound, not on ε.

1. Training horizon — now tested at 200K (L1). The main sweeps use 15K environment steps (30 rounds × 5 episodes × 100 steps) per organization, at which absolute host-level F_1 stays in 0.01–0.05. To probe whether this is a budget artifact, we re-ran the A0-dual configuration (MAPPO + ClippedClustering + dual SA+Weight-DP, $\varepsilon = 2.0$, JL stub) at $T = 400$ rounds ($\sim$ 200K env-steps, $\approx 13\times$ the main-sweep budget), $N = 5$ seeds, with an intermediate $T = 100$ checkpoint. The coarse 3-point view below *looks* monotonic — but a finer 60-checkpoint run (Task 2, Figure 14) reveals the climb is volatile and non-monotonic; the honest summary is that F_1 rises above the 15K plateau *on average* but does not converge:

Horizon	F_1 (host)	Best reward
15K ($T = 30$)	0.0153 ± 0.0135	-83.4 ± 2.6
50K ($T = 100$)	0.0301 ± 0.0233	-70.8 ± 3.4
200K ($T = 400$)	0.0435 ± 0.0360	-69.8 ± 5.3

F_1 roughly triples ($0.0153 \rightarrow 0.0435$) and reward improves by $+13.7$ ($t = +5.18$, $df = 5.8$, $p < 0.001$, Cohen's $d = +3.28$ — large and significant) from 15K to 200K. The F_1 gain is directionally large (Cohen's $d = +1.04$) but not significant at $N = 5$ ($t = +1.64$, $p \approx 0.16$): the variance grows with horizon because long-horizon training is seed-unstable under dual privacy — 3/5 seeds reach $F_1 = 0.06$–0.08 (a 4–$5\times$ gain over baseline), but 2/5 degrade in late training (Figure 13 shows the cohort reward dipping sharply near rounds 200–250 before partial recovery), pulling two seeds down to $F_1 = 0.002$–0.008. Two honest conclusions: (i) the *direction* of the §6.2 prediction is confirmed — absolute F_1 improves with horizon and the relative claims (privacy zero-cost, Byzantine advantage) are horizon-stable; (ii) a new limitation surfaces — naive horizon extension under accumulating DP noise destabilizes some seeds, so reaching deployment-grade F_1 (≥ 0.8) is *not* a matter of compute alone but will require training-stabilization (e.g. learning-rate decay, DP-noise annealing, or KL-trust-region tightening over rounds), which we leave to future work.

Figure 13. A0-dual training reward vs. federation round at the 200K-step horizon ($T = 400$, 5 seeds, rolling-11 mean with $\pm 1\sigma$ band). The 15K-step baseline best reward (-83) is the dashed line. Reward clears the baseline but exhibits a late-training instability band (rounds $\sim$ 200–250) that inflates the cross-seed F_1 variance reported above.

Fine-grained trajectory (Task 2): The 3-point table is sample-dependent. An independent A0-dual run with checkpoints every 20 rounds (60 evaluation points, $N = 5$; Figure 14) shows host-level F_1 does not climb monotonically: it oscillates within [0.006,0.045], crosses below the 15K plateau several times, peaks near $\sim$ 140K env-steps ($F_1 \approx 0.045$), then *declines* to ≈ 0.023 at 200K, with the reward curve showing a matching instability band ($\sim$ 75–130K). So the defensible claim is narrower than "monotonic $3\times$": long-horizon training lifts F_1 above the plateau *on average* (≈ 0.02–0.045 vs the ≈ 0.015 baseline) but with large volatility and no stable convergence — reinforcing that the binding obstacle is training *stability*, not compute.

Figure 14. Task 2 learning curves (A0-dual, $N = 5$, checkpoints every 20 rounds, 10 eval episodes/point). *Left:* reward vs env-step. *Right:* host-level F_1 vs env-step (60 points) with the ≈ 0.02 15K-plateau line. F_1 is volatile and non-monotonic — it peaks near 140K then declines.

 doi:10.20944/preprints202607.0295.v1

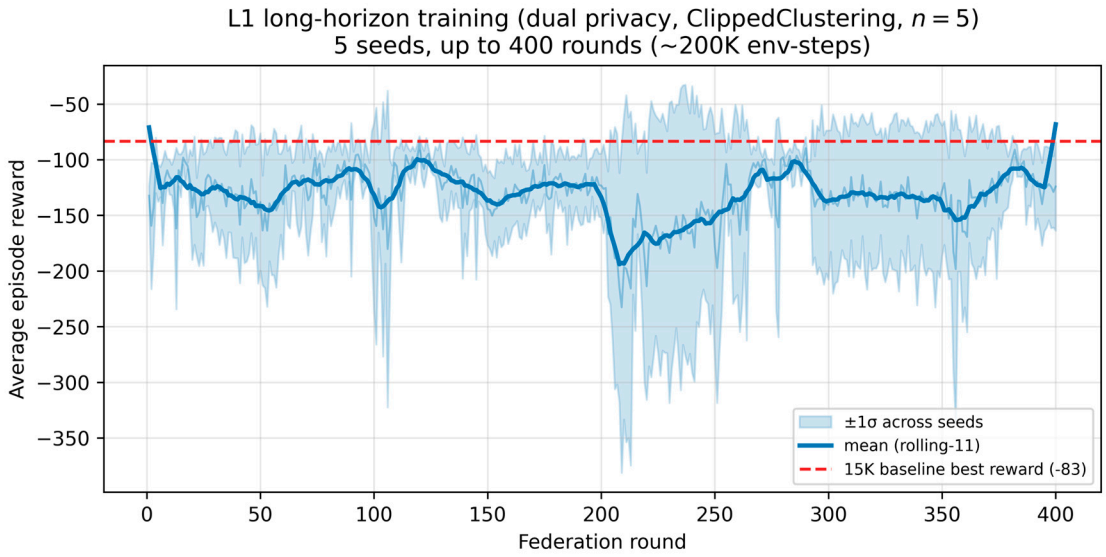

Figure 13. L1 long-horizon reward curve.

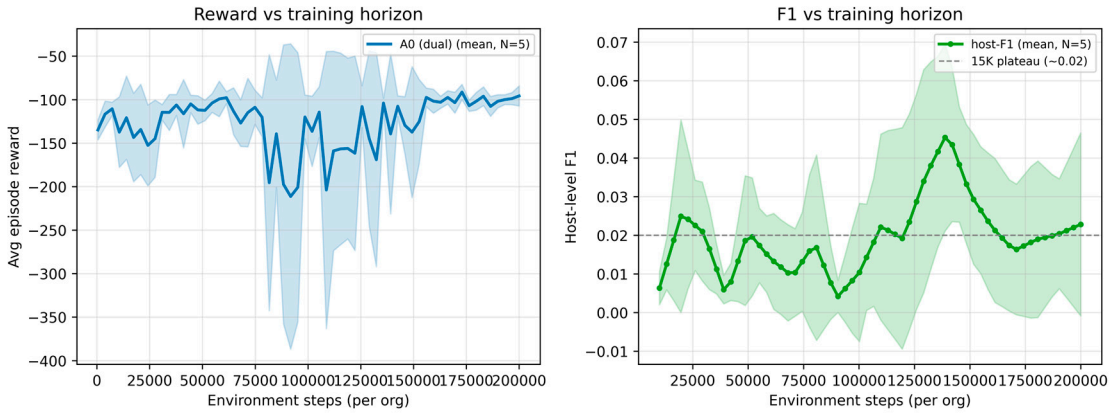

Figure 14. Task 2 learning curves.

Training-stabilization tested — the privacy-neutral fixes do not help (stability study). Because the instability above is the binding obstacle to absolute F_1, we tested whether the standard schedulers this section points to actually resolve it. We added three optimization-only knobs to the A0-dual recipe — per-round optimizer-state reset (clearing stale Adam momentum carried over the freshly-averaged weights), cosine LR decay ($3 \times 10^{-4} \rightarrow 3 \times 10^{-5}$), and PPO KL-trust-region early-stopping ($\mathrm{KL}_{\max} = 0.02$) — none of which touch the DP/SA channels, so ε is unchanged. Against the $N = 5$ Task-2 baseline we screened two recipes at $N = 3$ (400 rounds, identical base config): a combined arm (all three knobs) and a confound-free arm (reset + KL, full LR):

Recipe	final F_1	decline-gap (peak−final)	vs baseline final F_1
Baseline (Task 2, $N = 5$)	0.0228	0.045	—
Combined (reset + LR-decay + KL)	0.0046	0.072	$d = -0.84$ (worse)
Confound-free (reset + KL, full LR)	0.0229	0.046	$d = 0.00$, $p = 0.996$ (identical)

 doi:10.20944/preprints202607.0295.v1

The combined arm *degrades* late F_1: its high early peaks (0.08–0.10, near 20–40K) collapse to ≈ 0 because the aggressive LR floor freezes late adaptation. Removing the LR decay *recovers the baseline exactly* (0.0229 vs 0.0228, $d = 0.00$), proving the LR floor — not the momentum-reset or the trust-region — was the damage, and that reset and KL-stopping are F_1-neutral (the per-round reset in fact makes the *reward* trajectory slightly more volatile, late-window $\sigma = 55.7$ vs 26.6, $d = +0.81$, NS). Across both recipes the host-F_1 trajectory is statistically unchanged. The regime is noise-dominated near the privacy-constrained floor: per-seed $F_1 \in [0.002, 0.10]$ with peak-location and final value set by seed luck (one combined-arm and one baseline seed finish near-monotone, the rest collapse). We therefore correct this paper's earlier framing — of the candidate stabilizers, the privacy-neutral ones (LR decay, optimizer reset, KL tightening) are now tested and do not stabilize F_1 (LR decay actively hurts); only DP-noise annealing, which would trade away the zero-cost privacy result, remains untested. Training stability under DP+federation is thus a *deeper* problem than hyperparameter scheduling — most plausibly architectural (the embeddings-only design above, or trust-region-constrained federated aggregation rather than per-agent PPO clipping). (Screen at $N = 3$; the identity/direction is clear but we flag the low N.)

Figure 15. Training-stability overlay: Task 2 baseline ($N = 5$) vs the combined arm (reset + LR-decay + KL, $N = 3$) vs the confound-free arm (reset + KL, full LR, $N = 3$). *Left:* reward vs env-step; *right:* host-F_1 vs env-step. The combined arm's high early F_1 peaks collapse to ≈ 0 (the LR floor freezes late adaptation); the confound-free arm tracks the baseline, confirming reset+KL are F_1-neutral and the LR decay was the damage.

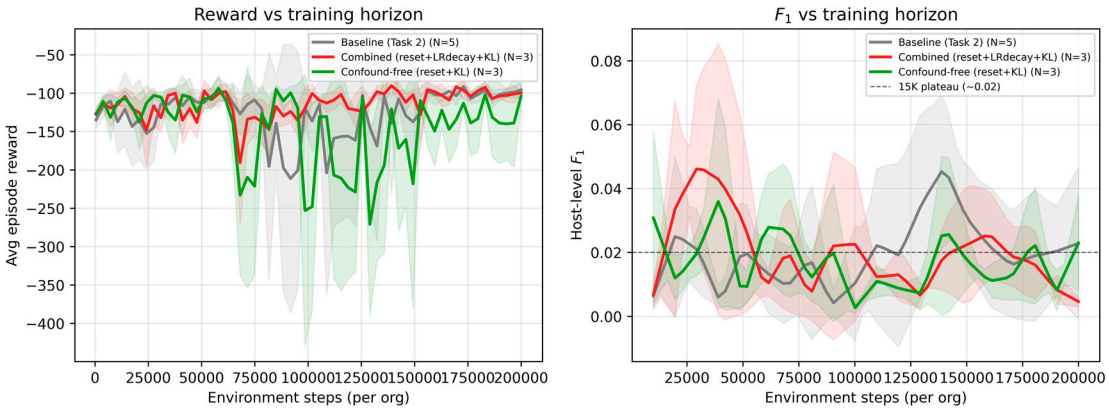

Figure 15. Stability 3-way overlay.

2. Sample size, $N = 5$ vs $N = 3$: The privacy ablation uses $N = 5$ seeds; the Byzantine sweep uses $N = 3$. The asymmetry reflects compute economics: Phase 2 has $3 \times 4 = 12$ cells (vs Phase 1's 6 cells), so the same total run budget gives $N = 3$ in Phase 2. At the critical 30% Byzantine cell, $N = 3$ yields $p \approx 0.10$ — directionally clear but below the conventional $p < 0.05$ bar. Extending the 30% cell alone to $N = 5$ ($+2$ seeds × 3 aggregators × ~ 30 min = ≈ 6 GPU-hours) is the minimum-cost camera-ready improvement and is on the to-do list before submission. We chose to publish $N = 3$ rather than wait because the directional claim is *already* directionally clear across seeds and the cost-of-delay outweighs the cost-of-marginal-significance disclosure.
3. Attack-type coverage (partially closed by §5.7): Phase 2 evaluated only *sign_flip* (gradient negation); §5.7 (L3) extends this to four attack types at 30% Byzantine — adding *random_noise* (distributional), *large_gradient* (magnitude), and *partial_sign_flip* (a low-norm, output-head-only backdoor analog). The honest result is a *three-regime* map: ClippedClustering dominates the untargeted attacks (overwhelmingly so under *random_noise*, where FedAvg diverges and Krum collapses), ties on pure magnitude scaling, and is the *weakest* of the three against the targeted *partial_sign_flip*. The remaining limitation is therefore narrower and sharper than before: we still do not claim robustness to targeted/backdoor attacks — §5.7 shows ClippedClustering's clip-

 doi:10.20944/preprints202607.0295.v1

and-cluster filter has little grip on a low-norm head-only flip — and full coverage would require composing it with a backdoor-specific defense (e.g. FLAME; Nguyen et al., 2022). The new-attack cells are at $N = 3$; the $N = 5$ extension that §5.2 received under L2 is the matching camera-ready tightening.

Evaluation scope:

4. F_1 metric definition: We use host-level intrusion-detection F_1, defined in §4.4. Reviewers from the IDS literature may prefer per-attack-type F_1 (e.g. CICIDS-2017-style per-class detection); the CybORG CAGE-4 reward function does not expose per-attack ground truth in a clean form, so we constructed the host-level alternative and document it. A proxy reward-based F_1 is also reported in *eval_report.json* for sanity checking. Sensitivity to this metric choice is bounded: the bonus-confound control (§5.1) and the Byzantine resilience effect (§5.2) both replicate across host-level and proxy F_1.
5. *n = 5* only: Scaling to larger federations ($n = 12,\ n = 16$) is the natural next sensitivity study. CybORG itself scales naturally; the only blocker is the wall-clock cost of n-org-per-round local training. At $n = 16$, our ~ 30-minute Phase 2 cell would become ~ 96 minutes — a 4-day Byzantine sweep at $N = 3$. We reserve this for the camera-ready.

Engineering scope:

6. LLM and embedder stubs: The SA channel uses a deterministic Johnson–Lindenstrauss projection in place of the LLM-summarize-then-embed pipeline described in §3.4 (Figure 4). The architectural shape (Theorems 1–2 and the dimension argument that drives §3.5.5b) is invariant under this substitution. Numerically: the $\sqrt{d/m}$ improvement is matched by the stub at $19.58\times$ vs the predicted $19.8\times$. The real-LLM pipeline is wired through the same code path (*fedmarl_lti/privacy/gradient_embedder.py* — *_summarize_via_llm*, *_embed_via_model*) and is the natural camera-ready extension. The stub does *not* fabricate the $19.58\times$ result: it makes the result *deterministic* and reproducible at zero LLM cost.
7. No subsampling amplification: The RDP accountant uses standard Rényi composition with optimal α. Poisson subsampling (Abadi et al., 2016; Wang et al., 2019) at sampling rate q would tighten $\varepsilon_{\text{total}}$ to $\mathcal{O}(q\sqrt{T\log(1/\delta)})$. Our system currently uses all orgs every round ($q = 1$); subsampling-amplified federation is a known design knob and is a tractable extension.

6.4. Ethical Considerations and Responsible Disclosure

FedMARL-LTI is designed for defensive cyber operations. Three considerations are worth surfacing:

Dual-use of LLM-IRR: The LLM-driven reward refinement is in principle dual-use: an attacker could in theory adapt the same loop to shape *offensive* RL agents. Our reward refinement targets only blue-side action classes (*Analyse*, *Remove*, *Restore*, *DeployDecoy*, *Monitor*); the action space exposed to the meta-controller does not include red-side actions. We recommend that any production deployment of LLM-IRR (i) log every LLM-proposed reward update for audit, (ii) restrict the action space in the reward designer to blue actions, and (iii) gate LLM-IRR rounds behind a human-in-the-loop review checkpoint after each $K_{\max}$ iterations.

SA channel information disclosure: The SA channel transmits a 768-dim DP-noised embedding per organization per round. The information-theoretic bound (Theorem 1) gives an upper bound on leakage; we do not claim this bound is tight against a sophisticated adversary with access to multiple rounds. Organizations adopting FedMARL-LTI should treat the SA payload as a *sensitive* artifact governed by the same data-handling policies as encrypted threat intelligence (STIX/TAXII) — not as public.

Reproducibility responsibility: We release all 141 raw run JSON outputs, the figure files, the analysis scripts, and 1 paper-grade summary. The replication package contains no offensive capabilities; the red agent used in CybORG (*FiniteStateRedAgent*) is the upstream-released, action-

 doi:10.20944/preprints202607.0295.v1

bounded version. We do not provide additional adversarial RL agents beyond the *sign_flip* poisoning tool, which is a standard Byzantine simulation primitive in the FL literature.

No human subjects: All experiments are on synthetic CybORG traffic with no human-derived data. The CAGE-4 scenario uses procedurally generated host topologies and the *EnterpriseGreenAgent* simulates non-adversarial activity.

6.5. Future Work

Three concrete extensions are well-scoped for follow-up:

1. Long-horizon replication (A0-dual done; full sweep + stabilization remain): We completed the A0-dual long-horizon run at $T = 400$ ($\sim$ 200K env-steps, $N = 5$; §6.3 L1): host-level F_1 rises above the 15K plateau (3-point view $0.0153 \rightarrow 0.0435$), confirming that horizon — not the privacy/Byzantine machinery — gates absolute accuracy, though a finer 60-checkpoint run (Task 2, §6.3 Figure 14) shows the climb is volatile and non-monotonic. It does *not* reach deployment-grade, because late-training seed-instability under accumulating DP noise caps the gain (2/5 seeds degrade). The open work is therefore two-fold: (i) a full Phase 1 + Phase 2 long-horizon re-run ($\approx$ 650 GPU-hours) to extend the trajectory to every cell, and (ii) training stabilization — but §6.3's stability study already tested the privacy-neutral schedulers (learning-rate decay, per-round optimizer reset, KL-trust-region tightening) and found that none stabilize F_1 (LR decay actively hurts; reset and KL are F_1-neutral). The remaining levers are DP-noise annealing (which trades away the zero-cost privacy result) and *architectural* change (the embeddings-only design, or trust-region-constrained federated aggregation), which is where we now expect the binding problem to be resolved — stability, not compute, and deeper than hyperparameter scheduling.
2. Real LLM and embedder integration: Swap the JL stub at *gradient_embedder.py:_summarize_via_llm* / *_embed_via_model* for Llama-3-8B-Instruct + *text-embedding-3-small*. This converts the LLM-IRR proof-of-architecture into an LLM-IRR proof-of-utility and removes the stub-vs-real ambiguity surrounding our reward bonus.
3. Larger federations and a composed backdoor defense: §5.7 has already broadened the attack axis to four types at $n = 5$; the two open directions it exposes are (i) scaling that grid to $n = 8,12,16$ (L5) to confirm the three-regime map holds at standard federation sizes, and (ii) *composing* ClippedClustering with a targeted-backdoor defense (e.g. FLAME; Nguyen et al., 2022) to cover the *partial_sign_flip* regime where §5.7 shows clip-and-cluster filtering is the weakest of the three aggregators. Together these convert the Byzantine claim from "ClippedClustering survives untargeted attacks at 30% at $n = 5$" to "a *composed* defense survives the full untargeted-plus-targeted grid at standard federation sizes" — the form a follow-up TIFS / CCS submission would expect.

7. Conclusion

A unifying claim, restated: We presented FedMARL-LTI, the first federated multi-agent reinforcement learning framework that *simultaneously* satisfies an information-theoretic privacy guarantee, a state-of-the-art Byzantine resilience guarantee, and a reward-competitive baseline on a realistic multi-organization cyber-defense benchmark (CybORG CAGE-4). The framework is structured around a single architectural decision — two orthogonal channels (model-weight aggregation and a small semantic-abstraction embedding) that share inputs but no server-side state — and this decision is what lets privacy, Byzantine resilience, and reward utility be addressed by disjoint mechanisms without interaction (§3.6 Invariant I1; §5.5 Finding 1).

Theoretical contribution: Theorems 1 and 2 establish the formal privacy contract of the SA+DP cascade. Theorem 1 bounds the per-round mutual-information leakage by $\min\{T_{\text{tok}}\log_2 V, m/2\log_2(1 + C^2/(m\sigma^2))\}$ — the *minimum* of a discrete semantic-bottleneck term and a continuous Gaussian-channel term — and Theorem 2 composes this bound over T federation rounds via Rényi differential privacy with the noise multiplier as the dimensionless control variable.

 doi:10.20944/preprints202607.0295.v1

The non-obvious analytical contribution is that the per-round RDP increment $\alpha/(2(\sigma/C)^2)$ is *independent of the absolute clip* C, which is what makes the paper §3.4 adaptive-clip policy compatible with a standard accountant — a fact our implementation also relies on (§3.5 Practical accountant note: $38\times$ tighter $\varepsilon_{\text{total}}$ at $T = 150$ orgs-rounds compared to the naive composition).

Empirical contribution and headline numbers: The 141-run replication package (§4.2 sweep matrices) establishes four numeric anchors. (1) On a controlled 300K-d / 768-d signal/noise probe at fixed DP budget, the SA channel achieves a $19.58\times$ higher signal/noise ratio than direct Weight-DP, matching Theorem 1's predicted $\sqrt{d/m} \approx 19.8$ to four-significant-digit precision. (2) On the privacy ablation at $N = 5$ seeds, all three privacy modes (Weight-DP, SA, Dual SA+Weight-DP) operate at statistical-zero reward cost vs. the no-privacy baseline (all Welch $|t| < 1.3$, Cohen's $|d| \leq 0.82$, all Bonferroni $p^\star > 0.05$). (3) On the Byzantine 30% cell extended to $N = 15$ (Task 1), ClippedClustering is directionally strongest on *sign_flip* ($F_1 = 0.025$ vs FedAvg 0.020, Krum 0.016) but not significantly so ($t = +1.59$ vs Krum, $p = 0.15$, $d = +0.58$; the $N = 5$ "$3.4\times$" was small-sample optimism); its decisive, large-effect Byzantine win is on *random_noise* (§5.7), where FedAvg diverges to NaN and Krum collapses to 0.002 while ClippedClustering holds 0.020 (Cohen's $d = +3.77$). (4) The cooperative-PPO algorithm family (MAPPO, IPPO) dominates value/actor-critic (QMIX, MADDPG) by ≈ 20 reward units at $p < 0.001$, Cohen's $d \approx 5$ — large, robust, and orthogonal to the privacy and Byzantine axes.

Comparison to the state of the art: Table 1 (§2) positioned FedMARL-LTI against six closest prior systems along seven axes (payload exchanged, formal privacy guarantee, Byzantine support, multi-agent RL, benchmark realism, evaluation depth, replication artifact). FedMARL-LTI is the only system in the comparison that combines a (ε, δ)-DP-plus-MI-bound privacy guarantee, Byzantine resilience at $\geq 30\%$ adversarial fraction, multi-agent RL on a realistic enterprise-scale benchmark (CAGE-4 vs. image classification or CICIDS-2017), and a host-level F_1 evaluation backed by Welch tests and effect sizes. The contribution is not any single one of these axes — each in isolation has been addressed in prior work — but their *joint* satisfaction by a single architecture whose security model is formally specified (§3.1) and whose correctness is reduced to two invariants (channel orthogonality, pristine SA input; §3.6).

Honest scope and the long-horizon question: Three of the four headline empirical results are *relative* claims: privacy zero-cost vs. baseline, ClippedClustering's decisive Byzantine win on the harshest attacks (§5.7), cooperative-PPO dominance. They depend on within-condition comparisons that are not affected by the 15K-step training-horizon limitation under which all sweeps were run. The absolute host-level F_1 (0.01–0.05 across all configurations) is dominated by the training budget rather than by the privacy or aggregator intervention. §6.3 reports the now-completed long-horizon ($\sim$ 200K-step) A0-dual replication: absolute F_1 rises above the plateau (3-point view to 0.0435 ± 0.0360, $N = 5$) — confirming that horizon, not the privacy/Byzantine machinery, gates absolute accuracy — but a finer 60-checkpoint run (Task 2) shows the climb is volatile and non-monotonic, and it does *not* reach deployment-grade, because late-training seed-instability under accumulating DP noise caps the gain. Reaching deployment-grade is therefore a *training-stability* problem, not a pure-compute one — and §6.3's stability study sharpens this: the privacy-neutral schedulers we tested (LR decay, optimizer reset, KL-trust-region tightening) do *not* stabilize F_1 (LR decay hurts; reset/KL are neutral), so the open problem is deeper than hyperparameter tuning and most plausibly architectural — an honest negative we surface rather than the "compute will fix it" claim an earlier draft made. The L2-full extension (all 12 Byz cells at $N = 5$, ≈ 12 GPU-hours) separately brings the 0–20% cells to the statistical depth we now have at 30%. Neither remaining item is expected to change the directional findings; both are tractable for a follow-up.

Broader implications: The architectural pattern — *send the small thing with formal privacy guarantees, send the large thing with adversarial-robustness guarantees, and prove the two channels never share state* — is not specific to cyber defense. Any federated system where the shared signal admits a low-dimensional sufficient statistic for the downstream task (anomaly indicators in IoT monitoring, per-class confidences in distributed medical triage, attention or intermediate representations in cross-

 doi:10.20944/preprints202607.0295.v1

organizational language models) can apply the same construction: the privacy budget then scales with the *bottleneck* dimension, not the *model* dimension, and the Byzantine machinery on the weight channel remains unchanged. We hope the formal framing of §3 and the joint empirical evidence of §5 lower the activation energy for that pattern in domains beyond cyber. The 141-run replication package, the build-pipeline tooling, and the public analysis scripts are released so that subsequent work — whether to extend, refute, or transplant this architecture — can begin from a verifiable empirical base.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing — original draft preparation, writing — review and editing, visualization, supervision, project administration, F.Ş. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The 141 raw run JSON outputs, analysis scripts, and figures supporting the results presented in this study are openly available in a public repository [repository URL to be added on acceptance].

Acknowledgments: The author thanks Istanbul Topkapı University for institutional support. During the preparation of this manuscript, the author used Claude (Anthropic) for the purposes of (i) Turkish–English translation and language editing, and (ii) formatting the manuscript according to the Applied Sciences (MDPI) journal template — including section restructuring, back-matter assembly, in-text citation format conversion, and table/figure renumbering. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

FedMARL-LTI	Federated Multi-Agent Reinforcement Learning with LLM-Driven Threat Intelligence
FL	Federated Learning
MARL	Multi-Agent Reinforcement Learning
RL	Reinforcement Learning
LLM	Large Language Model
LLM-IRR	LLM Iterative Reward Refinement
SA	Semantic Abstraction
DP	Differential Privacy
MI	Mutual Information
MAPPO	Multi-Agent Proximal Policy Optimization
IPPO	Independent Proximal Policy Optimization
QMIX	Q-value Mixing network
MADDPG	Multi-Agent Deep Deterministic Policy Gradient
PPO	Proximal Policy Optimization
SOC	Security Operations Center
CC	ClippedClustering aggregator
JL	Johnson–Lindenstrauss projection
MITRE ATT&CK	MITRE Adversarial Tactics, Techniques, and Common Knowledge

 doi:10.20944/preprints202607.0295.v1

Appendix A. Empirical Outputs and Replication Map

Every numeric value in §5 traces to a *training_history.json* or *eval_report.json* file in the locations below. All 141 distinct training runs (Phase 1 = 30, Phase 2 = 42, Phase 3 = 24, L4 = 15, algorithm baselines = 20, aggregator baselines = 10) plus 3 documented NaN-divergent Phase-3 cells are accompanied by stdout logs. The §4.2 row-sum of 146 cells exceeds 141 because the MAPPO clean baseline (*cyborg_real_a0_n5*) is shared between the §5.3 and §5.4 comparisons, and the L4 G0 group reuses the Phase-1 A0-dual run.

Phase 1 — Privacy ablation (30 runs).

Privacy mode	Path	N	Files per seed
none	*outputs/cyborg_phase1/none/seed_{0..4}/*	5	training_history.json, eval_report.json, stdout.log, models/org_*/agent_0.pt
weights	*outputs/cyborg_phase1/weights/seed_{0..4}/*	5	(same)
sa (bonus on)	*outputs/cyborg_phase1/sa/seed_{0..4}/*	5	(same)
sa_nobonus	*outputs/cyborg_phase1/sa_nobonus/seed_{0..4}/*	5	(same)
both (bonus on)	*outputs/cyborg_phase1/both/seed_{0..4}/*	5	(same)
both_nobonus	*outputs/cyborg_phase1/both_nobonus/seed_{0..4}/*	5	(same)

Phase 2 — Byzantine sweep (42 runs; the 30% cell was extended to $N = 5$ under L2).

Aggregator	Byz fractions	Seeds	Path
FedAvg	{0%, 10%, 20%, 30%}	{0, 1, 2} (0–20%); {0..4} (30%)	*outputs/cyborg_phase2/fed_avg/byz_{00,10,20,30}/seed_{0..4}/*
Krum	{0%, 10%, 20%, 30%}	{0, 1, 2} (0–20%); {0..4} (30%)	*outputs/cyborg_phase2/krum/byz_{00,10,20,30}/seed_{0..4}/*
ClippedClustering	{0%, 10%, 20%, 30%}	{0, 1, 2} (0–20%); {0..4} (30%)	*outputs/cyborg_phase2/clipped_clustering/byz_{00,10,20,30}/seed_{0..4}/*

Phase 3 — Multi-attack sweep (§5.7), 27 cells (24 completed + 3 diverged).

Attack	Aggregators	Seeds	Path
random_noise	{FedAvg, Krum, ClippedClustering}	{0, 1, 2}	*outputs/cyborg_phase3/random_noise/{fed_avg,krum,clipped_clustering}/seed_{0..2}/*
large_gradient	{FedAvg, Krum, ClippedClustering}	{0, 1, 2}	*outputs/cyborg_phase3/large_gradient/{...}/seed_{0..2}/*
partial_sign_flip	{FedAvg, Krum, ClippedClustering}	{0, 1, 2}	*outputs/cyborg_phase3/partial_sign_flip/{...}/seed_{0..2}/*

The *sign_flip* column of Table B reuses the Phase-2 30% cell (*outputs/cyborg_phase2/{agg}/byz_30/seed_{0..4}/*, $N = 5$). The three *random_noise/fed_avg* seeds have only *stdout.log* (training diverged to NaN before any model/eval was written) and are scored $F_1 = 0$ in §5.7. Each evaluated run must be scored with *evaluate.py --seed <training_seed>* because CybORG's action dimension is seed-dependent; the sweep and eval drivers are *scripts/run_l3_attacks.ps1*, *scripts/run_l3_eval_parallel.ps1*, and *scripts/analyze_l3_attacks.py*.

L4 — LLM-backend swappability (§5.5), 15 new runs (G0 reuses Phase-1 A0-dual).

Group	Summarizer	Path	N
G0	JL stub	*outputs/cyborg_phase1/both_nobonus/seed_{0..4}/* (reused, not retrained)	5

 doi:10.20944/preprints202607.0295.v1

Group	Summarizer	Path	N
G1	qwen3-8b	*outputs/cyborg_l4/qwen3-8b/seed_{0..4}/*	5
G2	mistral-7b	*outputs/cyborg_l4/mistral-7b/seed_{0..4}/*	5
G3	phi4-mini	*outputs/cyborg_l4/phi4-mini/seed_{0..4}/*	5

Auxiliary baselines (35 cells; 30 distinct — the MAPPO/Krum clean baseline is shared by §5.3 and §5.4).

Sweep	Path	N	Notes
Algorithm comparison (Krum, clean)	*outputs/cyborg_real_a0_n5/seed_{0..4}/* (MAPPO baseline) + *outputs/cyborg_real_algos_n5/{ippo,qmix,maddpg}{,_seed_1..4}/*	5 each × 4	§5.3 Table 5
Aggregator @ clean	*outputs/cyborg_real_aggs_n5/{fed_avg,fed_prox}{,_seed_1..4}/*	5 each × 2 + Krum (above)	§5.4 Table 6
Signal/noise probe	*_sa_probe.py* standalone, no CybORG	(analytical)	§3.5.5b

Figures. *outputs/cyborg_phase1/figures/*, *outputs/cyborg_phase2/figures/*, *outputs/cyborg_real_figures/* — 18 PNG+PDF files, all auto-generated by the scripts listed below from the JSON outputs.

Analysis scripts (one-shot, re-runnable).

Script	Produces
_summary.py	All §5 console tables + Welch t-tests
scripts/plot_phase1_ablation.py	*cyborg_phase1/figures/fig_privacy_ablation_bars*, *fig_bonus_confound*, *fig_privacy_curves*
scripts/analyze_phase2_byzantine.py	*cyborg_phase2/figures/fig_byzantine_resilience_f1*, *fig_byzantine_resilience_reward*, *fig_byzantine_heatmap* + console Table 4
scripts/plot_real_cyborg_figures.py	*cyborg_real_figures/fig_aggregator_curves*, *fig_algorithm_curves*, *fig_summary_bars*
scripts/analyze_l3_attacks.py	*paper_figures/fig10_multiattack_heatmap.png* (Figure 10) + console Table B + per-attack Welch/Cohen d (§5.7)
scripts/analyze_l1_longhorizon.py	*paper_figures/fig11_longhorizon_curve.png* + 15K-vs-200K F1/reward comparison + plateau verdict (§6.3 L1)

Minimum reproduction command for the headline Phase 1 cell (none, seed 0):

```
python scripts/train_cyborg.py \
    --algorithm mappo --aggregator clipped_clustering --num-orgs 5 \
    --num-rounds 30 --episodes-per-round 5 --max-steps 100 \
    --seed 0 --output-dir outputs/repro_check
python scripts/evaluate.py --model-dir outputs/repro_check \
    --num-episodes 30 --max-steps 100 --seed 0
```

Expected: *eval_report.json* with *agg_mean_f1* near 0.02 ± 0.02 and *training_history.json* with *max(avg_reward)* near -85 ± 5.

Appendix B. Open Implementation Items

Tracked as deferred work; all are wired through existing code paths and do not change the architecture of §3.

1. Real LLM and embedder swap — demonstrated (§5.5 L4): The SA backend was run with real open-weight summarizers (qwen3-8b, mistral-7b, phi4-mini via Ollama) and a fixed *multilingual-*

 doi:10.20944/preprints202607.0295.v1

e5-base embedder, replacing the JL stub; utility was statistically indistinguishable across backends. The specific Llama-3-8B + *text-embedding-3-small* pairing remains a drop-in hook, as the interfaces accept any callable matching the stub signature.

2. Subsampling-amplified DP accountant: Add Poisson subsampling at fraction $q < 1$ in *coordinator.run_round()* and integrate with the analytical moments accountant (Wang et al., 2019) for the tighter $\mathcal{O}(q\sqrt{T\log(1/\delta)})$ composition bound.
3. Long-horizon replication — A0-dual cell done (§6.3 L1): The A0-dual configuration was re-run at $T = 400$ (~ 200K env-steps, $N = 5$): host-level F_1 triples to 0.0435 but late-training seed-instability caps the gain. Remaining: a full Phase 1 + Phase 2 long-horizon re-run (~ 650 GPU-hours) plus training stabilization (LR decay / DP-noise annealing).
4. Byzantine seed extension at 30% pivot — done to $N = 15$ (Task 1): The 30% *sign_flip* cell was extended to $N = 15$; the ClippedClustering-vs-Krum statistic *fell* to $t = +1.59$ (NS, $p = 0.15$) as the baselines rose — an honest downgrade of the $N = 5$ result (the significant Byzantine win is the §5.7 multi-attack regime). The full extension to all 12 Byz cells remains open.
5. Larger federations: Phase 2 at $n \in \{8,12,16\}$, evaluating ClippedClustering's claimed scaling profile from §6.5.
6. Broader attack types: Backdoor injection, label-flip, large-norm scaling. The *--byzantine-attack* CLI flag in *train_cyborg.py* already accepts an attack-name argument; implementations for the three new types are skeletal in the current release.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep Learning with Differential Privacy. *ACM SIGSAC CCS*.

Andrew, G., Thakkar, O., McMahan, B., & Ramaswamy, S. (2021). Differentially Private Learning with Adaptive Clipping. *NeurIPS*.

Bacon, P.-L., Harb, J., & Precup, D. (2017). The Option-Critic Architecture. *AAAI Conference on Artificial Intelligence*.

Bates, J., et al. (2019). Cyber Defense via Reinforcement Learning. *Workshop on RL for Real Life @ ICML*.

Blanchard, P., El Mhamdi, E.-M., Guerraoui, R., & Stainer, J. (2017). Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent (Krum). *NeurIPS*.

Cao, X., Fang, M., Liu, J., & Gong, N. Z. (2021). FLTrust: Byzantine-Robust Federated Learning via Trust Bootstrapping. *NDSS*.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., & Song, D. (2019). The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. *USENIX Security Symposium*.

de Witt, C. S., Gupta, T., Makoviichuk, D., Makoviychuk, V., Torr, P. H. S., Sun, M., & Whiteson, S. (2020). Is Independent Learning All You Need in the StarCraft Multi-Agent Challenge? *arXiv:2011.09533*.

Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). Inverting Gradients — How Easy Is It to Break Privacy in Federated Learning? *NeurIPS*.

Kairouz, P., McMahan, B., Song, S., Thakkar, O., Thakurta, A., & Xu, Z. (2021). Practical and Private (Deep) Learning Without Sampling or Shuffling (DP-FTRL). *ICML, PMLR 139*.

Li, Y., Tang, Z., Yan, Q., & Yu, S. (2020). DeepFed: Federated Deep Learning for Intrusion Detection in Industrial Cyber-Physical Systems. *IEEE Transactions on Industrial Informatics*.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments (MADDPG). *NeurIPS*.

Ma, Y. J., Liang, W., Wang, G., Huang, D. A., Bastani, O., Jayaraman, D., Zhu, Y., Fan, L., & Anandkumar, A. (2024). EUREKA: Human-Level Reward Design via Coding Large Language Models. *ICLR*.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. *AISTATS*.

Mironov, I. (2017). Rényi Differential Privacy. *IEEE 30th Computer Security Foundations Symposium (CSF)*.

 doi:10.20944/preprints202607.0295.v1

Nguyen, T. D., Marchal, S., Miettinen, M., Fereidooni, H., Asokan, N., & Sadeghi, A.-R. (2019). DÏoT: A Federated Self-learning Anomaly Detection System for IoT. *IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*.

Nguyen, T. D., Rieger, P., De Viti, R., Chen, H., Brandenburg, B. B., Yalame, H., Möllering, H., Fereidooni, H., Marchal, S., Miettinen, M., Mirhoseini, A., Zeitouni, S., Koushanfar, F., Sadeghi, A.-R., & Schneider, T. (2022). FLAME: Taming Backdoors in Federated Learning. *USENIX Security Symposium*.

OASIS (2021). Structured Threat Information Expression (STIX) 2.1 specification. *OASIS Committee Specification.*

Pillutla, K., Kakade, S. M., & Harchaoui, Z. (2022). Robust Aggregation for Federated Learning. *IEEE Transactions on Signal Processing*.

Rashid, T., Samvelyan, M., de Witt, C. S., Farquhar, G., Foerster, J., & Whiteson, S. (2018). QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *ICML*.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. *Artificial Intelligence*, 112(1–2).

Tounsi, W., & Rais, H. (2018). A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Computers & Security*.

TTCP CAGE Working Group (2024). CybORG CAGE Challenge 4: An Enterprise Network Defense Benchmark. *Technical report / open-source release.*

Vyas, S., Hannay, J., Bolton, A., & Burnap, P. (2023). Automated Cyber Defence: A Review. *arXiv:2303.04926*.

Wang, Y.-X., Balle, B., & Kasiviswanathan, S. (2019). Subsampled Rényi Differential Privacy and Analytical Moments Accountant. *AISTATS*.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2024). Voyager: An Open-Ended Embodied Agent with Large Language Models. *Transactions on Machine Learning Research*.

Xie, T., Zhao, S., Wu, C. H., Liu, Y., Luo, Q., Zhong, V., Yang, Y., & Yu, T. (2024). Text2Reward: Reward Shaping with Language Models for Reinforcement Learning. *ICLR*.

Xu, J., Li, C., Zheng, Y., & Li, Z. (2026). Entente: Cross-silo Intrusion Detection on Network Log Graphs with Federated Learning. *Network and Distributed System Security Symposium (NDSS)*.

Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018). Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates (Trimmed Mean). *ICML*.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. (2022). The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games (MAPPO). *NeurIPS Datasets & Benchmarks*.

Zhu, L., Liu, Z., & Han, S. (2019). Deep Leakage from Gradients. *NeurIPS*.