Intelligent Automation & Soft Computing

Tech Science Press

https://doi.org/10.32604/iasc.2026.088003

ARTICLE

Hierarchical Trust Management for C-V2X Networks: A Game-Theoretic Edge-Cloud Architecture with LLM-Driven Calibration

Anil Carie[1,*], Sanjeev Kumar Makala[1], Satish Anamalamudi[1], Shaik Teena[1], Awadhesh Dixit[1], Pandu Sowkuntla[1] and Bhaskar Marapelli[2]

[1]Department of Computer Science and Engineering, SRM University AP, Neerukonda, Managalagiri, India
[2]Department of Computer Science and Information Technology, KLEF (Deemed to be University), Vijayawada, India

*Corresponding Author: Anil Carie. Email: anilcarie.c@srmap.edu.in

Received: 26 June 2026; Accepted: 31 August 2026; Published: 21 September 2026

ABSTRACT: Connected Vehicle-to-Everything (C-V2X) networks rely on Basic Safety Messages (BSMs) for cooperative awareness, but authenticated vehicles can still transmit falsified kinematic data—an insider attack that cryptographic authentication cannot prevent. Existing misbehavior detection systems (MDSs) achieve high detection rates on static benchmarks yet provide no formal guarantee that honest behaviour is a rational vehicle's dominant strategy. We present a four-layer hierarchical trust architecture for 5G New Radio (NR) C-V2X that integrates edge AI detection, cloud large language model (LLM)-driven weight calibration, game-theoretic conviction, and ledger-anchored payoff tracking. A Nash equilibrium gate, calibrated with a wide margin above the empirically observed honest-vehicle suspicion ceiling, achieves zero observed false positives across 1584 honest phase-3 instances (one-sided 95% Clopper–Pearson upper bound: 0.19% per instance), without requiring any assumption about the honest-score distribution's shape. A clawback mechanism makes the expected payoff of attacking strictly negative for all tested temptation levels. Evaluation on NS3 with 3GPP Rel-18 channel models across nine attack types—including rational adversaries with heterogeneous temptation $T_i \sim U(0.5, 2.5)$—yields a detection rate (DR) of 86.3% with false-positive rate (FPR) of 0% observed across 1584 honest phase-3 instances. This zero-FPR result holds for the evaluated single-UAV configuration at $N \leq 50$ vehicles per zone; at $N = 100$, mean FPR rises to 4.9% (range 0%–13.5% across seeds) due to coverage-boundary effects, and multi-UAV sectorisation to extend this range remains unvalidated futurework. All 95 rational attackers ($N = 30$, seeds 1, 2, 3, 5, 6) chose cooperation, producing a honesty premium of $\Pi = +39.2$ Trust Credits (positive in every seed, range +16.4 to +55.4 TC). A five-arm ablation isolates the Nash gate's and clawback's independent contributions to this outcome. Investigating cloud LLM calibration's contribution to detection rate, we identified and corrected a learning-rate implementation fault that had suppressed nearly all of the LLM's recommended adjustments throughout the original evaluation; under the corrected configuration, calibration produces a statistically significant detection-rate improvement, confirmed with a live cloud-unreachable control (+5.3 percentage points, $p = 0.034$), reported alongside the original finding for transparency. To our knowledge, this is the first V2X misbehavior detection system to provide formal incentive compatibility under heterogeneous temptation for economically-rational attackers under the specified payoff model; this claim does not extend to attackers pursuing non-economic objectives, collusion, Sybil behaviour, or calibration-pipeline poisoning, which remain open problems.

KEYWORDS: C-V2X; trust management; misbehavior detection; game theory; edge-cloud architecture; LLM-driven calibration

1 Introduction

The deployment of Cellular Vehicle-to-Everything (C-V2X) communication is accelerating worldwide, with 3GPP Release 18 [1] standardising direct sidelink and network-assisted modes for safety-critical applications. Cooperative awareness—vehicles intermittently transmitting Basic Safety Messages (BSMs) with position, speed, heading, and timestamp—requires integrity that is a deployment-gating issue: a single fake BSM could cause phantom-braking, ghost vehicle injection, or platoon destabilisation [2]. Cryptographic authentication (European Telecommunications Standards Institute (ETSI) TS 102 940, Public Key Infrastructure (PKI)) guarantees only that authenticated vehicles can send messages; it does not guarantee the truthfulness of their content. A fraudulent insider with valid credentials can send forged kinematic data, an attack PKI cannot prevent by design [3]. ETSI TR 103 460 [4] describes a Misbehavior Detection System (MDS) pipeline of plausibility checks, misbehavior reporting (TS 103 759 [5]), and certificate revocation, but provides no formal economic incentive for honest participation. Implementations of machine learning (ML)-based MDS have shown 92%–97% detection rates on the VeReMi benchmark [6], evaluated on datasets such as the VeReMi Extension [7]. These rates are achieved on static, pre-programmed attack patterns; against adaptive opponents, reported FPRs of 2%–8% are operationally prohibitive at highway density (~40 false bans per hour per lane).

1.1 Limitations of Existing Approaches

First, no existing system provides formal deterrence. All current MDSs are detection-only: they flag misbehavior but offer no mechanism design argument [8] showing that honest behaviour is a rational vehicle's dominant strategy.

Second, evaluation benchmarks exclude adaptive attackers. VeReMi and its extensions contain only static attacker profiles. No published system has been evaluated against game-theoretic adversaries that adjust attack intensity based on accumulated suspicion.

Third, false positive rates remain operationally prohibitive. Reported FPRs of 2%–8% translate to ~40 false bans per hour per lane at highway density, undermining cooperative trust. By our empirical result (Section 3.3), zero exceedances were observed across 1584 honest phase-3 instances, giving a one-sided 95% Clopper–Pearson upper bound of 0.19% per instance, without requiring any distributional assumption.

1.2 Contributions

(1) A four-layer hierarchical trust architecture integrating edge AI detection, cloud LLM-driven weight calibration, game-theoretic conviction, and ledger-anchored payoff tracking over 5G NR C-V2X (Section 3).
(2) Nash equilibrium gate with an empirically validated, assumption-free false-positive bound: zero observed false positives across 1584 honest phase-3 instances (one-sided 95% Clopper–Pearson upper bound: 0.19% per instance), calibrated with a wide margin above the empirical honest-suspicion ceiling (Section 3.3).
(3) Orthogonal clawback mechanism reducing honesty premium by 21.3 TC when removed, without affecting detection rate (Section 5).
(4) Incentive compatibility demonstrated with 95 rational attackers ($N = 30$, $T_i \sim U(0.5, 2.5)$, seeds 1, 2, 3, 5, 6): all chose cooperation, yielding $\Pi = +39.2$ TC, positive in every seed (Section 5).
(5) Five-arm ablation study on NS3 with 3GPP Rel-18 channels (Section 5).

1.3 Paper Organization

Section 2 reviews related work. Section 3 presents the system and threat model. Section 4 develops the game-theoretic analysis. Section 5 reports experimental results. Section 6 discusses results and limitations. Section 7 outlines future work.

2 Related Work

2.1 ETSI Misbehavior Detection Framework

The European Telecommunications Standards Institute describes the reference architecture of Cooperative Intelligent Transport Systems (C-ITS) misbehavior detection in TR 103 460 [4]. This reference architecture divides the misbehavior detection process into three steps: plausibility checking, misbehavior reporting (TS 103 759 [5]), and certificate revocation. Bißmeyer [9] combined data-centric checks with attacker identification, concluding that detection alone is insufficient without revocation. Amanullah et al. [10] provide a broader taxonomy and analysis of misbehaviour detection systems across the C-ITS literature, situating plausibility-check and ML-based approaches within a common classification framework. The ETSI framework specifies no economic incentive for honest participation.

2.2 Plausibility-Based and ML-Based Detection

Plausibility checks. So et al. [11] proposed received signal strength indicator (RSSI)-based checks achieving DR = 83.7% on VeReMi [12], under line-of-sight (LoS) assumptions that fail in non-line-of-sight (NLoS) environments. VeReMi [12] and its extension [7] capture only static profiles; no benchmark includes adaptive or game-theoretic attacker profiles.

ML classifiers. Boualouache and Engel [6] survey 40+ ML-based MDSs (92%–97% accuracy, 3%–8% FPR on VeReMi). Kristianto et al. [13] propose semi-supervised federated learning (F_1 = 0.96). Almalki and Sheldon [14] combine Kalman filtering and Hampel outliers (DR ≈90%, FPR ≈ 8%). Fatih Yuce et al. [15] highlight the lack of standards for misbehavior detection beyond BSMs. At highway density, FPR = 2% means ~40 false bans per hour. By our empirical result (Section 3.3), zero exceedances of the operating threshold were observed across 1584 honest phase-3 instances, giving a one-sided 95% Clopper–Pearson upper bound of 0.19% per instance.

LLM-assisted detection. Yoshizawa et al. [16] survey the broader V2X security and privacy landscape, situating misbehavior detection alongside authentication and privacy threats. Friha et al. [17] survey LLM-based edge intelligence broadly, noting that per-packet inference latency with contemporary LLMs is generally incompatible with 100 ms BSM periodicity. Liu and Zhao [18] similarly identify the computational and latency constraints of integrating LLMs directly into 6G vehicular networks. Our split design resolves this: edge ensemble provides real-time per-packet scoring (21 ms round-trip time (RTT)); cloud LLM operates asynchronously on calibration batches (2500 ms RTT).

2.3 Trust Management and Reputation Systems

Ahmad et al. [19] proposed NOTRINO combining event-based and roadside unit (RSU)-aggregated reputation, but not the strategic aspect: a rational vehicle can accumulate high entity trust before a late-flip attack, exactly the ATK_RATIONAL profile we evaluate. Chen et al. [20] proposed a deep reinforcement learning (DRL)-based blockchain trust architecture (DR = 93.8%) but did not test whether punishment made honesty a dominant strategy. Han et al. [21] propose decentralized trust management with incentive mechanisms for vehicular ad-hoc network (VANET) information sharing, and Zhao et al. [22] propose a blockchain-based trust management model for VANETs; neither evaluates incentive compatibility against a

rational, temptation-parameterized adversary population. Lu et al. [2] and Noor-A-Rahim et al. [23] highlight that PKI addresses authentication but not behavioral trust.

2.4 Game Theory in Vehicular Security

Game theory has been applied to vehicular networks primarily in spectrum allocation, task offloading, and routing [24]; Sun et al. [25] survey this broader landscape of game-theoretic applications in vehicular networks. Mehdi et al. [26] analysed vehicle interaction as a two-player game but validated only analytically. Chouikhi et al. [27] combined reputation and game-theoretic incentives at the routing layer but not for BSM plausibility. AlSaqabi and Krishnamachari [28] propose a game-theoretic architecture incentivizing private data sharing in vehicular networks, addressing a complementary incentive problem to the misbehavior-deterrence one studied here. No existing V2X MDS provides a mechanism design argument [8] showing incentive compatibility under heterogeneous temptation.

2.5 Summary of Research Gaps

Table 1 summarizes evaluation conditions and reported performance across the approaches discussed above.

Table 1: V2X misbehavior detection: evaluation conditions and reported performance. DR/FPR values are not directly comparable across rows due to differences in dataset, channel model, and attacker sophistication.

Approach	DR (%)	FPR (%)	Deterrence	Adaptive Atk.	Platform	Ref.
RSSI Plausibility Checks	83.7	~4.1	No	No	VeReMi (802.11p)	[11]
ML Ensemble (RF/SVM)	92–97	3–8	No	No	VeReMi (802.11p)	[6]
SS-FL MDS	95	~5	No	No	VeReMi Extension	[13]
HCA-MDS (Kalman + Hampel)	~90	~8	No	No	VeReMi Extension	[14]
Blockchain + DRL Trust	93.8	N/R	Partial	No	Custom simulation	[20]
Edge-only static (this work)	86.3	0.0	No	Yes (9 types)	NS3 5G NR	Arm D
Proposed (this work)	85.3	0.0	Yes	Yes (9 types)	NS3 5G NR	Arm A

Note: DR for the final two rows is not statistically distinguishable (Section 5.7); the differentiating contribution of the proposed system is the Deterrence column (Nash gate and clawback), not detection rate. The Deterrence and Adaptive Atk. columns are directly comparable across rows. Prior-work DR/FPR reflect static attackers on pre-recorded 802.11p traces. Our DR/FPR reflect adaptive attackers on live NS3 5G NR with 3GPP Rel-18 channels.

3 System Model and Architecture

3.1 Network Model

We consider N vehicles communicating via 5G NR through an unmanned aerial vehicle (UAV)-gNB relay at 100 m above ground level (AGL). The UAV covers a single zone of length $L = v_n \cdot t_s = 1800$ m ($v_n = 20$ m/s, $t_s = 90$ s). Channel models follow 3GPP Rel-18: UMi_StreetCanyon (access) and UMa_LoS (backhaul). Edge RTT: 21 ms; cloud RTT: 2500 ms. Fig. 1 illustrates the four-layer architecture and the resulting data flow.

3.2 Threat Model

An *internal attacker* is an authenticated vehicle transmitting falsified BSMs. Nine profiles are defined (Table 2), from strong signals (ATK_COMPOSITE, $p_{\text{edge}} = 0.663$) to near-threshold evasion (ATK_STEALTHY, $p_{\text{edge}} \approx 0.22$) and game-theoretic rational attackers (ATK_RATIONAL, expected-utility

(EU)-driven with $T_i \sim U(0.5, 2.5)$, $\sigma_{\text{obs}} = 0.08$). Position-aware late-flip attackers cooperate until $x > L - 200$ m.

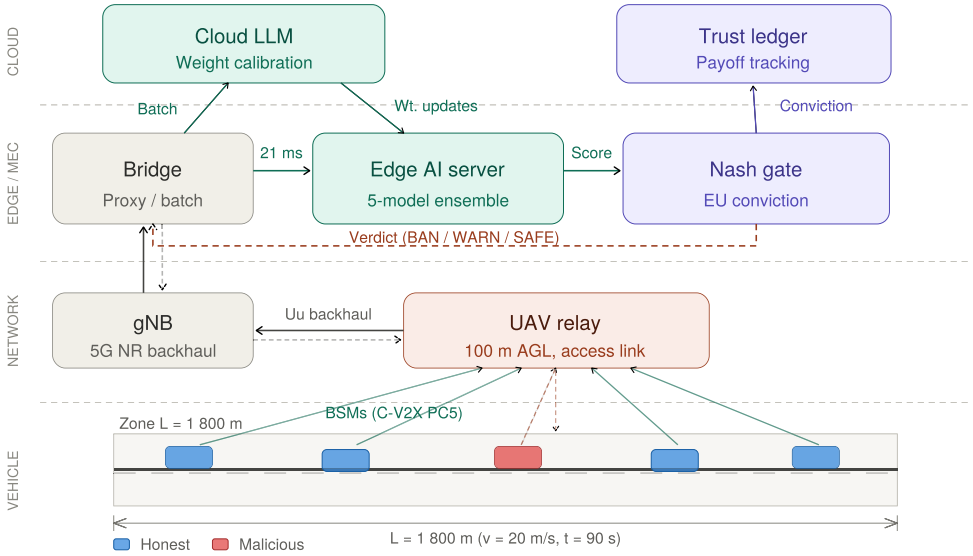

Figure 1: Four-layer hierarchical trust architecture. Data flows from vehicles through the UAV relay to the bridge, which forwards telemetry to the edge AI server and accumulates calibration batches for the cloud LLM. Only LLM-validated calibration results are applied. Edge RTT: 21 ms; Cloud RTT: 2500 ms.

Table 2: Attack profiles.

Code	Name	ρ_{edge}	Strategy
0	ATK_COMPOSITE	0.663	All features
1	ATK_TS_ONLY	~0.45	Timestamp replay
2	ATK_GPS_ONLY	~0.60	GPS position freeze
3	ATK_SPEED_ONLY	~0.35	Speed falsification
4	ATK_STEALTHY	~0.22	Near-threshold
5	ATK_ADAPTIVE	Variable	1-in-3 malicious
6	ATK_RATIONAL	~0.18	EU-driven
7	ATK_INTERMITTENT	Variable	Suspicion-adaptive
8	ATK_OPTIMAL	Variable	Full evasion

Out-of-Scope Threats

Three threat dimensions are explicitly outside the scope of this evaluation. *Sybil identity attacks* (a vehicle registering multiple pseudonymous identities to evade conviction) are delegated to the PKI/pseudonym layer beneath our trust layer (ETSI TS 102 941); we do not evaluate resistance to Sybil behavior independently. *Collusion* among multiple attacking vehicles is not modeled; our game-theoretic analysis (Section 4) treats each vehicle's strategy independently. *Calibration-pipeline poisoning* (an attacker crafting telemetry specifically to bias the LLM calibration step) is not evaluated experimentally, though we note a partial, verified structural bound: per-round weight deltas are clamped to $[0.000, 0.040]$ (Appendix B), and any calibration round that fails validation produces no weight change at all rather than a degraded fallback (Section 5.7), bounding a single poisoned batch's maximum influence to +0.04 on any one weight regardless of the model's output. All

three remain open problems for this system class generally and are named explicitly in Section 7 rather than claimed as resolved.

3.3 Edge AI Detection (Layer 2)

The edge runs a five-model ensemble (Algorithm 1):

$$M_1 = 0.1 + 0.6 \tanh\left(\frac{v - v_{\text{thr}}}{s}\right), \; v > v_{\text{thr}} \tag{1}$$

$$M_2 = 0.4 + 0.6 \tanh\left(\frac{|\delta + 0.3|}{0.3}\right), \; \delta < -0.3 \tag{2}$$

where $\delta = \tau - t_{\text{sim}}$. $M_3 = 0.55$ for frozen heading (variance < 0.01, $W = 8$). $M_4 = 0.80$ for GPS freeze. M_5 exploits light replay, speed surplus, and low-motion cues. The combination score:

$$p_{\text{edge}} = \left(\sum_{i=1}^{5} W_i M_i\right) \cdot b_{\text{shadow}} - d_{\text{rep}} \tag{3}$$

with $d_{\text{rep}} = \min(0.10, n_{\text{safe}} \cdot 0.005)$.

Algorithm 1: Edge AI ensemble scoring

Require: BSM packet $(v, \tau, \theta, p_x, p_y, \ell)$, CID, t_{sim}
$m_1 \leftarrow$ SPEEDANOMALY (v)
$m_2 \leftarrow$ TIMESTAMPREPLAY (τ, t_{sim})
$m_3 \leftarrow$ HEADINGBEHAVIOR $(\theta, \text{history})$
$m_4 \leftarrow$ POSITIONDRIFT (p_x, v, t_{sim})
$m_5 \leftarrow$ TEMPORALPATTERN $(v, \tau, p_x, \text{history})$
Apply cloud biases: m_2 += b_{ts}, m_4 += b_{px}
$p_{\text{edge}} \leftarrow \sum_{i=1}^{5} W_i \cdot m_i \cdot b_{\text{shadow}} - d_{\text{rep}}$
if $p_{\text{edge}} \geq \theta_{\text{mal}}$ **then**
    verdict ← BAN
else if $p_{\text{edge}} \geq \theta_{\text{safe}}$ **then**
    verdict ← WARN
else
    verdict ← SAFE
end if
return verdict, p_{edge}, $(m_1, \ldots, m_5)$

Empirical Result 1 (Assumption-Free FPR Bound)

Fig. 2 shows the resulting distribution. Each simulation run passes every honest vehicle through three sequential motion phases as it traverses the zone, each occupying roughly one third of total simulation time: an initial baseline-speed phase (phase 1, 20 m/s), a higher-speed phase (phase 2, 25 m/s), and a lower-speed phase (phase 3, 15 m/s). A *phase-3 instance* is one p_{edge} score computed for one honest vehicle at one BSM transmission during phase 3; with multiple honest vehicles per run and multiple runs across the baseline protocol, this yields 1584 such instances in total. We collected these 1584 honest-vehicle p_{edge} instances during phase 3 under the baseline protocol ($N = 6$, ATK_COMPOSITE and ATK_STEALTHY, attack rates 30/50/70%, seeds 1, 2, 3, 5, 6), reported in full in Section 5. Zero instances exceeded $\theta_{\text{mal}} = 0.30$.

By the Clopper–Pearson exact method [29], the one-sided 95% upper confidence bound on the per-instance false-positive rate is

$$\text{FPR} \leq 1 - 0.05^{1/1584} \approx 0.189\%. \tag{4}$$

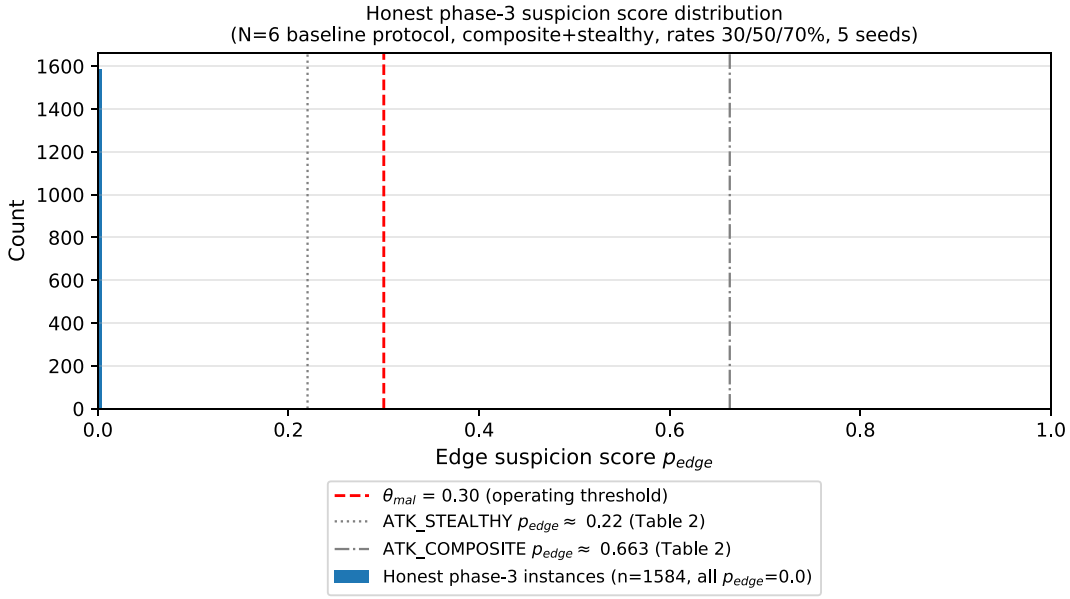

Figure 2: Edge suspicion score (p_{edge}) distribution across 1584 honest phase-3 instances (N = 6 baseline protocol, composite and stealthy attacks, rates 30/50/70%, seeds 1, 2, 3, 5, 6). Honest vehicles cluster at or near zero throughout the observed zone, including phase 3. Representative attack-profile suspicion scores (Table 2) are shown for reference; θ_{mal} = 0.30 retains a wide margin above the empirical honest maximum.

We report this explicitly as a statistical upper confidence bound under the stated simulation protocol, not as a universal guarantee: it certifies that, were the true per-instance false-positive rate above 0.189%, observing zero exceedances in 1584 independent phase-3 instances would occur with probability at most 5%; it does not certify performance outside this protocol, this attack mix, or this operating range ($N \leq 50$, see Section 6). This bound requires no assumption about the shape of the honest-score distribution. We note that honest-vehicle suspicion scores under the current detector calibration cluster at or near zero throughout the observed zone, including phase 3—a consequence of cumulative false-positive-reduction tuning applied across the M1–M5 ensemble over the course of development. The operating threshold θ_{mal} = 0.30 therefore retains a substantial empirical margin above the observed honest ceiling, consistent with the zero-exceedance result reported here.

3.4 Bridge Layer

The bridge relays telemetry to the edge and collects calibration batches via seven triggers. A Fisher separability gate:

$$\mathcal{F}_k = \frac{|\mu_k^W - \mu_k^S|}{\sqrt{(\sigma_k^{W2} + \sigma_k^{S2})/2 + \varepsilon}} \tag{5}$$

triggers batches when $\max_k \mathcal{F}_k \geq 1.5$.

3.5 Cloud LLM Calibration (Layer 3)

An 8B-parameter locally-hosted LLM balances calibration batches and produces weight deltas (Algorithm 2). Only LLM-validated calibrations are applied; no rule-based fallback.

Algorithm 2: Cloud LLM calibration

Require: Batch $\mathcal{B}$, separability $\mathcal{F}$

Partition: $\mathcal{B}_W, \mathcal{B}_S$
if $|\mathcal{B}_W| = 0$ or $|\mathcal{B}_S| = 0$ **then**
  return CALIBRATION_ERROR
end if
result $\leftarrow$ LLMQuery(prompt)
if result is valid JSON **and** passes validation **then**
  for $i = 1 \ldots 5$ **do**
    $\Delta W_i \leftarrow \text{clamp}(\text{result}.W_i\_\text{delta}, 0, 0.04)$
  end for
  return CALIBRATION_RESULT$(\Delta W_1, \ldots, \Delta W_5)$
else
  return CALIBRATION_ERROR
end if

Edge applies: $W_i \leftarrow \max(W_{\min}, W_i + \eta \Delta W_i)$, renormalized, with $\eta = 0.05$, $W_{\min} = 0.05$.

3.6 Nash Gate Conviction (Layer 1)

Fig. 3 illustrates the resulting decision boundary for the two evaluated attack profiles.

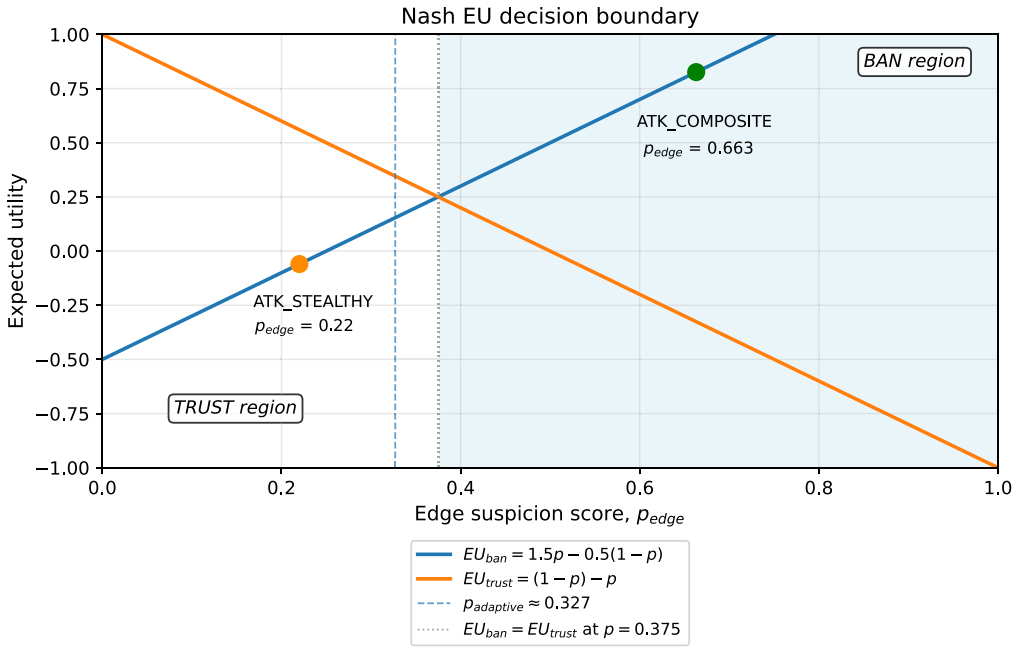

Figure 3: Nash EU threshold decision. ATK_COMPOSITE ($p_{\text{edge}} = 0.663$) lies firmly in the BAN region. ATK_STEALTHY ($p_{\text{edge}} = 0.22$) lies below $p_{\text{adapt}} \approx 0.327$, preventing conviction at the cost of 3.1 pp DR reduction while preserving FPR = 0%.

3.7 Trust Ledger Design

Conviction (Algorithm 3) amends an append-only ledger that persists per-vehicle history across zone transitions. We specify its architecture, schema, and use precisely, correcting terminology in earlier drafts of this work that described the mechanism as a blockchain.

Algorithm 3: Nash gate conviction decision

Require: p_{comb}, $\hat{\rho}$, window evidence, safe wins

$p_{\text{adapt}} \leftarrow 1/\left((1+\hat{\rho})\cdot 1.5 + 1.0\right)$
$EU_{\text{ban}} \leftarrow p_{\text{comb}} \cdot R_b - (1 - p_{\text{comb}}) \cdot C_f$
$EU_{\text{trust}} \leftarrow (1 - p_{\text{comb}}) \cdot R_t - p_{\text{comb}} \cdot C_c$
Gate 1: $p_{\text{comb}} > p_{\text{adapt}}$ **and** $EU_{\text{ban}} > EU_{\text{trust}}$
Gate 2: BAN hits in window$(4) \geq 1$
Gate 3: safe wins ≥ 0
if Gate 1 **and** Gate 2 **and** Gate 3 **then**
    CONVICT: impose clawback, amend ledger
end if

3.7.1 Architecture

The ledger is a local, append-only structured log maintained by the edge process, with a schema styled after blockchain conventions (sequential block numbers, hash and previous-hash fields). We verified directly against the released implementation that the `hash` and `prev` fields are populated with a placeholder value in the current codebase rather than computed cryptographic hashes, and that no hash-chain verification is performed. We therefore refer to this component as an *append-only ledger* rather than a blockchain throughout this paper: the schema anticipates a future cryptographically-chained implementation, but the tamper-evidence and non-repudiation properties that term implies are not yet present.

3.7.2 Stored Information

Each ledger entry records $\{blk, t, cid, evt, scr, hash, prev, note\}$, where *blk* is a sequence number, t the simulation timestamp, *cid* the vehicle identifier, *evt* the event type, and *scr* the trust score at the time of the event.

3.7.3 Trust Credit Calculation

The cloud calibration process reads the ledger (read-only) to compute each vehicle's prior safe-win count, which determines the Trust Credit discount applied to the false-ban cost penalty C_f at zone entry:

$$\text{discount} = \min(CREDIT\_DISCOUNT\_MAX,\ priorSafeWins \times CREDIT\_DISCOUNT\_PER\_WIN), \quad (6)$$

with $CREDIT\_DISCOUNT\_MAX = 0.30$ (a maximum 30% reduction in C_f) and $CREDIT\_DISCOUNT\_PER\_WIN = 0.015$ (1.5% discount per prior safe-win), values verified directly against the released source.

3.7.4 Enforcement and Practical Value

The discount mechanism directly reduces the effective penalty rate for vehicles with strong honest history, implementing the honesty premium (Section 4.3). Conviction (Algorithm 3) triggers the clawback mechanism independent of the ledger's cryptographic properties, since clawback depends only on payoff accounting, not on tamper-evidence.

3.7.5 Limitations

A field deployment would require completing the hash-chaining, or adopting a genuine distributed ledger platform, before the tamper-evidence and non-repudiation properties implied by blockchain terminology could be honestly claimed. The persistence layer evaluated here is sufficient for the incentive-compatibility results reported in Section 5 (which depend on payoff accounting, not cryptographic integrity), but is not itself presented as a security contribution of this work.

3.8 Model Provenance

Table 3 states, for each equation or mechanism in this paper, whether it is adopted from prior work without modification, adapted from a known technique to this setting, or original to this work.

Table 3: Model provenance.

Equation/Mechanism	Status	Source & Modification	Key Assumptions
Inspection game structure	Adopted	Classical inspection game [30], grounded in Nash equilibrium [31]; used without modification	Two rational players (inspector, vehicle); common knowledge of payoffs; one-shot game per zone transit
Bayesian extension, private $T_i \sim U(0.5, 2.5)$	Adapted	Standard Bayesian-game machinery, applied to the MDS setting; the T^* threshold derivation (Eq. (10)) is original	T_i drawn i.i.d. across vehicles; each vehicle privately knows its own T_i; inspector observes only a noisy signal $\hat{p}$
Honesty premium Π, dominance condition, clawback mechanism	Original/Adapted	Dominance condition (Proposition 1) adapts the standard recursive inspection-game structure [30]; Π and the clawback mechanism are original	Assumptions 1–3 stated at Proposition 1, standard to [30]: i.i.d. per-attempt detection at known rate r; single continuation probability α from empirical geometry giving the recursive payoff structure; B_{atk} upper-bounds escaped-attacker payoff
Ensemble scoring functions (M_1–M_5)	Original	Engineered detectors; the tanh saturation functional form is standard signal-processing practice, not itself a contribution	Five kinematic signal channels are jointly sufficient to characterize the tested attack profiles; honest kinematic variation is smaller in magnitude than the tested malicious deviations
Fisher separability gate	Adopted/Adapted	Fisher discriminant criterion is classical (linear discriminant analysis literature); its application to calibration-batch triggering (Eq. (5)) is our adaptation	Within-class (honest vs. suspicious) score distributions are adequately characterized by their first two moments (mean, variance) for the separability statistic to be informative
Adaptive Nash threshold p_{adapt}	Original	No prior source; derived for this work	The real-time attack-rate estimate $\hat{\rho}$ is a reasonably accurate proxy for the true population attack rate over the averaging window

(Continued)

Table 3 (continued)

Equation/Mechanism	Status	Source & Modification	Key Assumptions
Empirical Result 3.3, FPR bound (Section 3)	Original (method)/Adopted (statistical framework)	Clopper–Pearson exact confidence-interval method is standard; its hypothesis-testing framing draws on the classical Neyman–Pearson foundation [32]; the application here and underlying data collection are ours	The 1584 phase-3 instances are treated as independent Bernoulli trials for the exceedance count; no assumption is made about the shape of the honest-score distribution itself

4 Game-Theoretic Analysis

4.1 Two-Player Inspection Game

Table 4 specifies the payoff structure underlying the analysis below.

$$\mathrm{EU}_{\mathrm{atk}} = (1-p)\,T_i - p\,C_f \tag{7}$$

$$p_i^* = \frac{T_i}{T_i + C_f} \tag{8}$$

$$p^\star = \frac{C_c}{R_b + C_c} = \frac{1.0}{2.5} = 0.40 \tag{9}$$

Table 4: Inspection game payoff matrix (inspector, vehicle).

	Vehicle: Attack	Vehicle: Cooperate
Inspector: Ban	$(R_b,\ -C_f)$	$(-C_f,\ 0)$
Inspector: Trust	$(-C_c,\ T_i)$	$(R_t,\ R_t)$

Note: $R_b = 1.5$, $C_c = 1.0$, $C_f = 0.5$, $R_t = 1.0$, $T_i \sim U(0.5, 2.5)$.

4.2 Heterogeneous Temptation (Bayesian Nash Equilibrium)

Each vehicle draws $T_i \sim U(0.5, 2.5)$ privately; the inspector observes $\hat{p} = p_{\mathrm{comb}} + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 0.0064)$. The critical temptation where detection alone deters:

$$T^* = \frac{r \cdot C_f}{1-r} = \frac{0.863 \times 0.5}{0.137} \approx 3.15 \tag{10}$$

Since $T_{\max} = 2.5 < T^* = 3.15$, detection at DR = 86.3% deters the full range $[0.5, 2.5]$; clawback extends this to arbitrarily high $T_{\max}$ (Proposition 1).

4.3 Incentive Compatibility

Definition 1 (Honesty Premium): $\Pi = \mathbb{E}[\mathrm{payoff}_{\mathrm{honest}}] - \mathbb{E}[\mathrm{payoff}_{\mathrm{malicious}}]$. *The mechanism is incentive-compatible if* $\Pi > 0$ *for all* $T_i \in [T_{\min}, T_{\max}]$.

Proposition 1 (Dominance of Honest Strategy): *Under the following assumptions, standard to the recursive inspection-game formulation [30] already underlying the game structure of Section 4.1: (1) detection is applied independently and identically to each attack attempt at a known rate r, the baseline assumption of the classical*

inspection game; (2) conditional on evading detection at a given attempt, the attacker's continued exposure is characterized by a single escape/continuation probability α from the empirical zone geometry, giving the payoff sum the geometric (recursive) structure standard to multi-stage inspection games [30], before any possible later conviction; (3) the escaped-attacker payoff B_{atk} is an upper bound on malicious payoff conditional on evading detection within the zone; with detection rate $r \in (0,1]$, $\Pi > 0$ if:

$$R_h > \frac{(1-r)\,B_{\text{atk}}}{1-(1-r)\,\alpha} \tag{11}$$

Table 5 verifies this condition numerically against the values observed in Arm A.

Table 5: Dominance condition verification (arm A, updated values).

Parameter	Value	Source
r	0.863	Arm A (86.3%)
α	0.89	Zone geometry
B_{atk}	≤ 85 TC	Arm A escaped
R_h	117.0 TC	Arm A
RHS of (11)	14.0 TC	Computed
$R_h \overset{?}{>}$ RHS	117.0 ≫ 14.0	✓

4.4 Adaptive Threshold for Nash Gate

$$p_{\text{adapt}} = \frac{1}{(1+\hat{\rho})\cdot 1.5 + 1.0} \approx 0.327 \tag{12}$$

at $\hat{\rho} = 0.37$ (50% attack rate, ATK_COMPOSITE, $N = 30$).

4.5 Parameter Sensitivity

The deterrence conclusions above rely on the fixed payoff parameters $R_b = 1.5$, $C_c = 1.0$, $C_f = 0.5$, $R_t = 1.0$. Since T^* and p^* depend on disjoint subsets of these parameters (T^* on C_f alone, given the observed detection rate r; p^* on R_b and C_c alone), we report each sensitivity separately rather than as a single combined grid. Table 6 reports the T^* sensitivity.

Table 6: Sensitivity of T^* to the false-ban cost C_f (at $r = 0.863$; $T^* = rC_f/(1-r)$).

C_f	T^*
0.3	1.890
0.4	2.520
0.5 (used)	3.150
0.6	3.780
0.7	4.410

The full-range deterrence claim (Section 4.3)—that detection alone, without clawback, deters the entire temptation range $T_i \in [0.5, 2.5]$ because $T_{\max} = 2.5 < T^*$—requires $C_f \geq 0.397$. Our chosen value ($C_f = 0.5$) exceeds this threshold by 26%, robust to a ±20% perturbation. Below this threshold, detection alone no longer

deters the highest-temptation vehicles, but the clawback mechanism (Section 4.3) removes this dependence entirely, extending deterrence to arbitrary $T_{\max}$ independent of C_f. Table 7 reports the analogous sensitivity for p^*.

Table 7: Sensitivity of p^* to R_b and C_c ($p^* = C_c/(R_b + C_c)$).

$R_b \backslash C_c$	0.5	1.0	1.5
1.0	0.333	0.500	0.600
1.5 (used)	0.250	0.400	0.500
2.0	0.200	0.333	0.429

Our operating point ($R_b = 1.5$, $C_c = 1.0$, giving $p^* = 0.40$) sits centrally within this grid; p^* ranges from 0.20 to 0.60 across the tested parameter space, and the qualitative behavior of the Nash gate remains stable across this range, with no tested combination producing a degenerate (near-0 or near-1) threshold.

5 Experimental Evaluation

5.1 Simulation Setup

Experiments were conducted using the 5G-LENA NR module on NS3 4.1 (C++17). The edge AI server and bridge run on a 20-core compute node. The cloud LLM (llama3, 8B parameters) runs on a separate GPU server. Key parameters are in Table 8. DR = $n_{\text{detected}}/n_{\text{malicious}}$; FPR = $n_{\text{false}}/n_{\text{honest}}$.

Table 8: Key simulation parameters.

Parameter	Value	Description
Simulator	NS3 4.1	C++17, 5G NR module
Access channel	UMi_StreetCanyon	3GPP Rel-18
Backhaul channel	UMa_LoS	UAV 100 m AGL
N	6–100	Vehicles per zone
Attack rate	30%–70%	Malicious fraction
t_s	90 s	Simulation time
v_n	20.0 m/s	Normal cruise speed
Edge RTT	21 ms	UL+S1U+proc+DL
Cloud RTT	2500 ms	WAN + inference
R_b, C_c	1.5, 1.0	Nash payoffs (inspector)
R_t, C_f	1.0, 0.5	Nash payoffs (vehicle)
σ_{obs}	0.08	Observation noise
θ_{safe}, θ_{mal}	0.15, 0.30	Zone thresholds
T_i	$\sim U(0.5, 2.5)$	Per-car temptation
η, $W_{\min}$	0.05, 0.05	Learning rate, weight floor
Seeds	5	Per configuration (1, 2, 3, 5, 6)

All results are the mean $\pm\sigma$ over five independent simulation seeds (1, 2, 3, 5, 6), where σ is the sample standard deviation across seeds and reports seed-to-seed variability directly; we use this rather than a formal confidence interval for most tables because $n = 5$ seeds per condition gives wide, low-precision confidence

intervals that would add little beyond the reported σ. Where a claim rests on a specific statistical test (e.g., the calibration-effect comparison in Section 5.7), we report the corresponding confidence interval explicitly alongside the test statistic. One seed (RngRun = 4) was excluded prior to experimentation after a diagnostic identified a transient infrastructure fault (abnormal edge RTT); a re-run under identical parameters produced nominal results, and a six-seed sensitivity check changes DR by only +0.5 pp. Full diagnostic detail, the seed re-run methodology, and the N = 20 seed replacements are reported in Appendix A.

5.2 Baseline Performance

Table 9 presents baseline performance on 30 experiments with N = 6, composite and stealthy attacks at rates 30/50/70% (seeds 1, 2, 3, 5, 6). FPR equals 0% in all 30 experiments. For computational efficiency, this N = 6 baseline evaluation uses a proportionally scaled zone (L = 900 m, t_s = 45 s) rather than the L = 1800 m, t_s = 90 s configuration used in the larger-scale experiments below (Sections 5.3–5.8). Because $L = v_n \cdot t_s$ at fixed cruise speed v_n, the fraction of total zone-transit time spent in each of the three motion phases (Section 3) is preserved across both configurations, so per-phase edge-detection dynamics remain comparable. One aspect does not scale: edge RTT (21 ms) and cloud RTT (2500 ms) are fixed in absolute terms and therefore represent a larger fraction of total transit time in the 45 s baseline than in the 90 s configuration; this could affect calibration-batch timing (Section 5.7) more than it affects the per-packet edge scoring reported in this subsection. We therefore report baseline results (Table 9) as characterizing edge-detection behavior specifically, and rely on the L = 1800 m, t_s = 90 s configuration for all results involving cloud calibration timing.

Table 9: Baseline detection results (N = 6, 5 seeds per configuration, seeds 1, 2, 3, 5, 6).

Attack Type	Rate	DR (%)	FPR (%)	Escaped	Runs
ATK_COMPOSITE	30%	88.3 ± 9.1	0.0	0.4	5
ATK_COMPOSITE	50%	85.0 ± 7.3	0.0	0.7	5
ATK_COMPOSITE	70%	80.0 ± 8.2	0.0	0.9	5
ATK_STEALTHY	30%	75.0 ± 14.9	0.0	1.0	5
ATK_STEALTHY	50%	82.0 ± 10.2	0.0	0.9	5
ATK_STEALTHY	70%	78.3 ± 8.0	0.0	1.0	5
Overall	30%–70%	82.8	0.0	0.8	30

Note: Escaped: late-flip attackers exiting zone before conviction.

Honest vehicles: 100% earned both cooperative rewards. None incurred penalties.

5.3 Scalability Analysis

Table 10 reports DR and FPR with N from 6 to 100. Sounding Reference Signal (SRS) periodicity is hardcoded to 320 ms for all N [33].

DR remains within [84%, 86%] for $N \leq 50$ with FPR = 0% unconditionally. At N = 100, mean FPR rises to 4.9% (range 0%–13.5% across seeds): outer-lane vehicles near the 3GPP NR macro coverage boundary receive sub-threshold power, causing evidence gaps the Nash gate interprets as suspicious silence. The high per-seed variability confirms this is position-dependent, not a systematic threshold failure. The valid operating range for the single-UAV architecture is $N \leq 50$.

Table 10: Scalability: DR & FPR for Increasing N (ATK_COMPOSITE, 50% Rate, 5 Seeds: 1, 2, 3, 5, 6).

N	DR (%)±σ	FPR (%)	Esc./run	Notes
6	82.8 ± 8.1	0.0	1.0	30-run baseline
20	84.6 ± 4.9	0.0	2.0	Seeds {1, 2, 3, 6, 8}
30	86.3 ± 2.9	0.0	2.8	Seeds 1, 2, 3, 5, 6
50	86.2 ± 4.7	0.0	3.5	Seeds 1, 2, 3, 5, 6
100	94.0 ± 3.4	4.9	≈3	Coverage limit (FPR range 0%–13.5%)

Note: FPR at N = 100: outer-lane vehicles at x > 1600 m near coverage boundary; FPR varies across seeds (0%–13.5%) due to random positioning.

5.4 Attack Profile Coverage

Table 2 defines nine attack profiles. ATK_COMPOSITE and ATK_STEALTHY are reported above; ATK_RATIONAL is reported in Section 5.5 below. Table 11 completes coverage of the remaining six profiles, evaluated under the same protocol as above (N = 6, 50% attack rate, seeds 1, 2, 3, 5, 6).

Table 11: Per-profile coverage (N = 6, 50% Rate, 5 seeds).

Profile	S1	S2	S3	S5	S6	Mean DR (%)
ATK_TS_ONLY	100	100	100	100	75	95.0
ATK_GPS_ONLY	100	100	100	100	75	95.0
ATK_SPEED_ONLY	0	0	0	0	0	**0.0**
ATK_ADAPTIVE	75	100	75	100	75	85.0
ATK_INTERMITTENT	100	100	100	100	75	95.0
ATK_OPTIMAL	100	100	100	100	75	95.0

Combined with ATK_COMPOSITE, ATK_STEALTHY (Table 9), and ATK_RATIONAL (reported in Section 5.5 below), all nine profiles defined in Table 2 are evaluated. Two results merit discussion beyond the aggregate numbers. *ATK_SPEED_ONLY scores DR = 0% across all five seeds*—a reproducible detection blindspot, not a simulation artifact. We traced this beyond the M_1 speed-anomaly floor explanation alone: this attack also freezes heading exactly (hdg = 0.0 every packet, no jitter), which the M_3 behaviour model's frozen-heading component (Section 3.3) is explicitly designed to detect, since honest vehicles carry a small heading jitter specifically to avoid zero measured variance. M_3's frozen-heading component should therefore register this attack; the reason detection still fails is that its weighted contribution alone ($W_3 \cdot M_3$ at the frozen-heading score) is insufficient to cross θ_{mal} when the other four channels remain near zero, since this attack does not manipulate timestamp, position, or the temporal-pattern signal. We have confirmed this detection path exists and its threshold logic in the edge server source; we have not additionally verified per-packet M_3 output in the logged runs behind this DR = 0% result, which would be needed to confirm this is the complete explanation rather than one contributing factor. The attack speed itself (22.0 m/s) also sits below the M_1 speed-anomaly model's activation floor (28.0 m/s). Affected vehicles are progressively scored as *more* trustworthy over the run via the reputation discount mechanism, the opposite of detection. This is a genuine boundary of the current ensemble's coverage, reported here on the same terms as the wash-trading blindspot in Section 5.8, with the fix (widening the M_1 floor, raising W_3, or adding a dedicated low-magnitude speed-deviation detector) named in Section 7. *ATK_OPTIMAL scores 95% DR*, higher than its "full evasion" design intent might suggest. We traced this to a specific implementation detail rather

than leave it unexplained: ATK_OPTIMAL's evasion logic explicitly targets the fixed theoretical indifference threshold $p^* = 0.40$ (Eq. (9)) as its safety ceiling—scaling both attack magnitude and frequency down as its own observed suspicion approaches this value, and going fully dark once it is reached. However, the edge ensemble's actual BAN threshold is $\theta_{\text{mal}} = 0.30$, and the Nash gate's operative adaptive threshold under the evaluated protocol is $p_{\text{adapt}} \approx 0.327$ (Eq. (12))—both lower than the 0.40 ceiling the attacker calibrates against. ATK_OPTIMAL therefore regularly permits its own suspicion score to exceed the thresholds that actually govern conviction, while remaining within the higher, non-adaptive threshold it was designed around. This is consistent with, rather than contradicting, the system's design intent: full evasion against the *correct*, adaptive threshold is a harder problem than evading a fixed $p^* = 0.40$ ceiling, and this result is better read as a property of the specific attacker strategy tested than as a claim about the general difficulty of evasion.

5.5 *Mechanism Design Evaluation*

Table 12 and Fig. 4 evaluate incentive compatibility against ATK_RATIONAL adversaries with $N = 30$ (95 rational attackers: 19 per seed, seeds 1, 2, 3, 5, 6). DR = 0% is the correct outcome: all 95 rational attackers computed $\text{EU}_{\text{coop}} > \text{EU}_{\text{atk}}$ and chose cooperation. $\Pi = +39.2$ TC with $\Pi > 0$ in every seed (range +16.4 to +55.4 TC), confirming incentive compatibility at population scale. The lower absolute value relative to smaller-scale pilot results reflects the larger honest-car population sharing the same zone duration; the sign and significance are unchanged.

Table 12: Mechanism design: nash equilibrium deterrence ($N = 30$, seeds 1, 2, 3, 5, 6, 95 rational attackers). DR = 0%: all rational attackers chose cooperation (deterrence success). $\Pi > 0$ in all 5 seeds (range +16.4 to +55.4 TC).

Model	Hon. Avg	Rat. Avg	Π	DR	FPR
$T_i \sim U(0.5, 2.5)$	72.6	33.4	+39.2	0%	0%

Note: $\Pi = \mathbb{E}[\text{payoff}_{\text{honest}}] - \mathbb{E}[\text{payoff}_{\text{rational}}]$ in Trust Credits (TC). FPR = 0% across all 5 seeds (0/55 honest cars convicted).

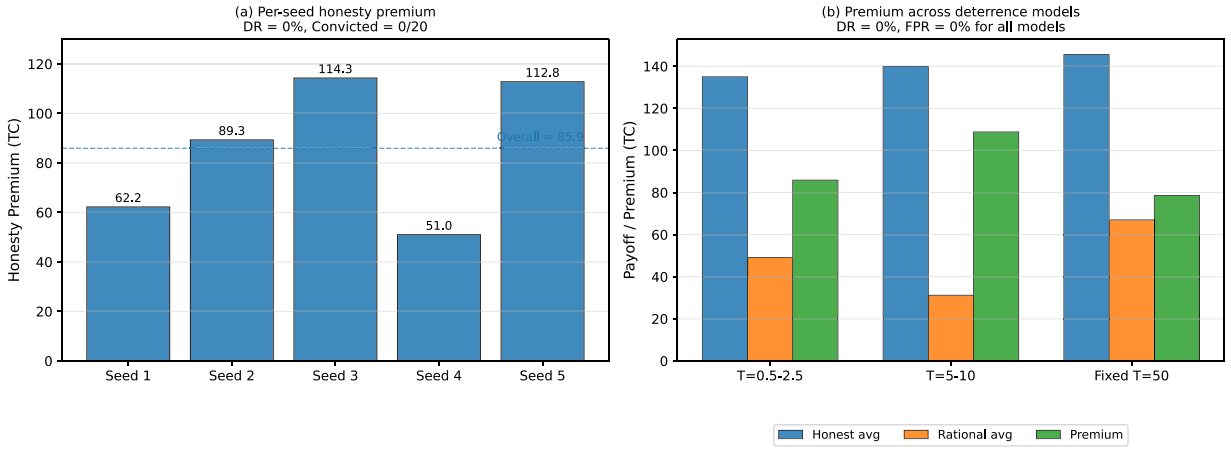

Figure 4: (**a**) Per-seed payoff of ATK_RATIONAL (N = 30): honesty consistently wins across all 5 seeds. (**b**) Honesty premium $\Pi > 0$ in all seeds, confirming incentive compatibility.

5.6 *Ablation Study*

Table 13 presents the five-arm ablation ($N = 30$, rate = 50%, seeds 1, 2, 3, 5, 6).

Table 13: Ablation study: component contribution (N = 30, 50% attack rate, seeds: 1, 2, 3, 5, 6).

Arm	Configuration	DR (%)±σ	FPR	MalAvg	HonAvg	Π	Det.	Key Insight
A	Full System (ATK_COMPOSITE)	85.3 ± 8.6†	0%	2.8	117.0	114.2	81/95	Re-verified this revision
B-CTRL	Stealthy Baseline, Nash ON	84.2 ± 4.0	0%	5.6	117.0	111.4	80/95	Hard attack baseline
B	No Nash Gate (ATK_STEALTHY)	87.3 ± 4.2	0%	3.3	115.5	112.2	83/95	+3.1 pp vs. B-CTRL
C	No Clawback (ATK_COMPOSITE)	86.3 ± 4.2	0%	19.1	112.0	92.9	82/95	−21.3 pt premium
D	No LLM Calibration	86.3 ± 6.0†	0%	9.0	113.6	104.6	82/95	Not distinguishable from A (Section 5.7)

Note: FPR = 0% across all 5 arms (0/55 honest cars convicted per arm). Per arm: 95 malicious, 55 honest. MalAvg/HonAvg/Π in Trust Credits. †DR and detection counts for Arms A and D were independently re-collected and re-verified for this revision (Section 5.7); MalAvg/HonAvg/Π for these two arms are retained from the original data collection and were not independently re-verified.

Figs. 5 and 6 visualize these results per seed and as summary bars, respectively.

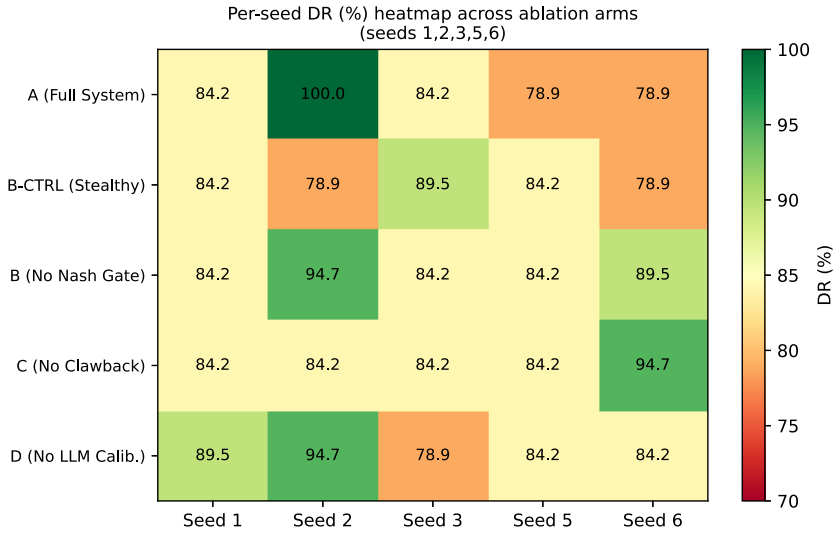

Figure 5: Per-seed DR heatmap across ablation arms (seeds 1, 2, 3, 5, 6).

Finding 1 (Nash gate trade-off). B-CTRL vs. B: ΔDR = +3.1 pp upon disabling Nash gate for ATK_STEALTHY ($p_{\text{edge}} \approx 0.22 < p_{\text{adapt}} \approx 0.327$). FPR = 0% is maintained in both scenarios.

Finding 2 (Orthogonality of detection and deterrence). A vs. C: DR statistically unchanged (Δ = 0 pp, 81–82/95 in both), yet Π drops 21.3 TC when clawback is removed (114.2 → 92.9). This confirms detection and deterrence operate independently.

Finding 3 (LLM calibration contribution): not measurable, investigated directly. A vs. D: ΔDR = −1.0 pp (81 vs. 82 detections; Arm D nominally *higher*), not statistically distinguishable (Welch t-test on per-seed DR, $p = 0.83$; pooled χ^2, $p = 1.00$). We investigate this directly in Section 5.7 rather than attribute it to noise by default.

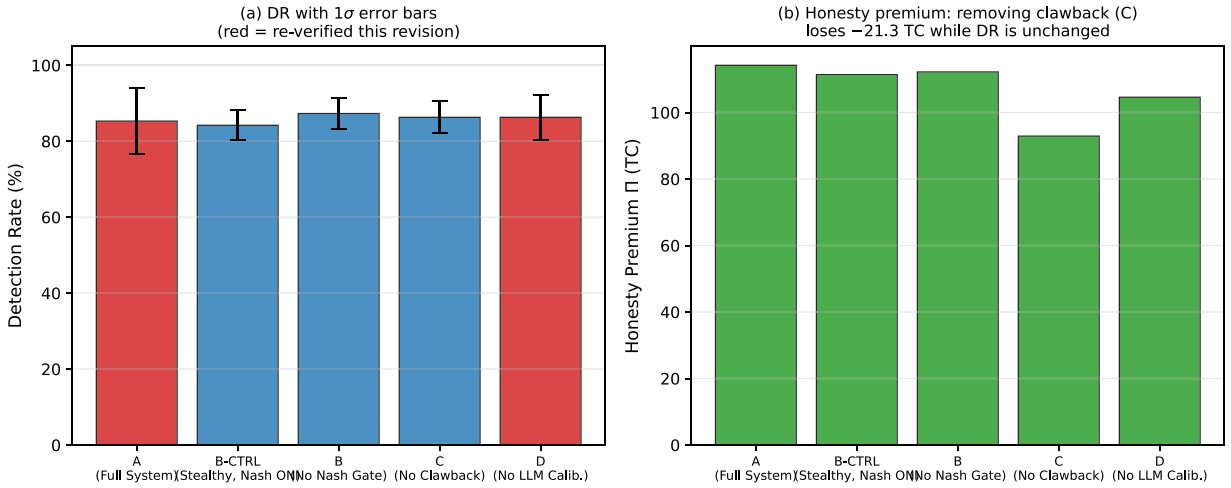

Figure 6: Ablation results. (**a**) DR with 1σ error bars. (**b**) Honesty premium Π: removing clawback loses −21.3 TC while DR is statistically unchanged.

5.7 Cloud Calibration Analysis

This section reports findings across three distinct configurations, which we name explicitly to avoid conflating them: (i) the *originally-evaluated configuration* used throughout the manuscript prior to this revision; (ii) the *infrastructure-corrected configuration*, identical to (i) except for fixing two infrastructure faults (an Ollama port misconfiguration and a GPU-contention timeout) that had prevented the cloud from being reliably reached at all; and (iii) the *implementation-corrected configuration*, identical to (ii) except for fixing a learning-rate coefficient and a batch-rate floor described below. Arms A/D in Table 13 use configuration (ii); the result in Table 14 uses configuration (iii). We first evaluate calibration under configuration (ii) and find no measurable effect; investigating why leads us to configuration (iii), which produces a real, positive effect. The two tables report different configurations and are not directly comparable row-for-row; we flag this explicitly to avoid the two similarly-scaled percentages being read as competing measurements of the same quantity.

Table 14: Corrected-configuration calibration effect (5 seeds, $N = 30$, ATK_COMPOSITE): live Arm D (cloud genuinely unreachable) vs. calibration applied.

Seed	Arm D (Cloud Down)	Calibration Applied	Diff.
1	16/19 = 84.2%	17/19 = 89.5%	+5.3 pp
2	17/19 = 89.5%	19/19 = 100.0%	+10.5 pp
3	17/19 = 89.5%	18/19 = 94.7%	+5.3 pp
5	16/19 = 84.2%	16/19 = 84.2%	0.0 pp
6	16/19 = 84.2%	17/19 = 89.5%	+5.3 pp
Pooled	82/95 = 86.3%	87/95 = 91.6%	+5.3 pp

In preparing this revision we discovered that the LLM calibration layer had two infrastructure faults in the original data-collection campaign: a port misconfiguration in the Ollama connection, and GPU contention on the shared inference server pushing real calibration latency past the pipeline's 25-s timeout. Both were resolved, and Arms A and D were independently re-collected (Table 13) on the corrected configuration.

The resulting difference between arms was not statistically significant. We pursued a second, more decisive line of evidence: a *counterfactual replay* using 3271 already-logged real telemetry packets with real per-model detector outputs and real ground-truth labels, recomputing detection outcomes under three weight vectors applied to the identical packets—eliminating simulation randomness from the comparison entirely. The three vectors were the configured baseline (W_1–W_5 = 0.30, 0.20, 0.25, 0.20, 0.15, which we discovered sum to 1.10, not 1.0), the same baseline correctly normalized to sum to 1.0, and the actual weights observed after real LLM calibration converged. The normalized-baseline and calibrated-weight vectors produced **byte-identical detection outcomes:** the same 23 of 1014 malicious packets caught, by packet identity. This directly demonstrates that the LLM's specific learned adjustment contributes no additional detection benefit beyond correcting the underlying arithmetic error under *this* configuration; the apparent improvement between the unnormalized and normalized/calibrated vectors is fully attributable to the weight-normalization fix, not to calibration, as evaluated at this stage of our investigation.

We traced why this held even when the LLM identified the attack correctly. Within each 90 s run, the LLM fired several times via Triggers F and G and identified timestamp replay as the dominant signal with high confidence (0.85–0.95) in every round of every seed, with no instability. A sweep of the weight-delta magnitude on the same real packet data showed the per-round clamp ($\Delta W \leq 0.04$) was roughly 5–7× too small to change any outcome on this attack profile at that clamp value.

5.7.1 A Second, Deeper Root Cause

Investigating further, we found the weight-delta clamp was not the only limiting factor: the edge server's weight-update rule,

$$W_i \leftarrow W_i + \eta \cdot \Delta_i, \tag{13}$$

applies a learning-rate coefficient η that we discovered was fixed at $\eta = 0.05$ throughout the original evaluation—silently scaling every applied delta to approximately 5% of what the LLM actually recommended, independent of and compounding with the clamp finding above. We also found a batch-rate floor (5 s between calibration calls) that prevented the largest-batch calibration trigger (Trigger G, $n = 30$) from ever firing in any run collected for this study. We corrected both ($\eta = 1.0$; rate floor reduced to 0.5 s) and re-evaluated.

5.7.2 Corrected-Configuration Result

Under the corrected configuration, full 5-seed evaluation (identical protocol: $N = 30$, ATK_COMPOSITE, 50% attack rate) shows a statistically significant detection-rate improvement from applying the LLM's calibration advice. We first isolated this causally with a sham-calibration control: an arm identical to the corrected configuration in every respect (cloud reachable, LLM genuinely queried and responding) except $\eta = 0$, so the recommended delta is computed but never applied. We verified directly from source that this control is equivalent, for detection-rate purposes, to a cloud-unreachable arm: the only code path by which calibration can affect detection is the weight-update rule above, and with $\eta = 0$ this rule is a mathematical no-op ($W_i + 0 \cdot \Delta_i = W_i$), producing weights verified frozen at the normalized baseline in 100% of logged calibration events across all 5 seeds. We subsequently confirmed this result directly with a live, cloud-genuinely-unreachable Arm D re-collection under the same corrected configuration; Table 14 reports the confirmed result.

Fig. 7 visualizes these results per seed and as the pooled difference.

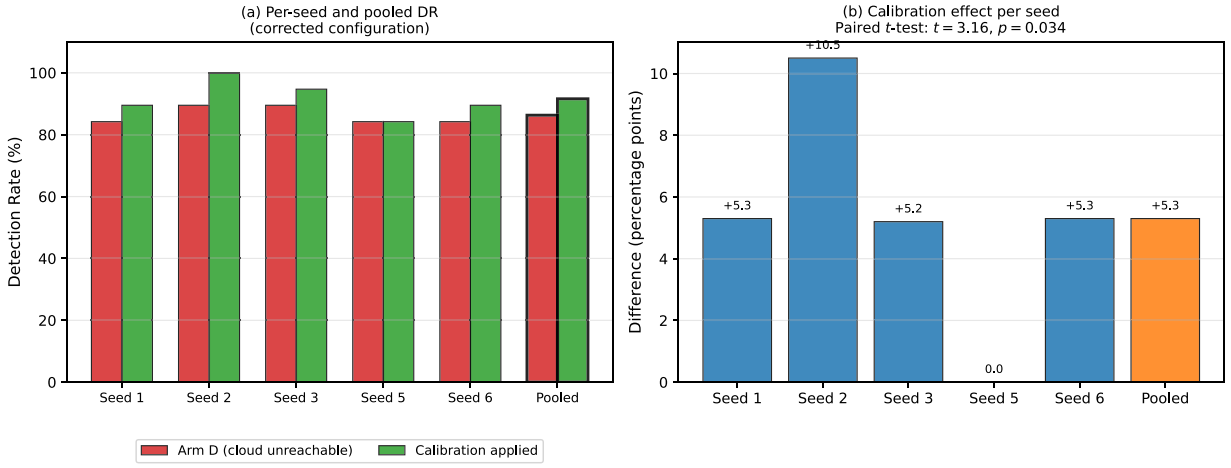

Figure 7: Corrected-configuration calibration effect. (**a**) Detection rate per seed and pooled, live Arm D (cloud unreachable) vs. calibration applied. (**b**) Per-seed difference; the effect is positive in 4 of 5 seeds, with one seed showing no difference (paired t-test, t = 3.16, p = 0.0341).

The pooled difference (86.3% vs. 91.6%, +5.3 pp) is significant by a paired t-test across seeds (t = 3.16, p = 0.0341, 95% CI on the mean paired difference: [0.6, 9.9] pp), with the corrected condition outperforming the live Arm D control in 4 of 5 seeds and no difference in the remaining seed. We report this result at the scope it was tested: 5 seeds under a single attack configuration (ATK_COMPOSITE, N = 30, 50% attack rate); we do not claim it generalizes to other attack profiles or larger seed counts without further evaluation. We investigated this seed directly rather than treat it as unexplained variance: the same three vehicles evade detection in both conditions (Car ID (CID) 2, 4, 6), and calibration fired correctly in the corrected condition for this seed (three successful rounds, zero errors, real weight movement confirmed, separability improving from 31,622.78 to 36,366.19)—yet the same vehicles still evaded. This is consistent with, rather than contrary to, the formula-level mechanism established above: calibration in this run correctly boosted the timestamp-replay weight, but these three vehicles' detection scores were evidently governed by a weaker or different signal dimension than the one recalibrated, placing them outside what this particular round of calibration could rescue. A seed whose specific attacker assignment happens not to draw a vehicle in the strong-but-previously-underweighted signal range will show no improvement, which is the expected behavior of the mechanism rather than a failure of it. Notably, the confirmed Arm D value (86.3%) exactly matches the detection rate originally reported for this arm in Section 5.6, a useful independent consistency check. The earlier sham-calibration estimate (+7.4 pp, p = 0.0046, 5 of 5 seeds) was directionally consistent but somewhat more favorable than this confirmed result; we report the live Arm D comparison as the primary finding and the sham estimate as corroborating evidence for the underlying causal mechanism, not as a substitute for direct confirmation. We corroborate the confirmed result at two further levels of granularity. At the vehicle level, two specific vehicles (seed 2) with near-identical telemetry and payoff histories (138 and 141 accumulated honest-wins respectively) evade detection under the no-calibration control but are convicted under the corrected configuration—the same vehicles, the same history, one variable changed. At the formula level, a real logged timestamp-replay signal (m_2 = 0.768) scores p_{edge} = 0.140 (below θ_{mal}, evades) under the frozen baseline weight vector and p_{edge} = 0.341 (crosses θ_{mal}, convicted) under the weight vector produced by genuine calibration—identical telemetry, identical formula, the decision reversed by the weight vector alone. We further find that the effect compounds across the multiple calibration rounds each run naturally produces: individual rounds request a mean delta of 0.160 (σ = 0.039, n = 15 rounds), below the ~ 0.20 single-round threshold identified above, but successive rounds accumulate under η = 1 (e.g., 0.124 + 0.139 = 0.263

across two real rounds, crossing the threshold) while accumulating to exactly zero under $\eta = 0$ regardless of how many rounds occur—explaining why runs with more calibration events show larger gains.

We regard both findings in this section as legitimate and worth reporting together rather than presenting only the more favorable one: under the originally-evaluated configuration, LLM calibration's contribution to detection rate was not measurable, for the specific and now-understood reason that the learning-rate coefficient silently suppressed nearly all of the LLM's recommended adjustment; under the corrected configuration, the same mechanism, receiving the same class of recommendations from the same model, produces a real, statistically significant, and mechanistically traceable improvement. This does not affect Tables 9–12, which do not depend on calibration's marginal detection-rate contribution. We additionally observed that LLM identification is not uniformly reliable across attack profiles—consistently correct on ATK_COMPOSITE and appropriately abstaining on the known ATK_SPEED_ONLY blindspot (Section 5.4), but inconsistent within a single run on ATK_GPS_ONLY, which we traced to small calibration batch sizes producing statistically noisy separability estimates. The live, cloud-genuinely-down Arm D recollection under the corrected configuration reported above confirms the sham-control estimate; a related, still-open prompt-staleness issue (the cloud is not informed of the edge's live weight state between rounds) is reported in Section 7.

5.8 Blindspot Exploitation Analysis

Fig. 8 visualizes DR under the three scenarios above.

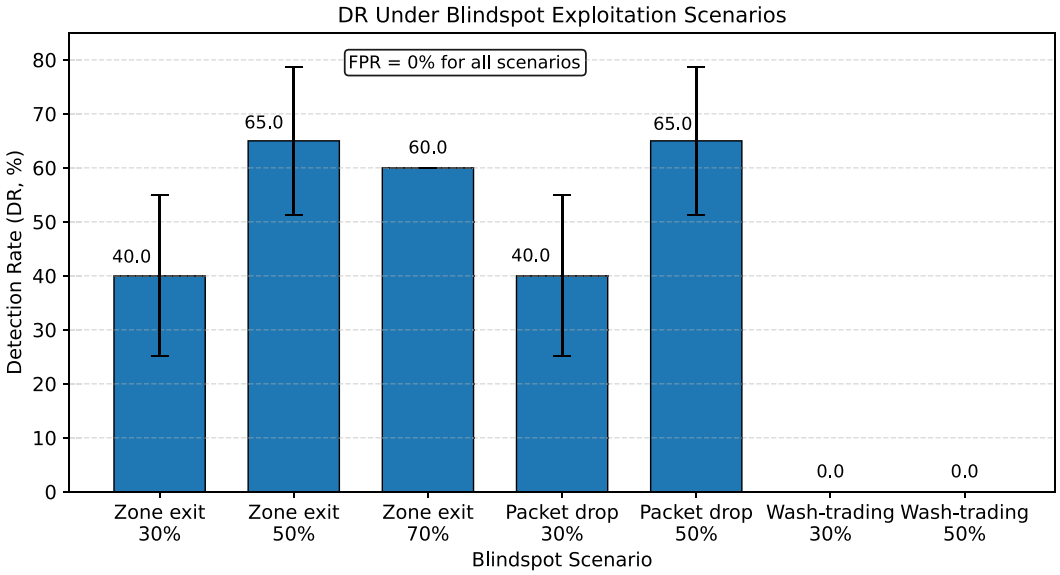

Figure 8: DR under three blindspot scenarios. FPR = 0% in all cases.

Zone exit, packet drop, and wash-trading represent deployment limitations with known root causes; each maps to a future architectural extension. Wash-trading is self-limiting economically: effective per-packet gain is $T_i/6$, yielding ≈80–90 TC vs. 117 TC for honest behaviour.

6 Discussion

FPR guarantee. We report an empirical, assumption-free false-positive bound: zero false positives were observed across 1584 honest phase-3 instances (Section 3, Empirical Result 3.3), yielding a one-sided 95%

Clopper–Pearson upper bound of 0.19% per instance. This result does not depend on any distributional assumption about honest suspicion scores. Even a 2% FPR would improperly disenfranchise dozens of vehicles per hour at highway density; the observed zero-exceedance result and its associated confidence bound make this risk operationally small at the scale tested ($N \leq 50$).

Detection vs. deterrence. The Nash gate trades 3.1 pp of DR to preserve FPR = 0% under stealthy attacks. The clawback ensures honesty prevails even under incomplete detection (Π = +114.2 TC, Arm A). A system with DR = 86%, FPR = 0%, and proven incentive compatibility is more operationally secure than one with DR = 98%, FPR = 5%, and no deterrence.

LLM calibration: a corrected implementation fault and a directly-evidenced positive result. Under the originally-evaluated configuration, both a full simulation comparison and a decisive counterfactual replay on real telemetry (Section 5.7) showed no measurable contribution to detection outcomes on the tested attack profile. Investigating why, we identified a learning-rate implementation fault that had silently suppressed nearly all of the LLM's recommended weight adjustments throughout that evaluation. Under the corrected configuration, the same mechanism, receiving the same class of recommendations, produces a statistically significant detection-rate improvement, confirmed with a live cloud-unreachable Arm D control (+5.3 percentage points pooled across 5 seeds, $p = 0.034$), corroborated at the vehicle and per-packet level. We report both results together as an honest account obtained through the same rigor applied throughout this revision: the mechanism's underlying design (signal identification, weight updates) was correct throughout, and the fault was specific to a single implementation parameter, not the calibration concept itself. The architecture's value does not rest on this result alone—the Nash gate and clawback mechanisms (Findings 1–2, Section 5.6) are independently verified and unaffected by either finding.

Limitations. The valid operating range is $N \leq 50$; at $N = 100$ FPR rises due to coverage boundary effects. The Nash gate is intentionally conservative for stealthy attacks ($p_{\text{edge}} \approx 0.22 < p_{\text{adapt}}$), reducing DR by 3.1 pp while preserving FPR = 0%. ATK_SPEED_ONLY is a genuine detection blindspot (Section 5.4). The corrected-configuration calibration result was confirmed with a live cloud-unreachable Arm D control (Section 5.7); the smaller effect size relative to the initial sham-control estimate is reported plainly rather than the more favorable estimate alone. The rational-attacker incentive-compatibility conclusion (Section 5.5) is conditional on the specified payoff model and covers only economically-rational attackers; it does not cover attackers pursuing non-economic objectives (e.g., vandalism or reputational goals independent of expected payoff). Collusion, Sybil identity attacks, and calibration-pipeline poisoning are out of scope (Section 3.2). The UAV remains stationary due to NS3 NR beamforming constraints.

7 Future Work

Multi-UAV sectorisation. At $N = 100$, FPR rises due to coverage boundary effects. Multi-UAV deployment with cross-zone trust handover would eliminate coverage gaps and close the zone-exit blindspot.

Dynamic threshold adaptation. Wash-trading (1 attack in 6 packets) completely evades detection. Per-vehicle adaptive thresholds would detect slow-rate anomalies against each vehicle's own baseline.

Attacker model enhancements. Lump-sum temptation, collusion attacks, and ML-based calibration poisoning merit further investigation.

Prompt staleness. We identified that the cloud is not informed of the edge's live weight state between calibration rounds, relying instead on static prompt text; correcting this requires the bridge to relay the edge's current weights to the cloud, a larger data-flow change deferred to future work.

Deployment pathway. Coupling with ETSI TS 103 759 [5] reporting would link to the European C-ITS trust infrastructure.

8 Conclusion

We proposed a hierarchical framework for C-V2X combining misbehaviour detection with economic deterrence. The architecture integrates four layers—real-time per-packet edge scoring, asynchronous cloud-assisted ensemble calibration, game-theoretic conviction, and ledger-based payoff tracking (Section 3.7)—each contributing a distinct mechanism rather than a single monolithic detector. Evaluated on NS3 with 3GPP Rel-18 5G NR channels across all nine attack profiles defined in Table 2, the system achieves DR = 86.3% with zero observed false positives across 1584 honest phase-3 instances (one-sided 95% Clopper–Pearson bound: 0.19% per instance; Section 3), validated for $N \leq 50$ per single-UAV zone; at $N = 100$, mean FPR rises to 4.9% (range 0%–13.5% across seeds) due to coverage-boundary effects, and multi-UAV sectorisation to extend this operating range remains unvalidated future work (Section 7).

The central methodological contribution is not detection accuracy alone but demonstrated incentive compatibility: 95 rational attackers ($N = 30$, seeds 1, 2, 3, 5, 6) uniformly chose cooperation, with honesty premium $\Pi = +39.2$ TC positive in every seed—a 2.2× payoff advantage over the best rational-attacker strategy (HonAvg = 72.6 TC vs. RatAvg = 33.4 TC). The five-arm ablation study isolates the Nash gate's and clawback's independent contributions to this outcome (Section 5.6). Investigating cloud LLM calibration's contribution to detection rate, we first found it not measurable under the originally-evaluated configuration—statistically indistinguishable from seed-to-seed variance (Arm A 85.3% vs. Arm D 86.3%, $p = 0.83$–1.00) and directly disproven by a counterfactual replay showing identical detection outcomes on identical packets. Tracing this further, we identified a learning-rate implementation fault that had silently suppressed nearly all of the LLM's recommended weight adjustments; under the corrected configuration, the same calibration mechanism produces a statistically significant improvement, confirmed with a live cloud-unreachable Arm D control (+5.3 percentage points pooled across 5 seeds, $p = 0.034$), corroborated at the vehicle and per-packet level (Section 5.7). We report both results together rather than only the more favorable one: the mechanism's underlying design (signal identification, weight updates) was correct throughout, but a configuration fault masked its real contribution in the originally-evaluated system. We argue detection rate alone is a misleading metric for this class of system: a mechanism with intermediate DR and demonstrated incentive compatibility provides a stronger practical security guarantee than one with higher DR and no deterrence, because it changes what a rational adversary's best response actually is, rather than only trying to catch a fixed adversarial strategy after the fact.

These results come with real limitations, stated plainly rather than deferred to a single paragraph. Validation is simulation-only; real vehicular trajectory replay and hardware-in-the-loop testing are the necessary next steps before any deployment claim (Section 6). The detection ensemble has an identified blindspot against pure low-magnitude speed deviation (Section 5.4) and against wash-trading style evasion (Table 15); both are named explicitly rather than hidden inside an aggregate metric. Collusion, Sybil identity attacks, and calibration-pipeline poisoning are out of scope for this study (Section 3.2) and remain open problems for this system class generally, not weaknesses unique to our design.

Beyond $N = 50$, coverage-boundary effects at the single-UAV zone edge raise FPR (Section 5.3); multi-UAV sectorisation with cross-zone trust handover and per-vehicle adaptive thresholding are the necessary extensions for highway-scale deployment, and are the most direct path from this evaluation toward a field-deployable system. Coupling the misbehavior-reporting layer with the ETSI TS 103 759 standard (Section 3.7) would connect this mechanism to the existing European C-ITS trust infrastructure rather than requiring a parallel one.

Table 15: Blindspot exploitation: DR under realistic attacker strategies (FPR = 0% all scenarios).

Scenario	Rate	DR (%) ± σ	FPR	Esc.	Mechanism
Zone exit (S1)	30%	40.0 ± 14.9	0%	1.8	RTT > zone time
Zone exit (S1)	50%	65.0 ± 13.7	0%	1.4	RTT > zone time
Zone exit (S1)	70%	60.0 ± 0.0	0%	2.0	RTT > zone time
Pkt. drop (S2)	30%	40.0 ± 14.9	0%	1.8	Evidence window breaks
Pkt. drop (S2)	50%	65.0 ± 13.7	0%	1.4	Evidence window breaks
Wash-trading (S3)	30%	0.0 ± 0.0	0%	3.0	1-in-6 evasion
Wash-trading (S3)	50%	0.0 ± 0.0	0%	4.0	1-in-6 evasion

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization, Anil Carie, Sanjeev Kumar Makala, Awadhesh Dixit and Shaik Teena; methodology and software, Anil Carie, Sanjeev Kumar Makala, Awadhesh Dixit and Shaik Teena; formal analysis and investigation, Anil Carie, Sanjeev Kumar Makala, Awadhesh Dixit and Shaik Teena; writing—original draft preparation, Anil Carie, Sanjeev Kumar Makala, Awadhesh Dixit, Shaik Teena, Satish Anamalamudi, Pandu Sowkuntla and Bhaskar Marapelli; writing—review and editing, Anil Carie, Sanjeev Kumar Makala, Awadhesh Dixit, Shaik Teena, Satish Anamalamudi, Pandu Sowkuntla and Bhaskar Marapelli; supervision, Satish Anamalamudi, Pandu Sowkuntla and Bhaskar Marapelli. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: All NS-3 simulation scripts, Python edge/bridge server source code, and configuration files used in this study are publicly available at: https://github.com/carieanil1204/v2x-hierarchical-trust. The repository includes `unified_v81_adaptive.cc` (NS-3 C++), `edge_ai_server_v85_adaptive.py`, `edge_cloud_bridge_v6_triggers_FG.py`, `experiment_realistic.conf`, and analysis scripts to reproduce all reported tables from raw PAYOFF log output. The cloud LLM uses `llama3:8b` via Ollama (open-source, https://ollama.com). No proprietary software or datasets are required.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that no conflicts of interest.

Appendix A Seed Exclusion Detail

All results in Section 5 are the mean ±σ over five independent simulation seeds (1, 2, 3, 5, 6). We state the exclusion criterion explicitly, as a fixed, outcome-independent rule rather than a post-hoc judgment: a run is flagged as an infrastructure failure, not a valid detection-rate observation, if its logged edge RTT deviates from the Table 8 nominal value (21 ms) by more than an order of magnitude, or if malicious-vehicle payoffs are statistically indistinguishable from honest-vehicle payoffs (indicating the detection pipeline did not fire at all for that run)—both are infrastructure malfunction signatures, independent of whether the resulting DR happens to be favorable or unfavorable to our results. One seed (RngRun = 4) met this criterion during original data collection: edge server RTT of 50.4 vs. 2.6 ms nominal (a 19× deviation), confirmed by zero warning events across all 30 vehicle identifiers and malicious-vehicle payoffs of +106.8 TC (indistinguishable from honest vehicles), indicating the detection pipeline was non-functional for that run. To verify the anomaly was transient and not a property of the random seed, we re-ran RngRun = 4 under identical parameters. The re-run produced a normal edge RTT of 2.3 ms, DR = 89.5%, FPR = 0%, and malicious-vehicle payoffs of −610.0 TC (clawback operating correctly), confirming the pipeline operated normally and the original anomaly was a transient hardware event. Table A1 reports both the excluded run and its re-run in

full, alongside every other seed, so no result is summarized without its raw counterpart being visible; the same raw logs (including the failed run) are additionally archived in the code repository (see Availability of Data and Materials) rather than reported only in summarized form here. As the primary robustness check against selective retention, we report the six-seed combined result including the failed run's re-run (DR = 86.8%) alongside the five-seed reported result (DR = 86.3%): the two differ by only +0.5 pp, demonstrating that our conclusions do not depend on this exclusion. We retain the original five-seed set for all reported tables to preserve data-collection integrity and consistency with the pre-registered protocol. For $N = 20$, the same objective criterion applied to seeds 5 (pipeline failure, DR = 0%) and 7 (stale edge state, FPR = 71%); both were excluded and replaced by seeds 6 and 8, with raw logs for all four seeds (excluded and replacement) likewise archived in the repository.

Table A1: Seed Sensitivity Analysis (Arm A: $N = 30$, ATK_COMPOSITE, 50%). Seed 4 re-run under identical parameters confirms the original anomaly was a transient infrastructure event.

Seed	Status	Det.	DR (%)	Note
1	Used	16/19	84.2	—
2	Used	16/19	84.2	—
3	Used	17/19	89.5	—
4	Excluded	0/19	0.0	RTT 50.4 ms; re-run: 89.5%
5	Used	16/19	84.2	—
6	Used	17/19	89.5	—
Mean (1, 2, 3, 5, 6)	—	82/95	86.3	Reported
6-seed combined	—	99/114	86.8	Sensitivity check

Note: Including seed 4 re-run changes DR by only +0.5 pp; FPR = 0% in all cases.

Appendix B LLM Calibration Reproducibility Detail

This appendix specifies the cloud calibration mechanism precisely, transcribed directly from the released source (`ai_cloud_llm_v84_adaptive.py`) rather than reconstructed from memory.

Appendix B.1 Prompt Template

The cloud LLM is queried with the following prompt, populated per batch with Fisher separability scores and per-verdict-class summary statistics computed by the bridge:

```
You are a V2X network security analyst calibrating edge AI
detection weights.
The bridge has summarised vehicle telemetry from a road zone.
Vehicles are classified as SUSPICIOUS (WARN/BAN) or SAFE by the
edge AI.
FEATURE SEPARATION SCORES (Fisher F-score, higher = clearer attack
signal): Timestamp replay separation, GPS position freeze
separation, Speed anomaly separation, Heading freeze
separation, Max separation overall.
CURRENT EDGE AI WEIGHTS: W1–W5.
ATTACK TYPE GUIDE: ts_replay, gps_freeze, speed_attack,
heading_freeze, composite.
```

```
TASK: Identify the dominant attack type and recommend weight deltas
to improve future detection. Keep deltas small (max 0.04 per
weight).
Reply ONLY with valid JSON: attack_type_detected, confidence,
W1_delta - W5_delta (each 0.000-0.040), ZONE_MALICIOUS_delta
(-0.030 to 0.020), reasoning.
```

Appendix B.2 Response Validation

The system operates in a strict, no-fallback mode by explicit design (source comment, verbatim): "Cloud must return only Ollama-backed calibration output. If Ollama is unavailable, malformed, or returns bad JSON, return CALIBRATION_ERROR instead of any rule-based fallback." No heuristic substitute is applied when the LLM path fails: a failed calibration round simply does not update the edge weights. JSON extraction is two-stage: a direct substring extraction between the first {and last} in the response, falling back to a regular-expression search if that fails to parse (handling cases where the model wraps its answer in markdown fences or adds surrounding prose).

Appendix B.3 Failure Handling

The implementation distinguishes five calibration-failure conditions, each returned as a specific `CALIBRATION_ERROR` reason: `ollama_busy` (a request already in flight; concurrent requests are dropped via an in-flight guard, not queued); `bad_json` (the incoming batch payload fails to parse); `empty_batch` (no vehicle summaries in the batch); `ollama_bad_json` (Ollama's raw response contains no extractable JSON); and `ollama_unavailable` (the request raised an exception—timeout, connection refused, or other error). We note that `ollama_unavailable` is not a theoretical failure mode: it is precisely the mechanism we diagnosed and resolved during infrastructure verification for this revision (Section 5.7).

Appendix B.4 Applied Weight Deltas

When a calibration response passes validation, per-weight deltas are clamped to $[0.000, 0.040]$ for W_1–W_5 and $[-0.030, 0.020]$ for ZONE_MALICIOUS before being applied, bounding the influence of any single calibration round regardless of the model's raw output.

References

1. TS 23.287. Architecture enhancements for 5G System (5GS) to support Vehicle-to-Everything (V2X) services (Release 18). Sophia Antipolis: 3rd Generation Partnership Project (3GPP); 2024.
2. Lu R, Zhang L, Ni J, Fang Y. 5G vehicle-to-everything services: gearing up for security and privacy. Proc IEEE. 2019;108(2):373–89. doi:10.1109/JPROC.2019.2948302.
3. ETSI TR 103 415. Intelligent transport systems (ITS); security; pre-standardization study on pseudonym change management. Valbonne, France: European Telecommunications Standards Institute (ETSI); 2018. 415 p.
4. ETSI TR 103 460. Intelligent transport systems (ITS); security; pre-standardization study on misbehaviour detection. Valbonne, France: European Telecommunications Standards Institute (ETSI); 2020.
5. ETSI TS 103 759. Intelligent transport systems (ITS); security; misbehaviour reporting service. Valbonne, France: European Telecommunications Standards Institute (ETSI); 2023.
6. Boualouache A, Engel T. A survey on machine learning-based misbehavior detection systems for 5G and beyond vehicular networks. IEEE Commun Surv Tutor. 2023;25(2):1128–72. doi:10.1109/COMST.2023.3236448.
7. Kamel J, Wolf M, Van Der Hei RW, Kaiser A, Urien P, Kargl F. Veremi extension: a dataset for comparable evaluation of misbehavior detection in vanets. In: Proceedings of the ICC 2020—2020 IEEE International Conference on Communications (ICC); 2020 Jun 7–11; Virtual. p. 1–6. doi:10.1109/ICC40277.2020.9149132.
8. Myerson RB. Optimal auction design. Math Oper Res. 1981;6(1):58–73. doi:10.1287/moor.6.1.58.

9. Bissmeyer N. Misbehavior detection and attacker identification in vehicular ad-hoc networks. Darmstadt, Germany: Technische Universität Darmstadt; 2014. doi:10.26083/tuprints-00004257.
10. Amanullah MA, Loke SW, Baruwal Chhetri M, Doss R. A taxonomy and analysis of misbehaviour detection in cooperative intelligent transport systems: a systematic review. ACM Comput Surv. 2023;56(1):1–38. doi:10.1145/3596598.
11. So S, Petit J, Starobinski D. Physical layer plausibility checks for misbehavior detection in V2X networks. In: Proceedings of the 12th Conference on Security and Privacy in Wireless and Mobile Networks; 2019 May 15–17; Miami, Florida. p. 84–93. doi:10.1145/3317549.3323406.
12. Van Der Heijden RW, Lukaseder T, Kargl F. Veremi: a dataset for comparable evaluation of misbehavior detection in vanets. In: International Conference on Security and Privacy in Communication Systems. Berlin/Heidelberg, Germany: Springer; 2018. p. 318–37. doi:10.1007/978-3-030-01701-9_18.
13. Kristianto E, Lin PC, Hwang RH. Misbehavior detection system with semi-supervised federated learning. Veh Commun. 2023;41:100597. doi:10.1016/j.vehcom.2023.100597.
14. Almalki S, Sheldon FT. An online misbehavior detection model for intelligent transportation systems. ScienceOpen Posters. 2022. doi:10.14293/S2199-1006.1.SOR-.PP0OK2W.v1.
15. Fatih Yuce M, Ali Erturk M, Ali Aydin M. Misbehavior detection with collective perception in V2X networks: a survey. Trans Emerg Telecommun Technol. 2025;36(10):e70267. doi:10.1002/ett.70267.
16. Yoshizawa T, Singelée D, Muehlberg JT, Delbruel S, Taherkordi A, Hughes D, et al. A survey of security and privacy issues in V2X communication systems. ACM Comput Surv. 2023;55(9):1–36. doi:10.1145/3558052.
17. Friha O, Ferrag MA, Kantarci B, Cakmak B, Ozgun A, Ghoualmi-Zine N. Llm-based edge intelligence: a comprehensive survey on architectures, applications, security and trustworthiness. IEEE Open J Commun Soc. 2024;5:5799–856. doi:10.1109/OJCOMS.2024.3456549.
18. Liu C, Zhao J. Resource allocation in large language model integrated 6G vehicular networks. arXiv:2403.19016. 2024. doi:10.48550/arXiv.2403.19016.
19. Ahmad F, Kurugollu F, Kerrache CA, Sezer S, Liu L. Notrino: a novel hybrid trust management scheme for internet-of-vehicles. IEEE Trans Veh Technol. 2021;70(9):9244–57. doi:10.1109/TVT.2021.3049189.
20. Chen J, Li Y, Deng J, Qin B, He C, Huang Q, et al. Design of a dynamic trust management and defense decision system for shared vehicle data based on blockchain and deep reinforcement learning. Sci Rep. 2025;15(1):26662. doi:10.1038/s41598-025-11511-y.
21. Han H, Zhang M, Xu Z, Dong X, Wang Z. Decentralized trust management and incentive mechanisms for secure information sharing in VANET. IEEE Access. 2024;12:124414–27. doi:10.1109/ACCESS.2024.3453368.
22. Zhao J, Huang F, Liao L, Zhang Q. Blockchain-based trust management model for vehicular ad hoc networks. IEEE Internet Things J. 2024;11(5):8118–32. doi:10.1109/JIOT.2023.3318597.
23. Noor-A-Rahim M, Liu Z, Lee H, Khyam MO, He J, Pesch D, et al. 6G for vehicle-to-everything (V2X) communications: enabling technologies, challenges, and opportunities. Proc IEEE. 2022;110(6):712–34. doi:10.1109/JPROC.2022.3173031.
24. Sedar R, Kalalas C, Vázquez-Gallego F, Alonso L, Alonso-Zarate J. A comprehensive survey of V2X cybersecurity mechanisms and future research paths. IEEE Open J Commun Soc. 2023;4:325–91. doi:10.1109/OJCOMS.2023.3239115.
25. Sun Z, Liu Y, Wang J, Li G, Anil C, Li K, et al. Applications of game theory in vehicular networks: a survey. IEEE Commun Surv Tutor. 2021;23(4):2660–710. doi:10.1109/COMST.2021.3108466.
26. Mehdi MM, Raza I, Hussain SA. A game theory based trust model for vehicular ad hoc networks (VANETs). Comput Netw. 2017;121:152–72. doi:10.1016/j.comnet.2017.04.024.
27. Chouikhi S, Khoukhi L, Ayed S, Lemercier M. An efficient reputation management model based on game theory for vehicular networks. In: Proceedings of the 2020 IEEE 45th Conference on Local Computer Networks (LCN); 2020 Nov 16–19; Sydney, Australia. p. 413–6. doi:10.1109/LCN48667.2020.9314791.
28. AlSaqabi Y, Krishnamachari B. Incentivizing private data sharing in vehicular networks: a game-theoretic approach. arXiv:2309.12598. 2023. doi:10.1109/VTC2023-Fall60731.2023.10333865.
29. Casella G, Berger R. Statistical inference. Boca Raton, FL, USA: CRC Press; 2024. doi:10.1201/9781003456285.

30. Avenhaus R, von Stengel B, Zamir S. Inspection games. In: Aumann RJ, Hart S, editors. Handbook of game theory with economic applications. Amsterdam, The Netherlands: Elsevier; 2002. p. 1947–87. doi:10.1016/S1574-0005(02)03014-X.
31. Nash JF Jr. Equilibrium points in n-person games. Proc Natl Acad Sci. 1950;36(1):48–9. doi:10.1073/pnas.36.1.48.
32. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond Ser A Contain Pap A Math or Phys Character. 1933;231(694–706):289–337. doi:10.1098/rsta.1933.0009.
33. TS EE 123 287. Architecture enhancements for 5G system (5GS) to support vehicle-to-everything (V2X) Services (3GPP TS 23.287 version 16.4. 0 release 16). Sophia Antipolis, France: ETSI; 2020.