Trust No One, Verify Everyone: An Identity-Centric Adaptive Framework for Zero Trust Architecture

Shreya Miskin[1], Sanchi Mane[2]
[1,2]*Cybersecurity Identity and Access Management Zero Trust Architecture*
doi.org/10.64643/ijirt.208446-459

***Abstract*—Traditional perimeter-based security relies on implicit trust and is increasingly ineffective against credential compromise, insider threats, and lateral movement in cloud and distributed environments. Zero Trust Architecture (ZTA) eliminates this implicit trust by continuously verifying identities, devices, and access requests. This paper investigates Identity and Access Management (IAM) as the foundation of ZTA, examining authentication, authorization, least-privilege access, human and non-human identities, and modern access-control models including RBAC, ABAC, ReBAC, and PBAC. It further integrates context-aware risk assessment and Continuous Access Evaluation (CAEP) to address dynamically changing security conditions. Based on this analysis, the study proposes an identity-centric adaptive Zero Trust framework that combines identity verification, contextual risk assessment, policy-based authorization, and continuous access evaluation into a unified access-decision mechanism. The proposed framework aims to enable adaptive access control, minimize unauthorized access and lateral movement, and strengthen security across cloud, hybrid, and distributed environments.**

***Index Terms*—Zero Trust Architecture (ZTA), Identity and Access Management (IAM), Context-Aware Risk Assessment, Continuous Access Evaluation (CAEP), Adaptive Access Control, Least-Privilege Access, Policy-Based Authorization.**

I. INTRODUCTION

1.1. Background and Motivation

Enterprise infrastructure has moved away from the fixed network perimeter that legacy security models assumed. Cloud-native applications, hybrid work, BYOD, and API-driven integration mean identities authenticate from heterogeneous, often untrusted networks. Zero Trust Architecture (ZTA) responds by requiring explicit, continuous verification of every subject requesting access, as codified in NIST SP 800-207. Machine and non-human identities (NHIs) service accounts, API keys, OAuth tokens, autonomous agents — now substantially outnumber human identities in the average enterprise (industry estimates range 45:1 to over 80:1) and are frequently unmanaged and long-lived.

1.2. Problem Statement

Successful authentication at time t0 does not imply a session, device, or credential remains trustworthy at a later time t1. The 2024 Midnight Blizzard (APT29) intrusion against Microsoft in which a compromised OAuth token retained valid access for months because it remained technically valid even as its behavioral usage pattern diverged from baseline illustrates precisely the failure mode point-in-time authentication cannot catch.

1.3. Research Gap

As detailed in Section 3, the literature provides strong individual treatments of identity verification, standardized continuous-evaluation signaling (OpenID CAEP), behavioral analytics, and NHI risk taxonomies, but comparatively little quantitatively-evaluated work fuses these signal classes into a single decision function with ablation evidence isolating each signal's contribution, evaluated under a methodology that avoids calibration/test leakage a gap a 2025 systematic review of 33 trust-scoring papers explicitly flags as a maturity gap in the field.

1.4. Research Questions

1. RQ1 Can a fused, identity-centric adaptive trust model improve access-control decision accuracy relative to static and context-only baselines, under a leakage-free evaluation methodology?
2. RQ2 Which signal classes contribute most significantly to correct decisions, and is identity alone sufficient?
3. RQ3 Does continuous, session-level reassessment

reduce the persistence of high-risk access after initial authentication, relative to a point-in-time baseline?
4. RQ4 How does detection performance degrade against an adversary who deliberately mimics legitimate signal values to evade detection?

1.5. Contributions Implemented and Empirically Evaluated

1. A weighted, five-signal adaptive trust-scoring function and four-tier decision policy, evaluated on a held-out test split with independently-calibrated thresholds (Sections 9–10).
2. A redesigned synthetic data-generation methodology producing overlapping, probabilistically-degraded, noisy signal distributions, explicitly built to avoid the circular-validation risk of cleanly-separable synthetic classes (Section 8).
3. A controlled, independently-calibrated ablation study isolating the marginal contribution of each signal class (Section 10.3).
4. A weight grid-search with an explicit calibration-vs-test overfitting check (Section 10.4).
5. Bootstrap confidence intervals and a McNemar significance test comparing AITS-ZTA against both baselines (Section 10.2).
6. An adversarial-evasion experiment measuring detection degradation as a simulated attacker mimics an increasing number of legitimate signals (Section 10.5).
7. A session-level, multi-timestep continuous-evaluation experiment with a time-to-containment (TTC) metric, directly and empirically supporting RQ3 within the simulation's scope (Section 10.6).

1.6. Contributions Proposed Architecture and Future Work

1. Temporal trust decay between observations (Section 6.4) — specified mathematically but not exercised in the current experiments.
2. Non-human-identity-specific step-up mechanisms (Section 5.3) — architecturally specified, not separately validated from human-identity mechanisms in the current simulation.
3. Real-world or red-team validation replacing the synthetic dataset (Section 15).
4. Privacy-preserving/federated trust computation (Section 12.1).

II. BACKGROUND AND THEORETICAL FOUNDATIONS

2.1. Zero Trust Architecture

NIST SP 800-207 defines Zero Trust as a set of concepts designed to minimize uncertainty in enforcing accurate, least-privilege, per-request access decisions on a network assumed compromised. Its Policy Engine / Policy Administrator / Policy Enforcement Point separation is the architectural seam this paper's trust-evaluation engine occupies.

2.2. Continuous Access Evaluation and CAEP

The OpenID Foundation's Continuous Access Evaluation Profile (CAEP), part of the Shared Signals Framework, formalizes real-time exchange of risk and security-state signals so access can be attenuated between authentication events.

CAEP has seen growing production adoption (including Microsoft Entra) and interoperability demonstrations through 2025, indicating continuous, event-driven evaluation is moving from concept to operational standard — the direction this paper's design targets compatibility with, without claiming to implement CAEP itself.

2.3. Risk-Based Authentication and Behavioral Analytics

Empirical studies of production risk-based authentication (RBA) systems show that complex, multi-signal combinations outperform simple heuristics in real deployments. User and Entity Behavior Analytics (UEBA) extends this with longitudinal behavioral baselines, directly relevant to detecting the credential-valid-but-behaviorally-anomalous pattern central to this paper's motivating incident.

2.4. Non-Human Identities

NHIs cannot complete interactive MFA, often hold broad standing privileges, and are reported by OWASP's 2025 NHI Top 10 and multiple industry sources as a rapidly growing, under-governed attack surface, with millions of exposed credentials documented in 2025 alone.

This motivates treating NHIs as a distinct subtype in the threat model (Section 4) and the simulation (Section 8), though full step-up-mechanism validation remains future work.

III. SYSTEMATIC LITERATURE REVIEW AND RESEARCH GAP

3.1. Method

Sources were identified via targeted search of ScienceDirect, arXiv, standards bodies (NIST, IETF, OpenID Foundation, OWASP) and current (2024–2026) industry/academic reports. This is a structured, representative review, not a full PRISMA-protocol systematic review — disclosed as a limitation (Section 14) rather than presented as exhaustive.

Table 1. Comparative literature summary (unchanged from initial review; see Revision Note for what changed in this version).

Source	Year	Focus	Dynamic Trust?	Behavior/NHI Signal?	NHI Coverage?	Evaluation
NIST SP 800-207	2020	Foundational ZTA reference architecture	Conceptual	No	Limited	Reference spec
Cloud Security Alliance, *Identity Is the Cornerstone of ZT*	2025	Identity as adaptive control plane	Conceptual	Partial	Partial	Industry report
Xu et al. Dynamic Trust Scoring for ZT at the Edge	2025—	Multi-factor context-aware trust scoring	Yes—	Partial—	No—	Methodology proposal
ScienceDirect SLR, *Scoring Algorithms for ZT-SDP*	2025 (33-paper review)	Systematic review of trust-scoring algorithms	Yes (surveyed)	Varies	Limited	Meta-analysis; flags lack of standardized attack evaluation
HC-ZTIA (Human-Centric ZT Architecture)	2026	Identity governance for Industry 5.0	Yes	Yes	Partial	Architecture + case illustration
Agentic AI Zero-Trust Framework	2025	Decentralized auth for AI agents	Yes	Partial	Yes (agents)	Framework proposal
OpenID Foundation, CAEP / SSF	2020–2025	Standardized continuous-event signaling	Yes (event-driven)	No	Partial	Interoperability demos
Wiefling et al., Real-World RBA Revisited	2023	Empirical study of production RBA	Yes	Partial	No	Empirical, real deployment
OWASP Non-Human Identities Top 10	2025	NHI-specific risk taxonomy	No (catalog)	N/A	Yes (primary)	Qualitative risk ranking

3.2. Identified Research Gap

Existing work provides strong individual treatments of identity verification, standardized signaling, contextual/behavioral risk, and NHI risk taxonomies, but limited quantitatively-evaluated work fuses these into a single decision function tested under an evaluation methodology that avoids calibration/test leakage, tests ablation variants independently-calibrated rather than at one shared threshold, or evaluates detection against a deliberately evasive adversary. This narrow, evidence-grounded combination not the individual components is this paper's positioning.

IV. PROBLEM FORMULATION AND THREAT MODEL

4.1. System Model and Assumptions

A Policy Decision Point (PDP) evaluates each access request from subjects S (human and non-human

identities) against resources R, and — distinct from conventional models — continues re-evaluating active sessions on event triggers and fixed intervals. We assume the identity provider's core authentication mechanism is trusted, telemetry sources deliver data with bounded latency (individual signals may still be degraded or spoofed — addressed below), and the enforcement point can act on PDP decisions within an operationally meaningful window.

4.2. Adversary Model

Table 2. Adversary classes and where each is empirically exercised in this revision (previously only the first five were modeled at all).

Adversary class	Capability	Primary target	Modeled in
External credential thief	Valid credentials via phishing/breach/info stealer	Human session, context signal	Static simulation + evasion
Compromised endpoint	Controls an enrolled device	Device trust signal	Static simulation
Session hijacker	Steals a live session token	Active session, behavior/context	Static simulation
Malicious insider	Legitimate credentials, authorized baseline access	Behavioral boundary	Static simulation
NHI/credential abuser	Exposed API key/token/OAuth secret	Non-human identity trust signal	Static simulation
Adaptive/evasive attacker	Deliberately mimics 1–4 legitimate signals	Detection threshold itself	Dedicated evasion experiment (10.5)
Multi-step session attacker	Escalates gradually across a live session	Continuous re-evaluation loop	Dedicated TTC experiment (10.6)

4.3. Trust-Engine Attack Surface

An adaptive trust system introduces its own attack surface: the signal-collection pipeline (spoofable telemetry), the weighting/threshold configuration (mis-calibration or deliberate manipulation risk), and the decision engine itself. This is analyzed in Section 11.8 and is treated as inherent to the approach, not eliminated by it.

V. PROPOSED AITS-ZTA FRAMEWORK

5.1. Architecture

Eight components feed a continuously re-invoked decision engine: (1) Identity Verification, (2) Device/Posture, (3) Context, (4) Behavioral, (5) Threat Intelligence layers feed the (6) Adaptive Trust/Risk Engine, which drives the (7) Policy Decision Point and (8) Continuous Monitoring/Reassessment loop, consistent with NIST SP 800-207's logical separation of decision from enforcement and compatible with CAEP-style event signaling.

5.2. Human and Non-Human Identity Handling (proposed, not separately validated)

For human subjects, Step-up maps to an interactive challenge. For NHIs, which cannot complete interactive challenges, Step-up is proposed to map to non-interactive mitigations — short-lived token reissuance with tightened scope or mandatory secondary attestation. The simulation (Section 8) includes NHI-abuse as a labeled attack subtype and evaluates detection of it, but does not separately implement or test the differentiated step-up mechanism itself; this remains a design proposal (Section 1.6).

VI. DYNAMIC TRUST MODEL

6.1. Trust Score

$T = wI \cdot I + wD \cdot Dv + wC \cdot C + wB \cdot B + wTh \cdot Th,\ \Sigma w = 1$

Initial weights (wI=0.30, wD=0.20, wC=0.15, wB=0.25, wTh=0.10) are expert-derived. Section 10.4 reports a grid search over the weight simplex, performed only on the calibration split, with performance on the untouched test split reported alongside to check for overfitting.

6.2. Temporal Trust Decay (proposed, not evaluated)

$T\_eff(t) = T(t0) \cdot e^{-\lambda(t-t0)} + T\_baseline \cdot (1 - e^{-\lambda(t-t0)})$

This decay model is specified for completeness but was not exercised in the present experiments, which use discrete per-timestep signal values in the session simulation (Section 10.6) rather than continuous decay

between observations. Evaluating this specific mechanism is future work (Section 15).

6.3. Access Decision Function

a(s,t) = Allow if T ≥ θ_high; Restrict if θ_med ≤ T < θ_high; Step-up if θ_low ≤ T < θ_med; Deny if T < θ_low. Section 10.6 reports how θ_high/θ_med/θ_low was derived from calibration-split score distributions after an initial fixed-offset derivation were found to misclassify typical legitimate sessions as Restrict — a genuine methodological finding disclosed rather than corrected silently.

VII. ADAPTIVE AND EXPLAINABLE ENFORCEMENT

The four-state policy (Allow → Restrict → Step-up → Deny/Revoke) allows blast-radius reduction before full denial. Each decision is accompanied by a rule-based rationale derived from each signal's weighted contribution to T an explanation mechanism, not a claim of formal explainable-AI (XAI) methodology, since no learned model is used.

VIII. SIMULATION-BASED IMPLEMENTATION

We implemented a Python 3 / NumPy / scikit-learn simulation testbed. This is stated explicitly as a synthetic simulation, not a production or red-team deployment.

8.1. Redesigned Synthetic Data Generation Addressing the Circular-Validation Risk

The original data generator (prior draft) drew legitimate signals from tight, high distributions (means ≈0.80–0.90) and attack signals from distributions pushed deterministically low (≈0.25–0.45) for whichever signal each attack subtype was defined to affect. Because the attack taxonomy and the trust model's signal set were designed together, this created a real risk of circular validation: the model appeared to perform strongly partly because the dataset was shaped around the same signals used to detect it. This is disclosed here rather than minimized, because it was the single most serious methodological weakness identified in the Part 1 audit. The regenerated dataset (used for every result in this revision) makes three changes: (1) legitimate events are drawn from a two-component mixture — 85% typical high-trust behavior and 15% atypical-but-legitimate behavior (travel, new device, off-hours work) with substantially degraded device/context/behavior signals despite being fully legitimate; (2) attack events apply signal degradation probabilistically (50–70% chance per affected signal, not guaranteed) and at variable magnitude, so a meaningful fraction of attacks look partially or fully normal on some or all signals insiders in particular deviate only mildly, overlapping substantially with the legitimate behavioral tail, and 40% of NHI-abuse events mimic normal API usage entirely; (3) independent, identically-distributed Gaussian measurement noise (σ=0.04) is added to every signal for every event regardless of label. Figure 1 shows the resulting trust-score distributions: legitimate and adversarial populations now overlap substantially (compare to a near-perfectly-separated distribution, which would indicate the earlier circular-validation risk had not been addressed).

Figure 1. Trust-score distributions for legitimate vs. adversarial test-set events under the redesigned, overlapping data generator. Substantial overlap around the calibrated threshold (0.77) is the intended, honest outcome of removing artificial separability.

8.2. Calibration/Test Split

Data (3,000 legitimate, 600 adversarial; 3,600 total) is split by stratified random sampling into a 60% calibration set (2,160 events) used exclusively for threshold selection, weight grid search, and ablation calibration, and a 40% held-out test set (1,440 events) used exclusively for the metrics reported in Section 10. No threshold, weight, or ablation decision is made using test-set labels.

8.3. Baselines

- Static RBAC — identity signal only, threshold calibrated on the calibration split (not fixed a priori).
- Context-only — equal-weighted identity + context, threshold calibrated identically.
- AITS-ZTA (full) — the five-signal weighted fusion model, threshold calibrated identically.

All three models are now calibrated using the same procedure (ROC/Youden's J on the calibration split), correcting the prior draft's unfair comparison of a calibrated proposed model against arbitrarily-thresholder baselines.

8.4. Reproducibility

Fixed seed (42), NumPy Generator API. The complete script (data generation, splitting, calibration, evaluation, ablation, grid search, sensitivity, evasion, session/TTC simulation, bootstrap CIs, McNemar test, and all figure generation) is provided as Appendix B and reproduces every number and figure in this paper.

IX. EXPERIMENTAL METHODOLOGY

9.1. Metrics

- TPR, FPR, Precision, F1 on the held-out test set.
- ROC-AUC and PR-AUC (PR-AUC reported because the dataset is class-imbalanced, 5:1 legitimate:adversarial).
- 95% bootstrap confidence intervals (2,000 resamples of the test set) for TPR, FPR, and F1.
- McNemar's test (continuity-corrected, paired on identical test events) comparing AITS-ZTA against each baseline.
- Time-to-containment (TTC), containment rate, and exposure-window comparison in the session-level experiment.
- Detection rate under adversarial evasion, at four escalating mimicry levels.

9.2. Threshold Calibration

For each model, the flagging threshold is swept across the score range on the calibration split only, selecting the value maximizing Youden's J (TPR − FPR). This threshold is then frozen and applied, unchanged, to the test split.

9.3. Ablation Procedure (Corrected)

Each of five nested signal combinations (identity only; +device; +context; +behavior; +threat intelligence) is treated as an independent model: weights are proportionally renormalized over the included signals, a threshold is calibrated for that specific combination on the calibration split, and metrics are reported on the test split. This directly corrects the prior draft's error of comparing an uncalibrated 3-tier full model against itself under one shared threshold, which had produced an uninterpretable non-monotonic result.

9.4. Weight Grid Search and Overfitting Check

A coarse grid search (steps of 0.1 across the weight simplex, 1,001 valid combinations summing to 1.0) selects the F1-maximizing weight vector on the calibration split. We then report that vector's F1 on both the calibration split (where it was selected) and the untouched test split, alongside the expert-derived weights' F1 on both splits, to make any overfitting gap explicit rather than presenting the grid-searched weights as validated.

9.5. Adversarial Evasion Procedure

A separate, held-out generator produces attackers that deliberately keep 1–4 of the five signals at legitimate-looking values (randomly selected per event) while leaving at least one signal genuinely degraded (perfect evasion across all five signals is not modeled, since a fully signal-indistinguishable event is definitionally unattributable as an attack by any behavioral system). Detection rate is measured at each evasion level for all three models using their previously calibrated thresholds.

9.6. Session-Level Continuous Evaluation Procedure

400 six-timestep sessions (200 benign, 200 attack) are simulated. Attack sessions inject anomalies starting at $t=2$ (context), $t=3$ (behavior), and $t \geq 4$ (threat intelligence), consistent with a gradual, non-evasive compromise.

At each timestep T(t) is recomputed and mapped to a decision via the three-tier policy (Section 6.3). Time-to-containment is the number of steps between anomaly onset and the first Step-up/Deny decision. A single point-in-time baseline (decision computed once at t=0 from identity +context, never re-evaluated) is run on the same attack sessions for comparison.

X. RESULTS

All results below are computed from the held-out test split unless otherwise stated, and every number has a corresponding line in the accompanying script (Appendix B).

10.1. Baseline Comparison (Held-Out Test Set, All Models Independently Calibrated)

Figure 2 (left/above) ROC curves and Figure 3 (below) Precision-Recall curves, held-out test set. AITS-ZTA dominates both baselines across the full operating range, not just at one chosen threshold.

Table 3. Held-out test-set performance, all thresholds calibrated identically on the calibration split only. Compare with the far more favorable (leakage-affected) numbers in the original draft this table is the corrected, defensible version.

Model	TPR	FPR	Precision	F1	ROC-AUC	PR-AUC
Static RBAC (identity only)	53.3%	33.3%	24.3%	0.334	0.619	0.233
Context-only (identity +context)	60.8%	36.5%	25.0%	0.354	0.672	0.291
AITS-ZTA (full, 5-signal)	72.9%	24.3%	37.6%	0.496	0.799	0.397

10.2.Confidence Intervals and Statistical Significance

Table 4. Bootstrap (2,000-resample) 95% confidence intervals on the test set. AITS-ZTA's TPR/F1 confidence intervals do not overlap either baseline, and its FPR interval is materially lower.

Model	TPR (95% CI)	FPR (95% CI)	F1 (95% CI)
Static RBAC	53.3% [46.9%, 59.5%]	33.3% [30.8%, 35.9%]	0.334 [0.289, 0.377]
Context-only	60.8% [54.6%, 67.0%]	36.5% [33.8%, 39.2%]	0.354 [0.312, 0.396]
AITS-ZTA (full)	72.9% [67.2%, 78.2%]	24.3% [21.9%, 26.7%]	0.496 [0.450, 0.538]

McNemar's test comparing paired, per-event correctness on the identical test set finds AITS-ZTA significantly outperforms both the context-only baseline ($\chi^2 = 87.00$, $p < 0.001$, 264 events corrected only by AITS-ZTA vs. 88 corrected only by context-only) and the static baseline ($\chi^2 = 54.27$, $p < 0.001$).

This supports RQ1: the improvement is unlikely to be an artifact of sampling noise, though it remains bound to this simulation's specific data-generating process (Section 14).

Figure 4. AITS-ZTA confusion matrix on the held-out test set (1,440 events).

10.3.Independently-Calibrated Ablation

Table 5. Ablation with each variant independently calibrated. ROC-AUC — a threshold-independent metric — now increases monotonically (0.619 → 0.625 → 0.683 → 0.766 → 0.799) across every added signal, confirming each signal class adds genuine discriminative information once the calibration artifact from the prior draft is removed.

Signal combination	TPR	FPR	F1	ROC-AUC
A: Identity only	53.3%	33.3%	0.334	0.619
B: Identity + Device	41.7%	25.8%	0.308	0.625
C: Identity + Device + Context	59.2%	33.9%	0.360	0.683
D: + Behavior	65.8%	24.8%	0.455	0.766
E: Full model (+ Threat Intel)	72.9%	24.3%	0.496	0.799

Figure 5. TPR/FPR by ablation combination. Combination B (identity+device) shows a small TPR dip relative to A at its own calibrated operating point even though its ROC-AUC is higher a genuine, disclosed finding: device posture alone shifts the optimal Youden threshold in a way that trades some sensitivity for a large specificity gain (FPR drops from 33.3% to 25.8%) rather than improving both simultaneously. This illustrates why ROC-AUC, not single-threshold TPR, is the correct metric for judging a signal's standalone contribution.

10.4.Weight Grid Search and Overfitting Check

Grid search over the weight simplex (steps of 0.1) selected weights (identity 0.30, device 0.10, context 0.10, behavior 0.20, threat 0.30), shifting weight from device/context toward threat intelligence relative to the expert-derived defaults. This vector achieved F1 = 0.556 on the calibration split (where it was selected) but F1 = 0.520 on the held-out test split — a 0.036 overfitting gap. The expert-derived weights scored F1 = 0.521 on calibration and F1 = 0.496 on test (a smaller 0.025 gap). Both gaps are modest, but the finding that grid-searched weights do not transfer perfectly to held-out data is reported directly rather than presenting the grid-searched configuration as a validated improvement; we retain the expert-derived weights as the paper's primary reported configuration for this reason.

10.5.Adversarial Evasion Experiment

Figure 6. Detection rate vs. evasion level (number of signals a simulated attacker deliberately keeps legitimate-looking, out of five).

Table 6. Detection rate under increasing adversarial evasion (300 evasive events per level, at least one signal always left degraded, since fully signal-indistinguishable events are not modeled).

Evasion level	Static RBAC	Context-only	AITS-ZTA (full)
1 signal mimicked	51.3%	59.7%	82.0%
2 signals mimicked	52.3%	61.3%	81.7%
3 signals mimicked	49.7%	60.7%	79.7%
4 signals mimicked	44.3%	59.7%	75.0%

AITS-ZTA's detection rate degrades gracefully from 82.0% to 75.0% as mimicry increases from one to four signals — rather than collapsing, and remains above both baselines at every level. This directly answers RQ4: the framework does not defeat a maximally evasive attacker (a 25% miss rate at the highest tested evasion level is a real, material gap, not a rounding error), but multi-signal fusion provides a measurable, degrading-not-collapsing margin over single- or dual-signal approaches.

We do not claim resilience to evasion beyond what is shown here; an attacker capable of controlling all five signals simultaneously is out of this experiment's scope and is noted as a limitation (Section 14).

10.6.Session-Level Continuous Evaluation and Time-to-Containment

Figure 7. Trust-score trajectory for one sampled attack session and one sampled benign session, with the (corrected — see 6.3) three-tier decision boundaries. The attack session crosses the Step-up/Deny boundary by t=3, one step after anomaly onset at t=2.

This experiment directly and empirically supports RQ3 within its scope: continuous re-evaluation contains 100% of the simulated gradual, non-evasive multi-step attacks, at a median of one evaluation step after anomaly onset, versus 26% for a baseline architecture that authorizes once and never reconsiders.

Table 7. Containment comparison, non-evasive gradual-compromise sessions. The point-in-time baseline's 26% figure reflects sessions that happened to be denied at t=0 before any anomaly occurred, not sessions that responded to the attack — by construction it cannot react after t=0.

Metric	AITS-ZTA (continuous)	Point-in-time baseline
Attack sessions simulated	200	200 (same sessions)
Sessions contained (Step-up/Deny after onset)	200 (100.0%)	52 (26.0%)
Median time-to-containment	1 evaluation step	N/A (no re-evaluation)
Mean time-to-containment	0.65 steps	N/A

This result should be read alongside Section 10.5: these attack sessions are not evasive, and Section 14 explicitly notes that combining the evasion and continuous-evaluation experiments an adaptive attacker operating across a live session was not implemented and is the most important remaining gap before this specific claim could be considered robust.

10.7.Corrected Three-Tier Threshold Derivation

An earlier version of this experiment set θ_{high} and θ_{low} as fixed offsets (±0.12/0.15) from the calibrated flagging threshold. This produced a disclosed artifact: because the redesigned legitimate-event distribution has mean full-model score ≈0.82 with meaningful spread, a substantial share of ordinary legitimate sessions fell below θ_{high} and were permanently classified Restrict rather than allow.

We corrected this by deriving θ_{high} from the 30th percentile of legitimate calibration-split scores and θ_{low} from the 70th percentile of adversarial calibration-split scores, so each boundary is anchored to the actual population it is meant to separate rather than an arbitrary offset.

The resulting band is narrow (θ_{high}=0.787, θ_{med}=0.772, θ_{low}=0.760), which itself is an honest finding: the overlapping, realistic dataset does not support widely-separated tiers, and an operational deployment would need materially better signal quality than modeled here to support a wider, more forgiving Restrict band.

10.8.Latency

Table 8. In-process Python computation time only (2,000 repetitions, single-threaded, warm-started), measured with perf_counter.

Model	Mean	Median	p95	Std dev
Static RBAC	0.0001 ms	0.0001 ms	0.0001 ms	0.0000 ms
Context-only	0.0001 ms	0.0001 ms	0.0002 ms	0.0001 ms
AITS-ZTA (full)	0.0002 ms	0.0002 ms	0.0003 ms	0.0001 ms

This is pure arithmetic-evaluation time for a five-term weighted sum and is, unsurprisingly, on the order of tenths of a microsecond regardless of model. It is explicitly NOT a measurement of production decision latency: it excludes signal collection, network I/O, telemetry aggregation, and policy-enforcement-point round trips, all of which would dominate real deployment latency and are not modeled in this simulation. The earlier draft's ~38ms figure was a simulated, not measured, illustrative placeholder and has been replaced here with an actual measurement labeled for exactly what it is.

XI. SECURITY ANALYSIS

Language throughout this section is deliberately bounded: the framework reduces detection blind spots and limits the persistence of high-risk access under the simulated conditions tested; it does not prevent or eliminate any attack class.

11.1.– 11.6. Per-Attack-Class Findings

- Credential theft: captured primarily via context/threat-intelligence degradation; Table 5 shows this signal combination materially improves ROC-AUC over identity alone.
- Compromised endpoint: device-posture signal is the sole indicator when a device is compromised but behavior/context remain normal; ablation combination B's ROC-AUC gain (0.619→0.625) is small, indicating device posture alone is the weakest single addition in this simulation and a real deployment should not rely on it in isolation.
- Insider threat: identity/device checks are legitimate by construction, so only behavioral degradation exists — this is the clearest evidence that identity-only control (Static RBAC) is structurally blind to this class.
- Session hijacking: behavior and context both degrade, giving the multi-signal model redundancy against this class specifically.
- NHI abuse: behavior/threat-intelligence degrade probabilistically; 40% of simulated NHI-abuse events mimic normal usage entirely and are, by design, difficult for any signal-based approach to catch — a limitation of behavioral detection generally, not specific to this framework.
- Adaptive evasion: Section 10.5 shows detection degrades from 82.0% to 75.0% as mimicry increases; a fully signal-controlling attacker is out of scope.

11.7. Trust Engine as a High-Value Target

The engine aggregating and acting on every session's signal is itself a concentrated target: compromise, denial-of-service, or signal-spoofing against it are not eliminated by the framework's design and are proposed to be mitigated (not tested here) via change-controlled threshold/weight updates, redundant signal sources, and a fail-conservative default (deny/restrict) on engine unavailability.

XII. PRIVACY, PERFORMANCE, AND SCALABILITY

Continuous behavioral/contextual monitoring collects sensitive data; data-minimization, bounded retention, and pseudonymization are proposed design requirements, not implemented or evaluated here. The trust-fusion function itself is O (1) per decision (Table 8) and unlikely to be a scalability bottleneck; the binding constraint in a real deployment would be the signal-collection pipeline's throughput at the NHI scale documented in Section 2.4 (tens of thousands to millions of identities), which this simulation does not model. AITS-ZTA's calibrated 24.3% FPR (Table 3) is a materially higher step-up/restrict burden than the earlier draft's (leakage-affected) 2.6% figure suggested this is the honest cost of the accuracy gain and should be weighed explicitly against organizational risk tolerance, not minimized.

XIII. DISCUSSION

The corrected results are less impressive on paper than the original draft's — TPR fell from a claimed 95.2% to a defensible 72.9%, and FPR rose from 2.6% to 24.3% — and that is the correct outcome of removing data leakage and dataset separability-by-construction rather than a weakening of the underlying design. ROC-AUC (0.799, threshold-independent) is arguably the more honest headline number, and it still shows a clear, statistically significant (Section 10.2) advantage for signal fusion over both baselines across the full operating range, not just at one convenient threshold. The ablation and evasion experiments together answer RQ2 and RQ4 with genuine nuance: behavioral and threat-intelligence signals matter most, identity alone is close to the ROC-AUC floor (0.619), and even the full model has a real, non-trivial failure boundary against a sufficiently evasive attacker. The session-level experiment gives RQ3 its first direct empirical support, but only for non-evasive, gradual compromise combining evasion with session dynamics remains open (Section 14).

XIV. LIMITATIONS

14.1. The dataset remains synthetic. The redesign (Section 8.1) removes the worst circular-validation risk but does not make the data real; absolute numbers should be read as illustrative of relative model behavior under this specific data-generating process, not as production accuracy claims.

14.2. The adversarial-evasion and session-level/TTC experiments were run separately; an attacker who is simultaneously evasive and operating across a multi-step session — arguably the most realistic sophisticated-adversary scenario — was not tested and is the most important remaining experimental gap.

14.3. Temporal trust decay (Section 6.2) is specified mathematically but not exercised; the session experiment uses discrete per-timestep values, not continuous decay between observations.

14.4. The weight grid search (Section 10.4) is coarse (0.1 steps) and shows a measurable, disclosed overfitting gap; a finer or cross-validated search was not performed given time constraints.

14.5. The literature review (~16 primary sources) is a structured, representative survey, not a full PRISMA-protocol systematic review, and does not include paywalled IEEE/ACM/Springer venues.

14.6. NHI-specific step-up mechanisms (Section 5.2) remain a design proposal; only NHI-abuse detection, not the differentiated enforcement mechanism, was tested.

14.7. The trust engine's own attack surface (Section 11.7) is discussed but not penetration-tested or formally verified.

14.8. All confidence intervals and significance tests are computed against this simulation's specific data-generating process and do not generalize to other datasets or deployments.

XV. FUTURE RESEARCH DIRECTIONS

15.1. Combined evasive + multi-step-session evaluation — the single most important next experiment identified by this revision.

15.2. Real-world or red-team-generated telemetry to replace the synthetic generator entirely.

15.3. Finer-grained or cross-validated weight optimization with formal held-out reporting at every stage.

15.4. Empirical evaluation of the temporal trust-decay mechanism (Section 6.2) using continuous, not discrete-timestep, session data.

15.5. Dedicated, separately-validated NHI step-up mechanism testing.

15.6. Formal or empirical analysis of the trust-engine's own attack surface (policy manipulation, engine unavailability, signal spoofing).

XVI. CONCLUSION

This revision replaces a methodologically compromised evaluation — threshold calibration and performance reporting on the same data, a synthetic dataset separable largely by construction, and an ablation study comparing calibrated against uncalibrated models — with a corrected pipeline: stratified calibration/test splitting, independently-calibrated baselines and ablation variants, bootstrap confidence intervals, a paired significance test, an adversarial-evasion experiment, and a genuine session-level continuous-evaluation experiment with a

time-to-containment metric. The corrected numbers are less flattering (TPR 72.9% vs. the original 95.2%; FPR 24.3% vs. 2.6%) but are defensible under exactly the kind of scrutiny a national-level conference judge is likely to apply. The central findings survive the correction: multi-signal fusion significantly outperforms static and context-only baselines (ROC-AUC 0.799 vs. 0.619/0.672, McNemar $p<0.001$), behavioral and threat-intelligence signals contribute the most, detection degrades gracefully rather than collapsing under adversarial evasion, and continuous re-evaluation empirically contains gradual multi-step attacks far more effectively than a point-in-time baseline (100% vs. 26%). The framework's real, disclosed limitations — synthetic data, untested evasive-plus-continuous attacks, and a 24% false-positive rate that would carry real operational cost — are presented as the necessary next steps, not omitted to preserve a stronger-sounding claim.

REFERENCES:

[1] Y. He, D. Huang, L. Chen, Y. Ni, and X. Ma, "A survey on zero trust architecture: Challenges and future trends," *Wireless Communications and Mobile Computing*, vol. 2022, Art. no. 6476274, 13 pp., 2022, doi: 10.1155/2022/6476274.

[2] M. L. Gambo and A. Almulhem, "Zero trust architecture: A systematic literature review," CoRR, abs/2503.11659, 2025, doi: 10.1007/s10922-025-09998-x.

[3] C. Shepherd, "Zero Trust Architecture: Framework and Case Study," Graduate Student Project, Cyber Operations and Resilience Program, Boise State University, Boise, ID, USA, 2022.

[4] "Zero Trust Architecture in Modern Cybersecurity: Effectiveness, Limitations, and Adoption Barriers," Iconic Research and Engineering Journals, vol. 9, no. 10, pp. 4994–4995, 2026, doi: 10.64388/IREV9I10-1716226.

[5] S. Gandikota, "Zero trust and AI-driven identity and access management for secure multi-cloud enterprise architectures," Journal of Computational Analysis and Applications, vol. 31, no. 4, pp. 2846–2860, 2023.

[6] "Zero Trust Identity and Access Management (IAM) in Multi-Cloud Environments," 2024.

[7] RSA, "Zero Trust Identity and Access Management," RSA, Solution Brief.

[8] Versa Networks, "Versa Zero-Trust Architecture Overview," Versa Networks, Solution Brief.

[9] I. Kovacevic, M. Stojkov, and M. Simic, "Authentication and identity management based on zero trust security model in micro-cloud environment," CoRR, abs/2410.21870, 2024, doi: 10.1007/978-3-031-50755-7_45.

[10] M. Mehata, "Dynamic Zero Trust Access Control: Fortifying Security in Mobile-Cloud Environment," M.S. thesis, Dept. Computer Science and Information Systems, Youngstown State University, Youngstown, OH, USA, 2025.

[11] S. Gandikota, "Zero trust and AI-driven identity and access management for secure multi-cloud enterprise architectures," Journal of Computational Analysis and Applications, vol. 31, no. 4, pp. 2846–2860, 2023.

[12] R. Chandramouli and Z. Butcher, "A zero trust architecture model for access control in cloud-native applications in multi-location environments," NIST Special Publication 800-207A, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2023, doi: 10.6028/NIST.SP.800-207A.

[13] OpenText, "Identity first step towards zero trust architecture," OpenText, Solution Flyer, 2023.

[14] AIRO, "Zero Trust Architecture," AIRO, 2026.

[15] Ampcus Cyber, "5 Key Pillars of Zero Trust Architecture," Ampcus Cyber, 2025.

[16] OpenText, "State of Zero Trust in the Enterprise: Shift to Identity-Powered Security," OpenText, 2023.

[17] GoodAccess, "A Practical Guide to Implementing Zero Trust Security," GoodAccess, 2024.

[18] B. Dey, "Impact of Zero Trust Architecture on Stock Exchange Network Security: A Quantitative Evaluation of Access Control and Incident Reduction," ASRC Procedia: Global Perspectives in Science and Scholarship, vol. 1, no. 1, pp. 2528–2569, 2025, doi: 10.63125/dcvska73.

[19] M. Mangla, "AI-driven zero trust architecture: A scalable framework for threat detection and adaptive access control," International Journal Science and Technology, vol. 2, no. 3, pp. 117–124, 2023, doi: 10.56127/ijst. v2i3.2274.

[20] "Web-Biometrics for User Authenticity Verification in Zero Trust Access Control," IEEE

Access, vol. 12, pp. 129611–129622, 2024, doi: 10.1109/ACCESS.2024.3413696.

[21] I. Uzougbo Onwuegbuzie and A. O. Augustine, "A review of authentication and authorization mechanisms in zero trust architecture: Evolution and efficiency," TechSphere Journal of Pure and Applied Sciences, 2025.

[22] D. Schönle and C. Reich, "Hardening ZTA by AI-driven observation: Survey and case studies of aerospace cyber defence architecture," Authorea, Jun. 23, 2025, doi: 10.22541/au.175070963.35708723/v1.

[23] S. Al-Tamimi, Q. A. Al-Haija, and K. Alrawashdeh, "Zero-trust architecture for securing Internet of Things (IoT) networks: A review," in Proc. 2024 5th Int. Conf. Communications, Information, Electronic and Energy Systems (CIEES), Veliko Tarnovo, Bulgaria, 2024, pp. 1–6, doi: 10.1109/CIEES62939.2024.10811176.

[24] K. Sabatini, "Zero-trust architecture enforcement through generative AI: Continuous identity verification and adaptive access control," 2026.

[25] O. Alexander, "Zero trust-based centralized authentication: A framework for continuous identity verification and adaptive enterprise access control," 2026.

[26] J. Rhoads and A. Smith, "Effectiveness of continuous verification and micro-segmentation in enhancing cybersecurity through zero trust architecture," SageScience Journal of Cybersecurity, vol. 4, no. 12, 2024.

[27] K. Rustam, "Architectural principles of zero trust privileged access management in modern corporate infrastructures," Colombo Scientific Journal of Computer Science & Information System, vol. 11, no. 5, pp. 33–39, 2026.

[28] T. J. Olorunlana, "Least privilege and access control principles in enterprise network security: A comparative analysis of the United States and global practices," International Journal of Science, Architecture, Technology, and Environment, vol. 2, no. 12, Dec. 2025, doi: 10.63680/ijsate1225001.001.

[29] M. Mehata, "Dynamic Zero Trust Access Control: Fortifying Security in Mobile-Cloud Environment," M.S. thesis, Youngstown State University, Youngstown, OH, USA, 2025.

[30] A. Smith and J. Rhoads, "Effectiveness of continuous verification and micro-segmentation in enhancing cybersecurity through zero trust architecture," SageScience Journal of Cybersecurity, vol. 4, no. 12, 2024.

[31] "Zero Trust Architecture," IEEE Xplore, 2026.