SpacSec 2026 | Technical Track

Application of Zero Trust Architecture for Information Security of Large Language Models in Corporate Infrastructure

Authors: O.S. Lauta, D.S. Mahov, N.A. Vasilev

Affiliations: Admiral Makarov State University of Maritime and Inland Shipping, Special Technology Center, Military Telecommunication Academy (Russian Federation)

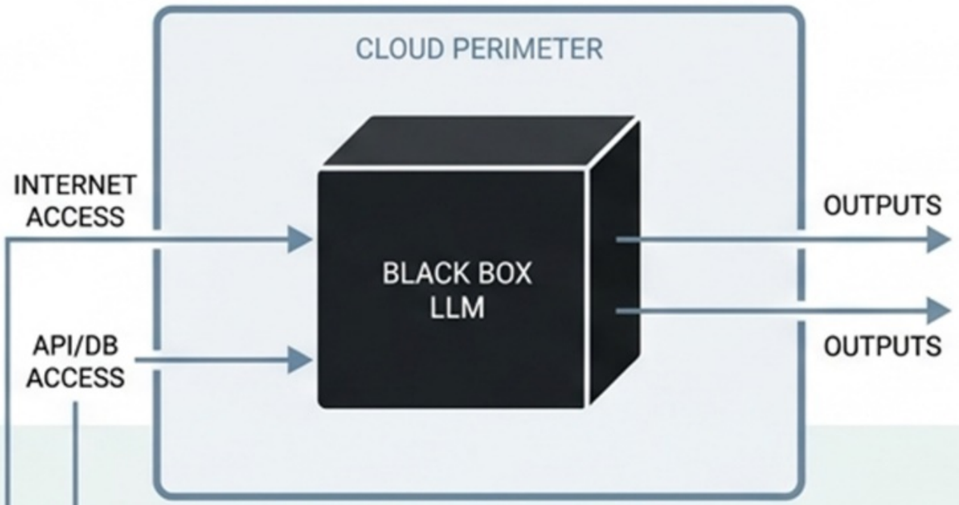
EXECUTIVE THESIS

99.155% Cumulative LLM Security Risk Reduction

α = 0.80 | β = 0.85 | γ = 0.90 | ρ = 0.75 $L \leq 200$ ms

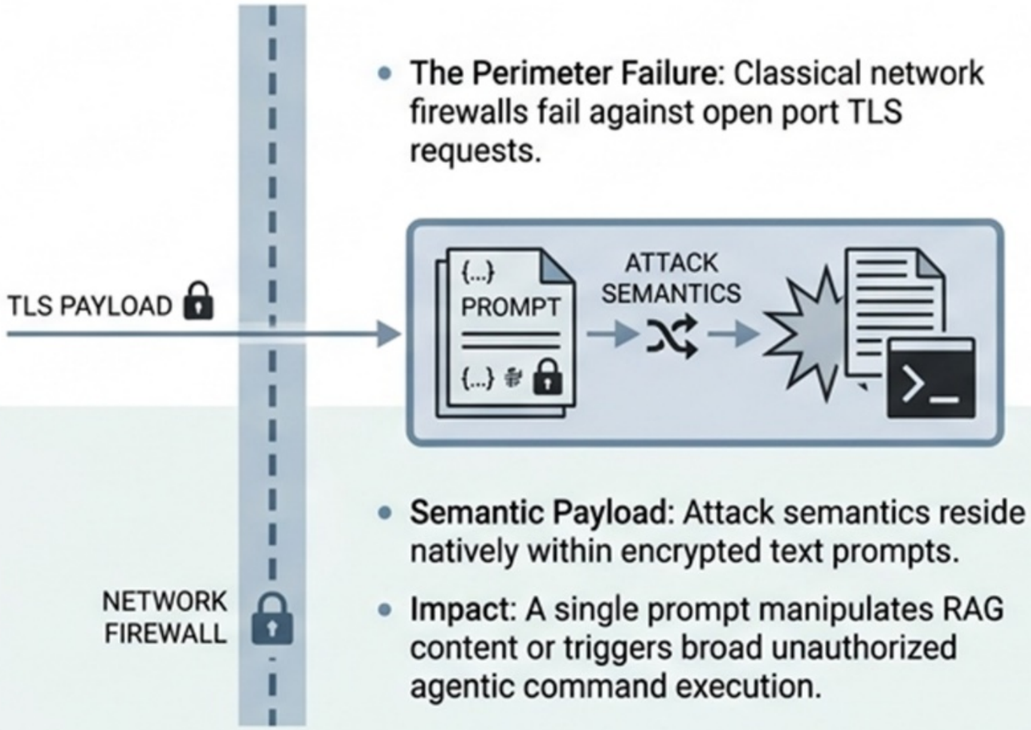
Implementation of a formalized, six-level Zero Trust (ZT) architecture successfully balances threat mitigation with strict system latency constraints.

Computing & Operational Constraints



1. **Deployment Isolation:** LLM services operate as black boxes requiring internet, API, and DB access.
2. **Processing Limits:** High-stakes corporate environments demand latency $L \leq 200$ ms.
3. **Opacity:** Internal decision-making processes cannot be natively debugged via traditional log tracing.

Semantic Cyber Threat Landscape



- **The Perimeter Failure:** Classical network firewalls fail against open port TLS requests.
- **Semantic Payload:** Attack semantics reside natively within encrypted text prompts.
- **Impact:** A single prompt manipulates RAG content or triggers broad unauthorized agentic command execution.

Baseline Limitations & Research Gap

Current State: Reliance on perimeter-only protection and qualitative checklists (e.g., OWASP, BSI/ANSSI).

Limitation: Lack of a unified quantitative model linking threat mechanisms to tunable controls, latency, and cost.

Formal Problem Definition

Objective: Minimize cumulative LLM information security risk via ZT parameter optimization.

Control variables:

$\alpha \in [0, 1]$ (Access policy granularity)

$\beta \in [0, 1]$ (DLP & semantic filter coverage)

$\gamma \in [0, 1]$ (Audit and containment completeness)

Optimization Objective

$$\min_{\alpha,\beta,\gamma} R_{ZT}(\alpha, \beta, \gamma, \rho{=}0.75)$$

Hard System Constraints

Latency: $L(\alpha, \beta, \gamma) \leq 200\,\mathrm{ms}$

Cost: $C(\alpha, \beta, \gamma) \leq 3.5 \times 10^6\,\mathrm{c.u.}$

Architecture Assumptions & Bounds

Entity Sets:

Agents	$A = \{a_1, a_2, \ldots, a_n\}$
Privileges	$P_i = \{p_{1i}, \ldots, p_{ki}\}$
Data Sources	D_i

Principle of Least Privilege

$$P_i^{\text{opt}} = \underset{P_i \subseteq P_{\text{all}}}{\operatorname{argmin}} \left(|P_i| \right)$$

$$\text{subject to } F_i(P_i) \geq F_{\text{required}}$$

Adversary Profile (7-Class Taxonomy)

Telemetry

T1 : Prompt Injection [$\lambda_1 = 2.50$ / hr]

T2 : Data Leakage [$\lambda_2 = 1.80$ / hr]

T3 : Privilege Escalation [$\lambda_3 = 1.20$ / hr]

T4 : Model Poisoning [$\lambda_4 = 0.40$ / hr]

T5 : Jailbreaking [$\lambda_5 = 1.50$ / hr]

T6 : Output Encoding/Obfuscation [$\lambda_6 = 0.80$ / hr]

T7 : Vector-Database Poisoning [$\lambda_7 = 0.35$ / hr]

Governing Risk Equations

Baseline Risk:

$$R_{\text{baseline}} = \sum_{j=1}^{7} \lambda_j \cdot I_j \cdot P_{0j}$$

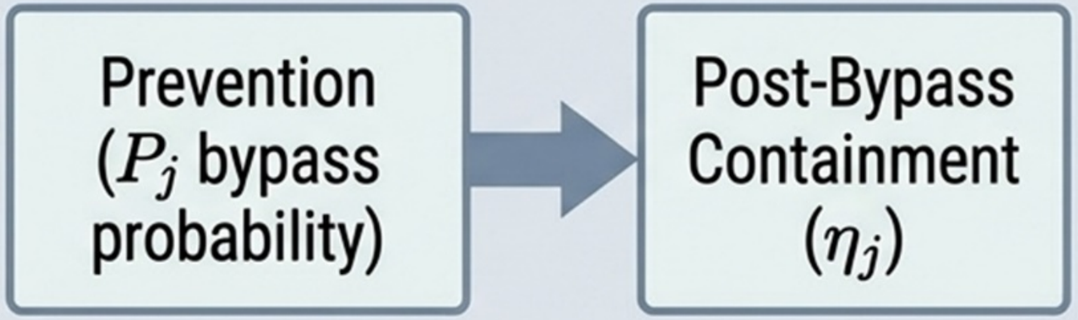
Zero Trust Risk:

$$R_{ZT} = \sum_{j=1}^{7} \lambda_j \cdot I_j \cdot P_{j(\text{success}|\alpha,\beta,\rho)} \cdot (1 - \eta_j(\gamma))$$

Detection Function:

$$\eta_j(\gamma) = 1 - e^{-k4_j\gamma}$$

Theoretical Novelty: Two-Stage Chain



Example (Prompt Injection):

$$P_1 = P_{01} \cdot \exp(-k_1\alpha) \cdot (1 - \beta)^{k_2}$$

Each threat is mathematically linked only to specific controls that directly suppress its mechanism.

Entry & Data Stack (Levels 1-3)

L1 **AI Gateway** | Request authentication, basic size filtering, scenario contextualization.

L2 **Policy & Filter Layer** | Semantic inspection, DLP controls (masking PII/financials). Governed by parameter β.

L3 **Data Contour & RAG** | Segmented indices by security zone. No direct LLM access to production DBs. Governed by parameter ρ.

Identity & Observation Stack (Levels 4-6)

L4 **Identity / IAM** | Contextual service accounts linking user → app → scenario → agent. Governed by parameter α.

L5 **Agent Tools** | API wrappers restricting mass/destructive actions, transforming agents into managed microservices.

L6 **SOC/SIEM Integration** | Centralized anomaly detection and logging. Governed by parameter γ.

Testbed Parameters

Model: Qwen2.5-14B-Instruct (Self-hosted corporate environment)
Exposure Window: 90-day baseline phase vs. 90-day protected phase
Workload: 100,000 completed LLM requests per phase (200,000 total)
Population: 5 fixed LLM agents, 60 authorized users

Validation Metrics

- **Incidence-rate ratios (RR)** with log-scale Poisson 95% CI
- **End-to-end response latency** tracking (ms)
- **Mean Time to Detect (MTTD)** evaluated via Welch two-sample test

System Calibration Variables

Damage Weights (I_j): **100k - 300k** c.u. range (normalized severity)

Baseline Success Rates (P_{0j}): Varying from **0.60** (T6) to **0.85** (T1)

Simulation Rigor: **10,000-run Monte Carlo** parametric analysis (Seed: 20260903)

Modeled Risk Reduction

$R_{baseline}$: **1,132,150.00** c.u./hour

R_{ZT}: **9,571.034** c.u./hour

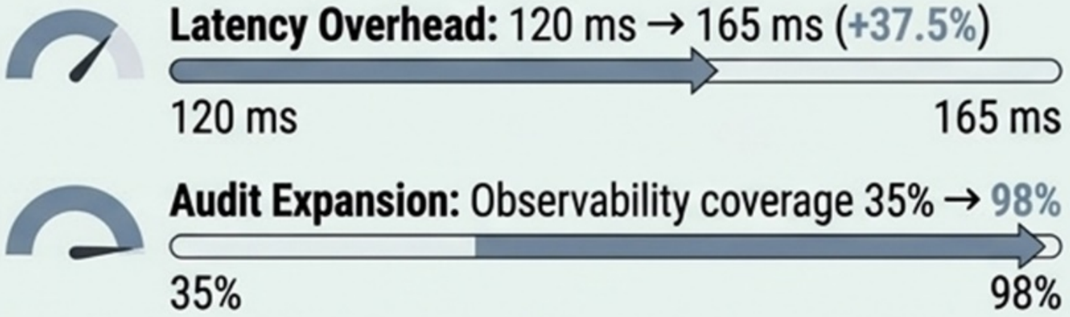
Total Modeled Gain: 99.155% reduction

(95% CI: 98.593% - 99.504% via Monte Carlo)

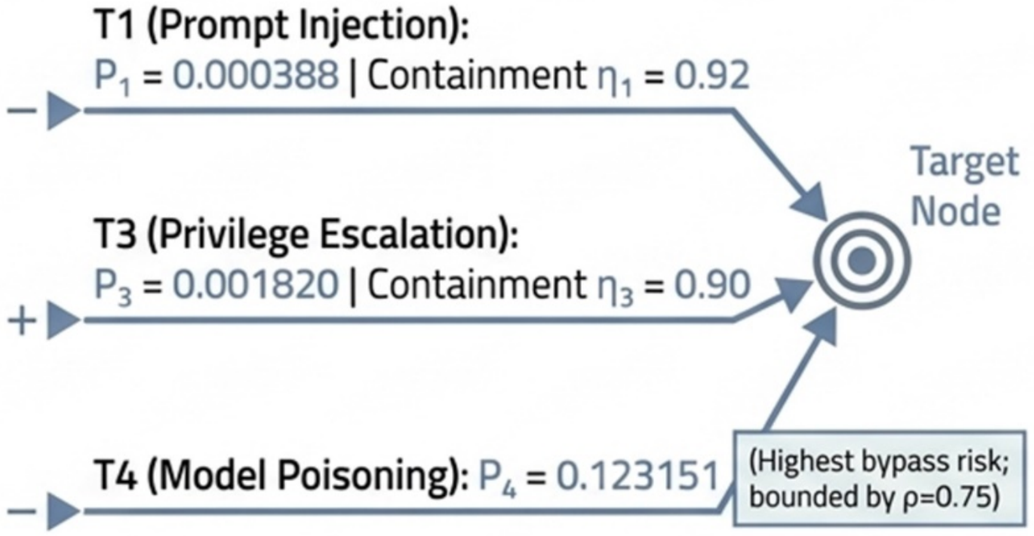
Empirical Field Metrics (90-Day Delta)

Metric	Baseline → Protected	[Delta]
Security Incidents:	72 → 9	[87.5% reduction]
Data-Exposure Events:	45 → 3	[93.3% reduction]
Unauthorized API Calls:	2,568 → 126	[95.1% reduction]
Privilege Violations:	201 → 12	[94.0% reduction]

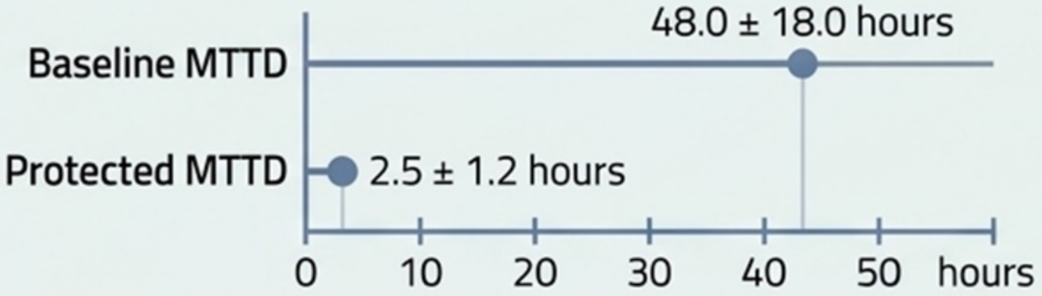
System Trade-Offs



Robustness Under Semantic Attack



Fault Tolerance & Recovery Time



Statistical Significance: yielding $p < 10^{-30}$ with

Welch two-sample test 95% CI of mean difference at 41.20 - 49.80 h.

Framework Benchmarking

Metric / Capability	Explainable ZT [2]	Multi-LLM Agentic [3]	OWASP Top 10 [12]	**Proposed Solution**
Quantitative Risk Model	No	No	No	**Yes**
Parameter Optimization	No	No	No	**Yes**
LLM Threat Coverage	Identity / Agent misuse	Distributed edge trust	10-category catalog	**7 Explicitly Modeled Classes**
Latency/Cost Trade-off	No	No	No	**Yes** (Fitted functions)
Primary Scope	Access explainability	Multi-agent network	Mitigation catalog	**Mathematical & Empirical ZT**

The Engineering Compromise

$$\begin{bmatrix} \alpha = 1.00 \\ \beta = 1.00 \\ \gamma = 1.00 \end{bmatrix}$$

✗ Absolute Risk Minimum: α=1.00, β=1.00, γ=1.00 (R_{ZT} = 3,878.34 c.u./h). (Rejected due to latency/cost).

✓ Deployed Compromise Point: α=0.80, β=0.85, γ=0.90.

Why?

Engineering selection utilized min-max normalization via compromise-programming distance ($w_R = 0.60$, $w_C = 0.40$).

Hard operational limits strictly honored: $L \leq 200$ ms, $C \leq 3.5 \times 10^6$ c.u.

Operational & Integration Constraints

- Provenance Cap: ρ fixed at 0.75 due to absence of independent latency/cost coefficients for provenance controls.
- Latency Optimizations (Unmeasured): Future overhead reductions require asynchronous SOC/SIEM export, semantic caching, and hardware-accelerated DLP pattern matching.

Three Proven Findings

1. A parameterized ZT architecture minimizes excessive privileges without degrading baseline business functionality.
2. Decoupling bypass probability (P_j) from detection efficacy (η_j) yields highly accurate predictive SOC models.
3. Risk optimization ensures LLM semantic security is achievable well within the 200 ms latency tolerance.

Future Research Roadmap

- Evaluate dynamic α, β, γ, ρ policies for multimodal architectures.
- Execute multi-site randomized tests to resolve baseline non-randomized limitations.
- Calibrate precise cost and latency functions specific to RAG/Model provenance controls.

SpacSec 2026 Session Q&A

System telemetry and Monte Carlo configurations (Seed 20260903) available per corporate disclosure guidelines.