SpacSec 2026

International Conference on Cyber Warfare,
Security & Space Computing

ITMO UNIVERSITY

Faculty of Secure IT

A Decentralized Semantic Attestation Method for Protecting Multi-Agent LLM Systems from Cascading Adversarial Injections

Sergey I. Serko

Master’s student · ITMO University

Co-author: Andrei V. Petrovski

Saint Petersburg, Russia

Saint Petersburg

26–28 October 2026

Cybersecurity · AI/ML

Zero-Trust Multi-Agent Protocol (ZTP)

Aim, focus and objectives

AIM	**Evaluate local semantic attestation as an early-containment mechanism for cascading indirect prompt injections.**
FOCUS	Semantic consistency of tool responses and inter-agent messages at each receiving-agent boundary.

OBJECTIVES

- **O1** Define local zero-trust verification at every receiving-agent boundary.
- **O2** Implement centroid-based and trajectory-aware semantic scoring.
- **O3** Quantify detection performance and false-positive limitations on InjecAgent.
- **O4** Compare simulated injection propagation under three defense conditions.

Scope: proposed architecture + semantic detector + containment simulation; no live multi-agent deployment was evaluated.

Research questions and hypotheses

RQ1	**How can semantic attestation be performed locally at an agent boundary?**
RQ2	**Does centroid drift discriminate injected from benign tool responses?**
RQ3	**Does trajectory-aware scoring reduce simulated propagation beyond centroid-only scoring?**

H1 Injected responses have higher centroid drift than benign responses.

H2 Adding trajectory drift reduces propagation beyond threshold-only scoring.

Working hypotheses derived from the paper's logic; results are interpreted conservatively.

Problem: cascading indirect prompt injection



Why the risk grows across agents

LLMs process natural-language data and instructions in the same context.

A transformed malicious instruction may appear to come from a trusted internal component.

Without re-verification, compromise can propagate to later agents and tools.

Security objective: re-establish trust locally before an incoming message can affect the next agent step.

Proposed method: local semantic attestation

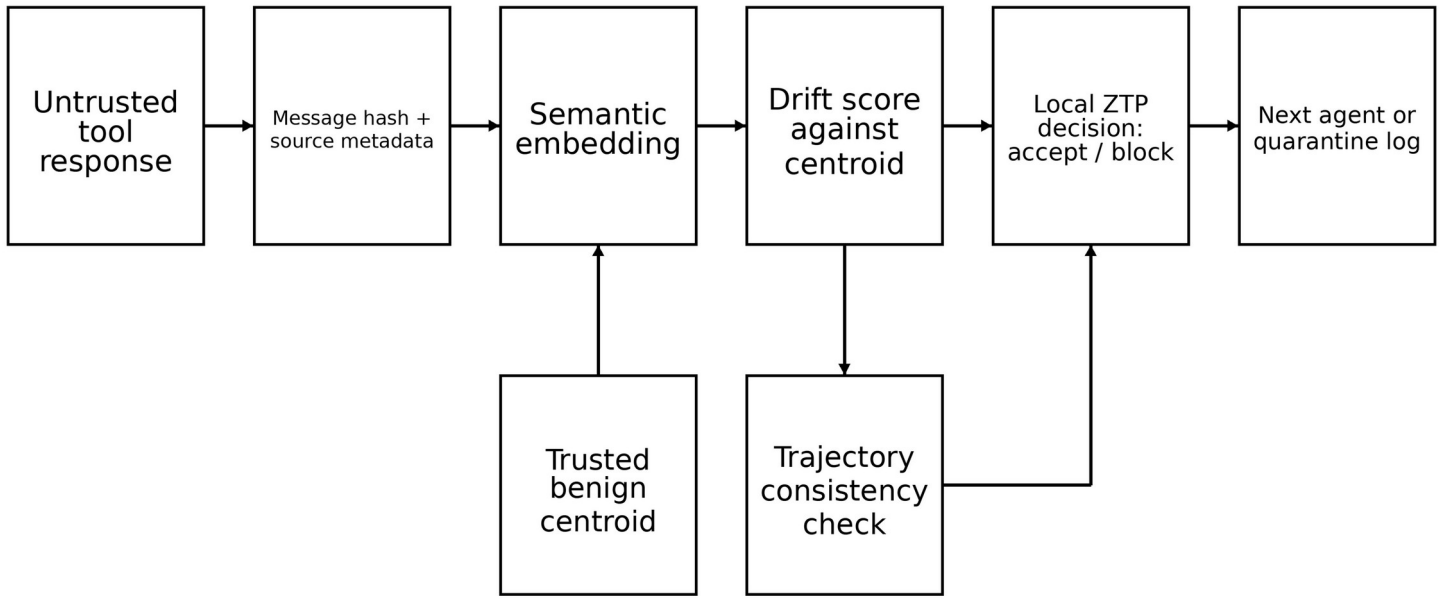


Figure 1. Decentralized ZTP semantic attestation at an agent boundary.

Zero-trust principle

Trust is re-established at every receiving agent boundary.

Evidence combined

- Source metadata and message hash
- Semantic drift from a trusted centroid
- Optional trajectory-consistency signal

Local action

Accept · Block · Quarantine · Escalate

Experimentally evaluated: semantic scoring and simulated accept/block decisions — not the complete operational protocol.

Data and evaluation design

InjecAgent base setting

Published indirect prompt-injection benchmark for tool-integrated agents

510	544	1,054	17
Direct Harm	Data Stealing	Reconstructed benign observations	Unique benign templates

Evaluation split

527 benign · calibration	527 benign · evaluation	1,054 injected · evaluation

1,581 evaluation observations = 527 benign + 1,054 injected

Implementation

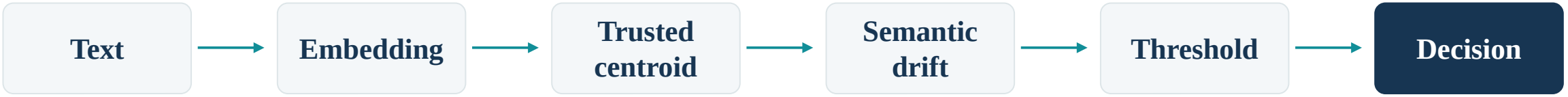
all-MiniLM-L6-v2 · 384 dimensions · no fine-tuning

Validity boundary

The same 17 templates occur in calibration and evaluation.

Prepared benchmark scenarios — not logs of real-world compromises.

Semantic scoring and decision rule



Individual-response detector

$$d_{(c)}(m) = 1 - \cos(e(m), c)$$

Centroid drift is an anomaly score — not an attack probability.

Operational threshold

$$\tau = 0.23$$

Block when d > τ (strict inequality)

Trajectory-aware chain score

$$d_ZTP = \max(d_{(c)}, 0.6 \cdot d_{(t)})$$

$d_{(t)}$ compares consecutive message embeddings.

Methodological caveat

The threshold was selected on the same evaluation split used to report threshold-dependent metrics.

Detection performance and false positives

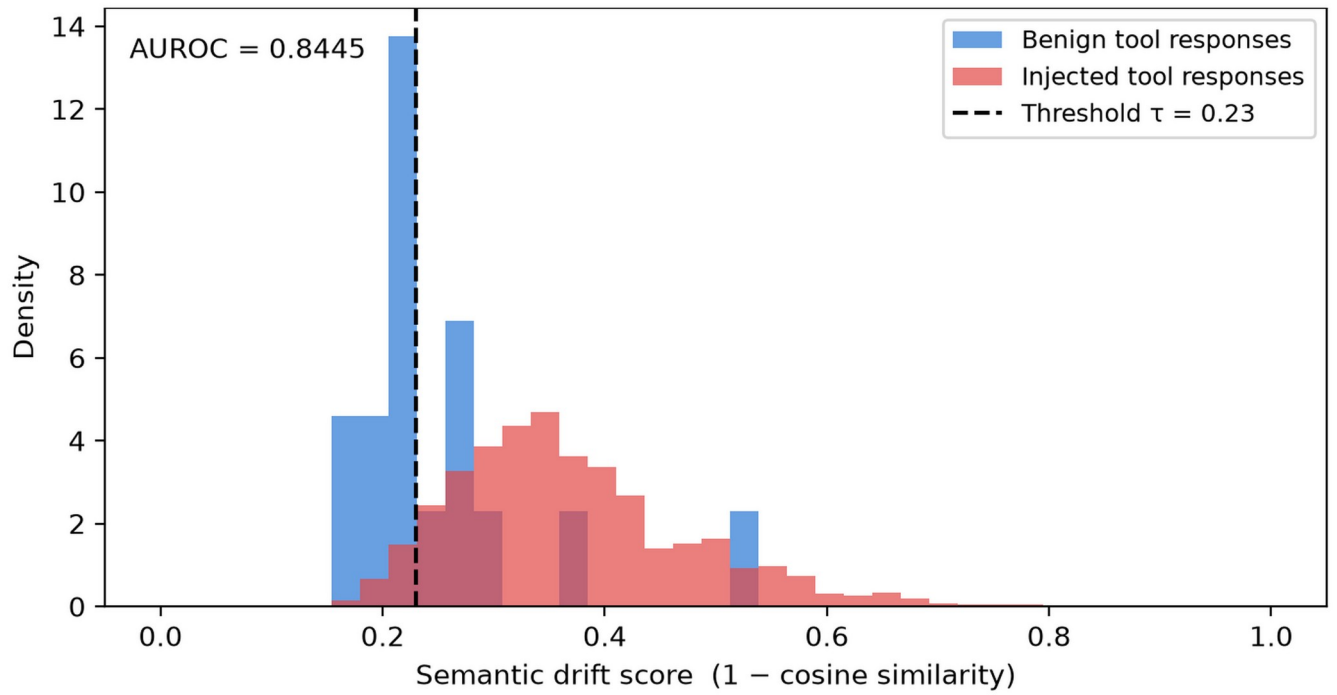


Figure 2. Drift-score distributions for benign and injected tool responses.

AUROC

0.8445

Threshold-independent ranking

F1-score

0.8777

At τ = 0.23

Recall

94.31%

994 of 1,054 injections detected

False-positive rate

41.18%

217 of 527 benign responses blocked

Confusion matrix

TP 994 · FN 60
FP 217 · TN 310

The detector identifies most injections, but the observed FPR is too high for universal hard blocking.

Semantic space and the limits of one global centroid

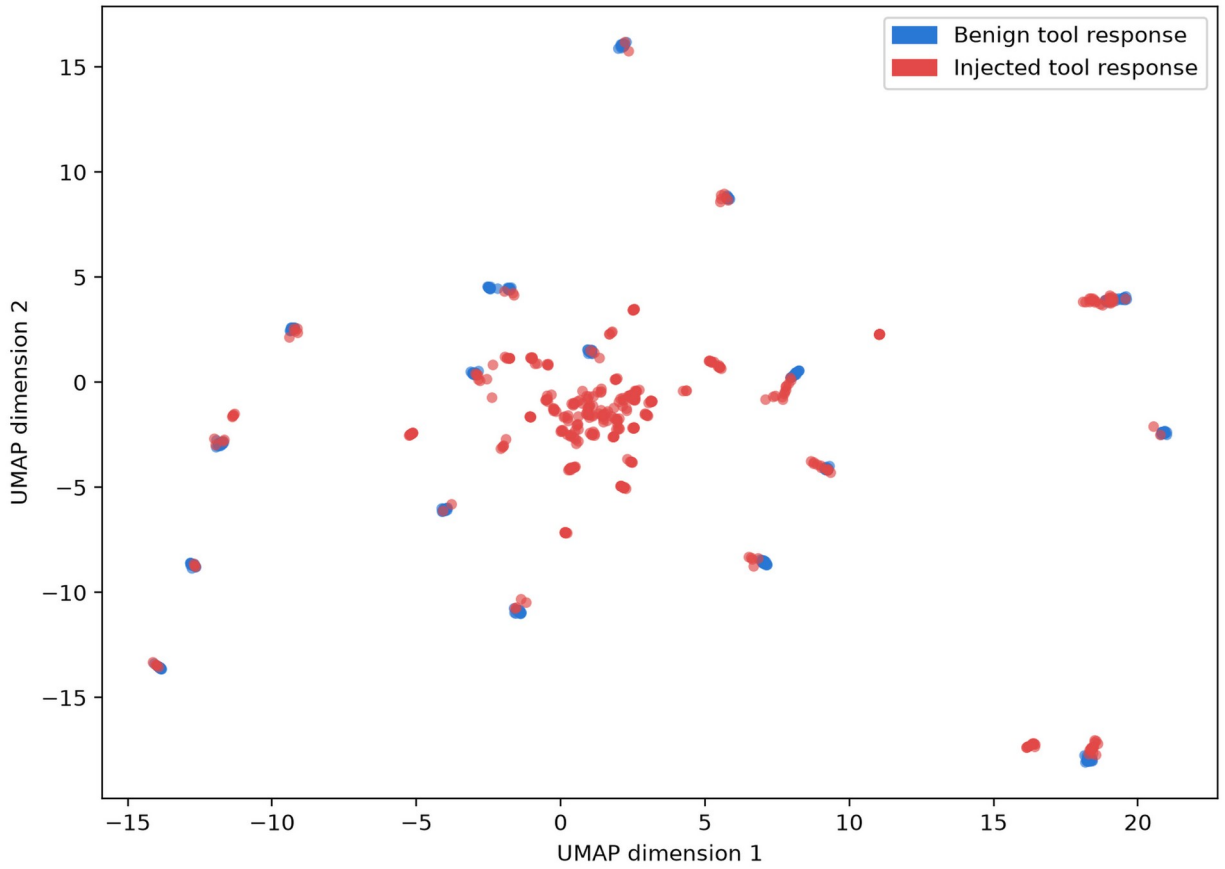


Figure 3. UMAP projection of benign and injected tool-response embeddings.

400 + 400

Benign and injected observations shown

17 benign templates

- Benign and injected points overlap substantially.
- Tool-response templates remain semantically dominant after injection.
- Different tools occupy different regions; one centroid is too coarse.

UMAP is a 2D view, not proof of class separability in 384 dimensions.

Simulated injection propagation across agent chains

80 chains per defense condition

No defense	Threshold-only	Trajectory-aware ZTP
100%	5.0%	3.75%
80 / 80	4 / 80	3 / 80

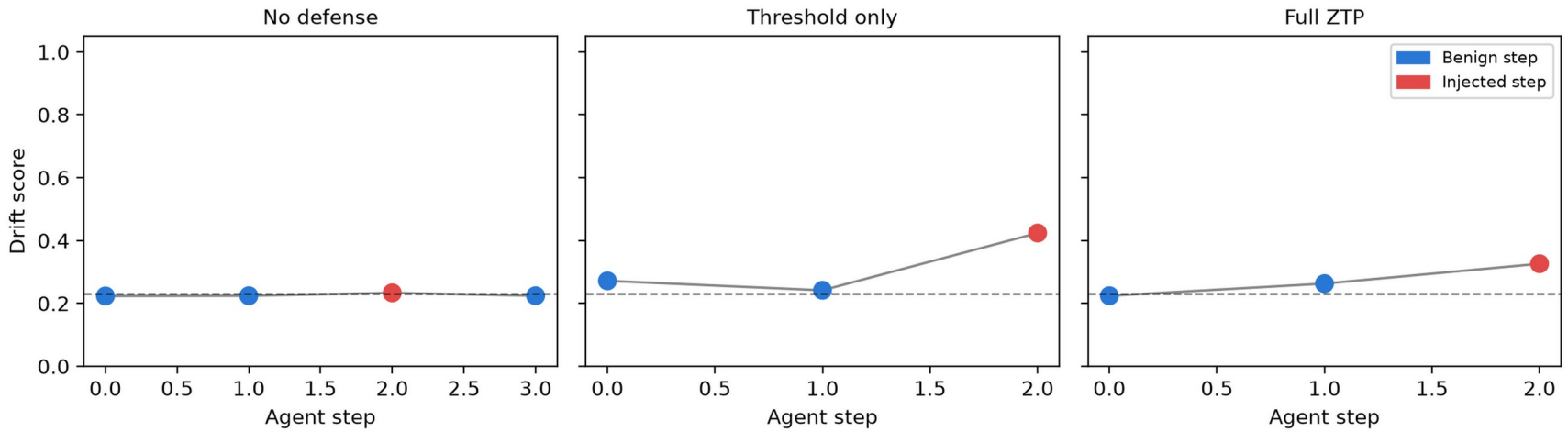


Figure 4. Representative simulated chain trajectories under the three defense conditions.

Containment simulation over benchmark text — not end-to-end attack success. The 5.0% vs 3.75% gap is one case.

Objectives and outcomes: what was solved, and with what evidence

Objective	Delivered by the study	Evidence boundary
O1	Local semantic-attestation architecture and decision flow at receiving-agent boundaries.	Not deployed or stress-tested in a live multi-agent system.
O2	Reproducible centroid and trajectory scoring with a transparent decision rule.	Provenance, quarantine and human escalation were conceptual controls.
O3	AUROC 0.8445; F1 0.8777; Recall 94.31%; template-level false-positive analysis.	FPR 41.18%; threshold fitted on evaluation data; only 17 benign templates.
O4	Aggregate propagation compared: 100% / 5.0% / 3.75%.	Different simulated chains; one-case gap; no live agent attack measurement.

H1 · Descriptively supported on this benchmark

H2 · Additional trajectory benefit not established

Next: tool-specific profiles · richer benign data · separate validation/test · live-agent evaluation