SpacSec 2026 | Technical Diagnostic Report

Deterministic Defense Methods Against Probabilistic Attacks

A Study of the Ineffectiveness of Traditional Approaches to Securing Large Language Models

Authors: M.V. Mitrofanov, N.N. Kramskoy, A.S. Kurakin
Affiliations: ITMO University & Special Technology Center

EXECUTIVE THESIS

Traditional deterministic cybersecurity methods demonstrate critically low efficiency against probabilistic transformer architectures.

Prompt injection is an inherent architectural property, not a software bug.

CRITICAL SYSTEMS CONSTRAINTS



THE ALIGNMENT TAX:

- **Parameter:** Capital Expenditure (Models costing hundreds of millions of dollars).
- **Constraint:** Reinforcement Learning from Human Feedback (RLHF) forces a strict compromise between utility and safety.
- **Failure Mode:** Strengthening syntactic restrictions inevitably reduces operational functionality.

CYBER THREAT LANDSCAPE

THREAT TELEMETRY:

- **Global Rank:** #1 in OWASP Top 10 for LLM Applications (2025).
- **Attack Vectors:** Non-standard encodings, multilingual injections, and semantic obfuscation.
- **Diagnostic Assessment:** Each wave of deterministic patches generates a new class of bypass attacks exploiting unexplored areas of the mathematical embedding space.

THE INDUSTRY FALLACY

- **Flawed Assumption:** Prompt-injection vulnerabilities are correctable implementation defects.
- **Current Approach:** Applying software patches and syntactic filters to fix probabilistic behavioral deviations.

FORMAL PROBLEM DEFINITION

- **Mathematical Root:** The vulnerability is embedded in the foundational equations of the architecture, not the codebase.
- **Optimization Objective:** Quantifying the degradation of deterministic defenses as a continus function of attack distribution shift.

$$D_{KL}(D_{attack} \parallel D_{train}) \to \infty$$

- **Key Variable:** Kullback-Leibler divergence (D_{KL}) between attacking prompts and the training distribution.

ARCHITECTURE BOUNDS

Parameter: Value lockups:

Model Class: Decoder-only transformer

Scale: 7-13B parameter class

Layers (L): 32

Hidden Dim (d_model): 4096

Attention Heads (n_head): 32

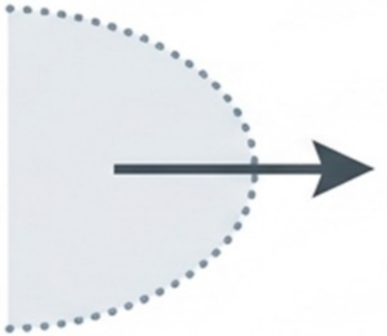
Context Window (L_max): 4096 tokens

Safety Vector: v_safety in R^4096

ADVERSARIAL PROFILE

- **Representation Engineering:** Direct intervention in latent space activations.
- **Crescendo Attacks:** Multi-turn context drift exploiting isolated per-turn checks.
- **OOD Masking:** Tokenization manipulation bypassing deterministic surface filters.

RLHF OPERATIONAL REALITY

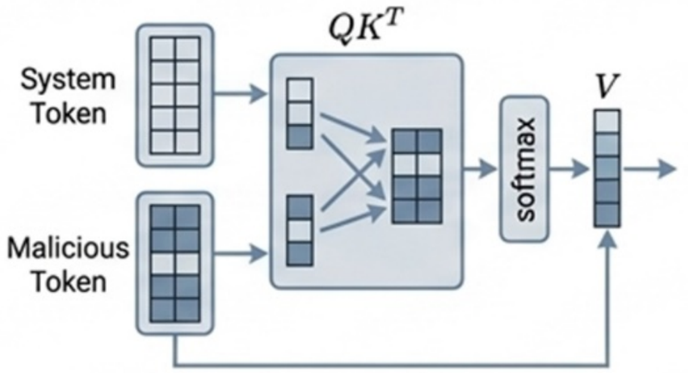


Mechanism: Forms probabilistic vector directions indicating preferred refusal behavior.

Vulnerability: Provides soft statistical boundaries, completely lacking hard architectural access control.

SELF-ATTENTION BUG

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



Zero variables exist for distinguishing token sources. System instructions and user tokens mix identically.

ACTIVATION STEERING

$$h_l^{\text{modified}} = h_l^{\text{original}} - \alpha\, v_{\text{safety}}$$

RLHF safety vectors are not immutable constants. They can be geometrically attacked and subtracted in the embedding space.

CONTEXT DRIFT (CDI)

$$\text{CDI} = \frac{1}{T}\sum_{t=1}^{T} \cos(h_t, v_{\text{malicious}})$$

Tracks the monotonic accumulation of malicious vectors across dialogue turns, completely bypassing isolated per-turn checks.

HARDWARE & SOFTWARE STACK

[COMPUTE_NODE]: Single NVIDIA A100 (80 GB GPU)

[FRAMEWORK_01]: Hugging Face Transformers 4.41

[FRAMEWORK_02]: PyTorch 2.2

[FRAMEWORK_03]: vLLM 0.4

[TARGET_SYSTEM]: Llama-2-7B-Chat

DATA PROCESSING PIPELINE



2,400 prompts (AdvBench, JailbreakBench, synthetic OOD)

5 D_{KL} bins (0 to 5.0) | 480 prompts per bin

Deterministic greedy decoding | temp = 0.0 | max_tokens = 256

ATTACK SUCCESS & FILTER EFFICIENCY

$$ASR = \frac{N_{success}}{N_{total}} * 100\%$$

$$ea_{det} = 1 - \frac{ASR_{filtered}}{ASR_{unfiltered}}$$

Prediction: $ea_{det} \to 0$ with adaptive OOD attacks.

SAFETY VECTOR DEGRADATION

$$\Delta_{safety} = \left\| \mathbf{v}_{safety}^{post} - \mathbf{v}_{safety}^{init} \right\|^2$$

Critical Threshold:
Degradation $> 0.5 * \left\| \mathbf{v}_{safety}^{init} \right\|^2$

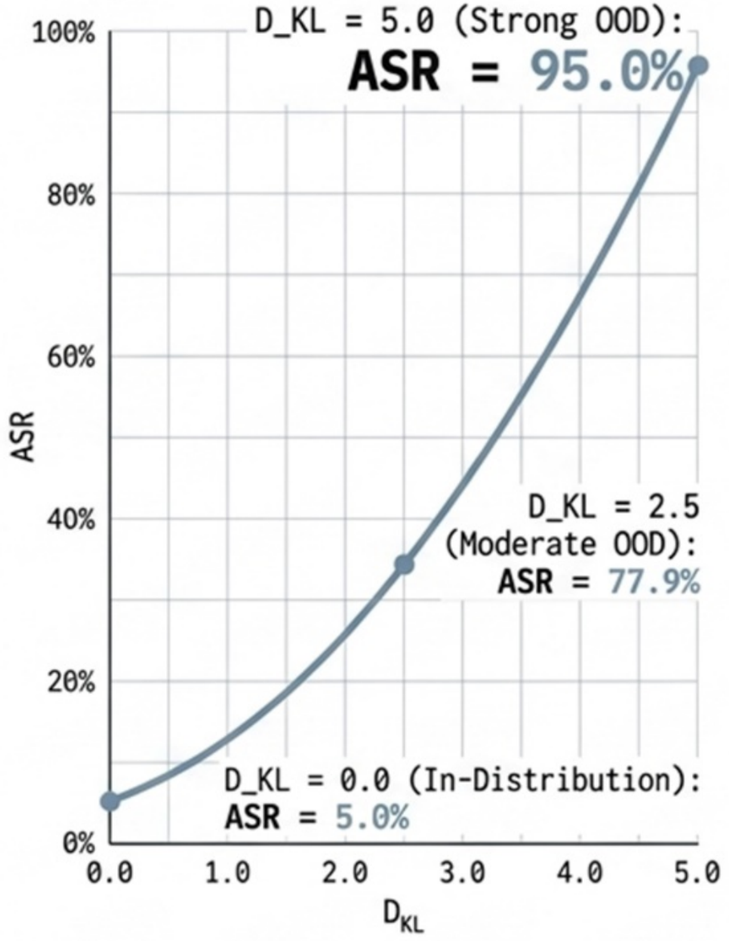
Interpretation: Indicates loss of >50% of the model's original protective capability.

AUTOREGRESSIVE INERTIA (LOCK-IN)

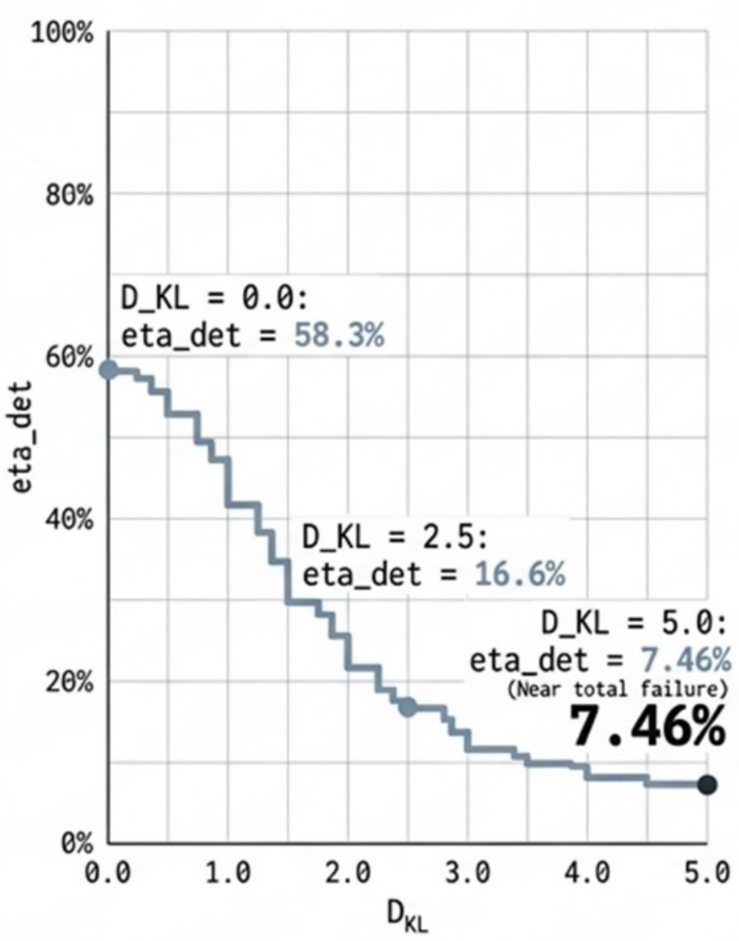
$$\Delta PPL > \tau$$

Once a model generates the first token of agreement (e.g., 'of course'), changing the trajectory causes a perplexity jump (Δ PPL), physically forcing the autoregressive engine to complete the malicious output.

EXPONENTIAL ASR GROWTH



FILTER EFFICIENCY DECAY



THE SCALABILITY PARADOX

DEFENSE SCALING:

O(n) linear complexity

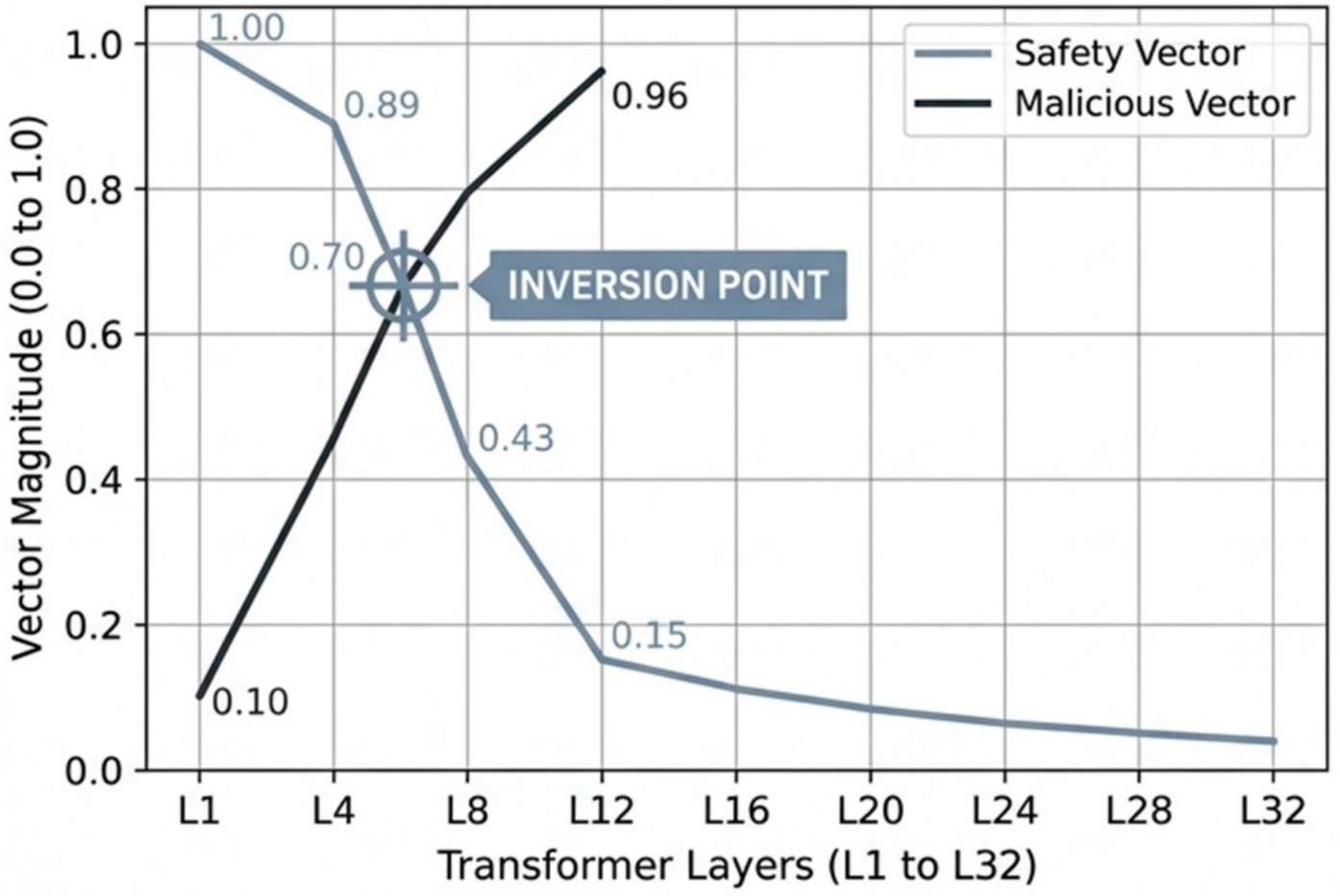
2,847 regex rules required to block just 1,000 known attacks.

ATTACK SCALING:

Combinatorial explosion

15,000+ variants generated instantly by combining basic languages and encodings.

LAYER-BY-LAYER DECAY (ACTIVATION STEERING)



DIAGNOSTIC READOUT

Event:
Safety vector crossover

Condition:
$v_{safety} = v_{malicious}$

Location:
Estimated layer $\ell \approx 6.4$

Relative Depth:
$l/L \approx 0.20$

CRITICAL FAILURE:
At just 20% network depth, 57% of protection is already lost. Deterministic surface filters cannot detect this latent-space manipulation.

COMPARATIVE EVALUATION: ARCHITECTURAL VULNERABILITIES

Vulnerability	Architectural Cause	Deterministic Defense	Defense Efficiency
Confused Deputy	Self-attention matrix does not distinguish token sources	Prefix/suffix separators	**Low Efficiency (bypass via injection)**
Lost in the Middle	Quadratic complexity O(n^2) of attention	Context length limitation	**Negative Efficiency** (-6.25% under test)
Activation Steering	Vector nature of latent space representations	Defenses Unavailable	**Absent Efficiency**
Crescendo Attack	Absence of global intent tracking across context	Isolated per-turn checking	**7.0% Efficiency** (CDI monotonically increases)
Tokenization Mismatch	Gap between token space and semantic space	Regex surface filters at input	**Low Efficiency (unbounded encoding variants)**

BOUNDARIES & GENERALIZABILITY

THE INVERSION POINT

- The specific intersection layer ($\ell = 6.4$) is bounded to the 32-layer configuration tested. However, the relative depth fraction ($\ell/L \approx 0.2 - 0.3$) is mathematically hypothesized to generalize across larger architectures.

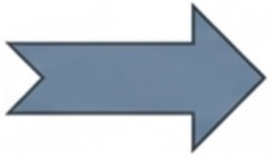
THE UTILITY TRADE-OFF

- Formal methods are required to determine Pareto boundaries. Beyond this limit, strengthening deterministic safety permanently degrades the system's core utility.

PRACTICAL FEASIBILITY & SHIFT

FAILURE OF SYNTACTIC FILTERS

- System protection requires transitioning away from surface-level textual blocks.



THE LATENT SHIFT

- Emerging paradigms (e.g., PICO) introduce separate, structurally constrained processing processing channels for system and user tokens, operating natively in the model's semantic representation.

THREE PROVEN DIAGNOSTICS

1. Deterministic filters degrade to a critically low 7.46% efficiency against strongly out-of-distribution (D_{KL}=5.0) attacks.
2. Prompt injection is a fundamental mathematical property (the Confused Deputy problem in self-attention matrices), not a patchable defect.
3. RLHF soft boundaries are geometrically reversible, losing 57% of protection by layer 8.

FUTURE RESEARCH ROADMAP

- Shift from syntactic input/output filters to representation-level defenses (e.g., KL-Divergence Optimizer).
- Develop hybrid architectures combining end-to-end differentiability with strict structural constraints.

DIAGNOSTIC CONCLUDED.

M.V. Mitrofanov, N.N. Kramskoy, A.S. Kurakin | ITMO University & Special Technology Center