ICLR 2026 · Blog Track

The Adversarial Conditioning Paradox

How Fine-Tuning Creates a Geometric Signature That Attacks Unknowingly Exploit

Khazretgali Sapenov University of Phoenix

Aidos Sapenov University of Toronto

The Problem

Adversarial attacks fool classifiers

They optimize against softmax output — no knowledge of internal geometry.

Yet they produce a geometric fingerprint

Attacked inputs have anomalously high Jacobian condition numbers (κ) at Layer 12 of fine-tuned BERT.

This signal is:

- **Absent at L1–L9** Only the final layer shows signal
- **Absent in base model** No fine-tuning → no signal
- **Attack-agnostic** 3 different attack families, same pattern

What is κ?

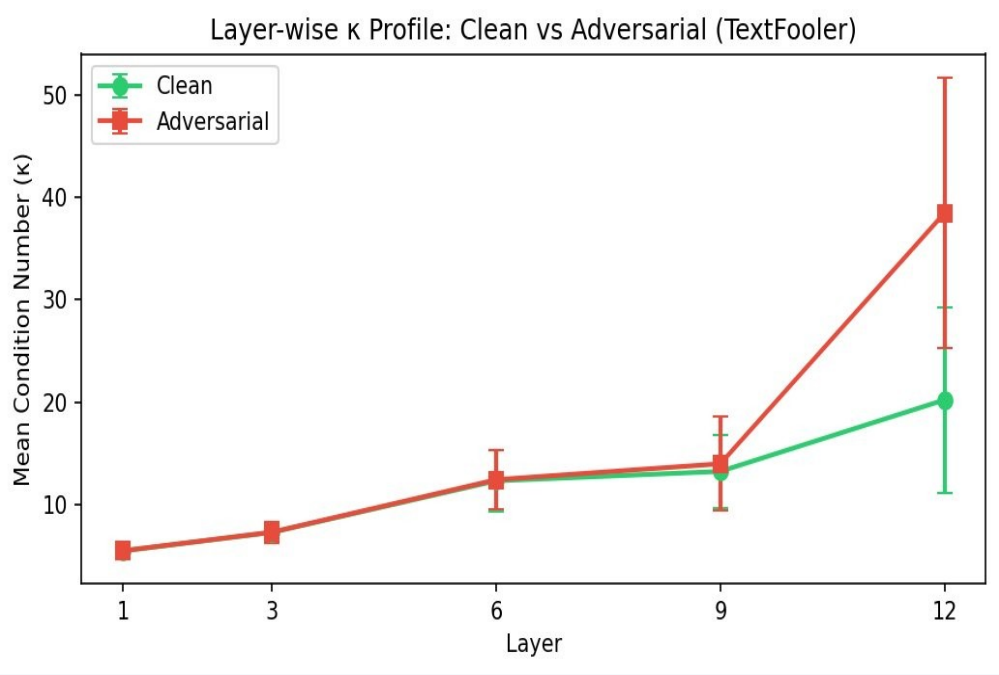
$\kappa = \sigma_{max} / \sigma_{low}$

The condition number of the layer-wise Jacobian — measures how uniformly the model responds to perturbations.

High κ = ill-conditioned: some directions amplified far more than others

Low κ = well-conditioned: all directions treated uniformly

Layer-wise κ Profile: Signal at Layer 12



AUC = 0.907

Layer 12 κ (TextFooler)

Cohen's d = 1.612

Clean κ: 20.2 ± 9.1

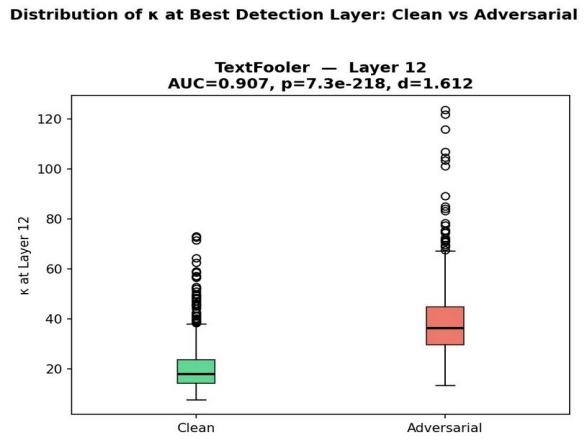
Adversarial κ: 38.5 ± 13.2

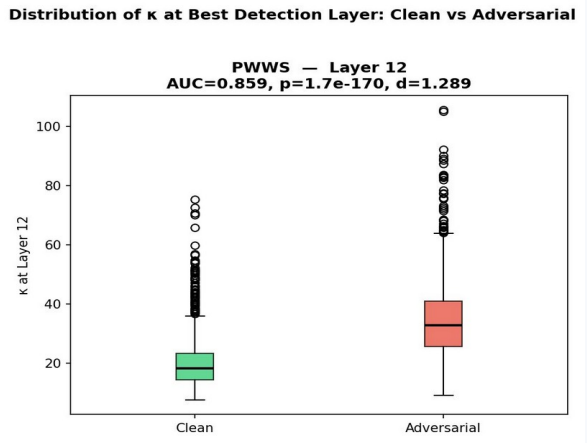
$p \ll 10^{-100}$

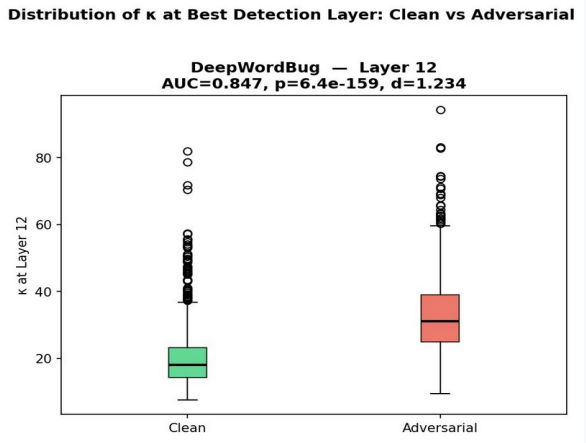
Cosine baseline: AUC = 0.464 (random)

Layers 1–9: no signal (AUC ≈ 0.5). Layer 12: sharp divergence. This is where fine-tuning imposed the classification boundary.

Three Attacks, One Geometric Signature







AUC = 0.907
d = 1.612 N = 1,000

AUC = 0.859
d = 1.289 N = 1,000

AUC = 0.847
d = 1.234 N = 1,000

Word-level semantic, word-level dictionary, and character-level attacks all produce the same Layer 12 κ elevation — because all must cross the same decision boundary.

Fine-Tuning Creates the Signal

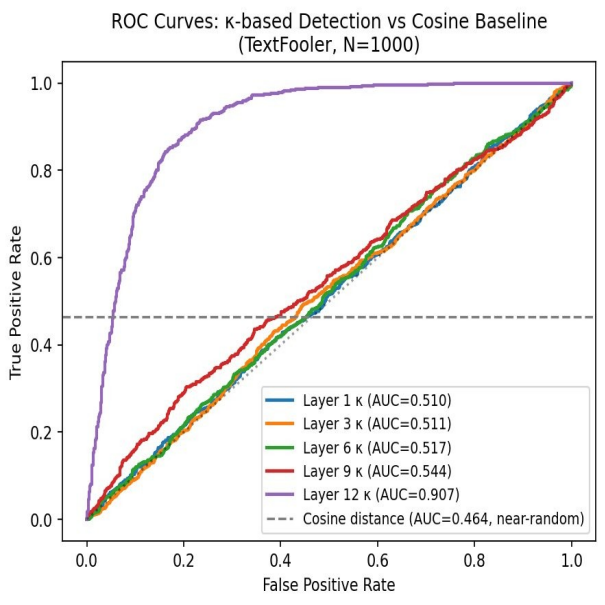
Same inputs, same architecture — only difference is whether the model was fine-tuned on SST-2.

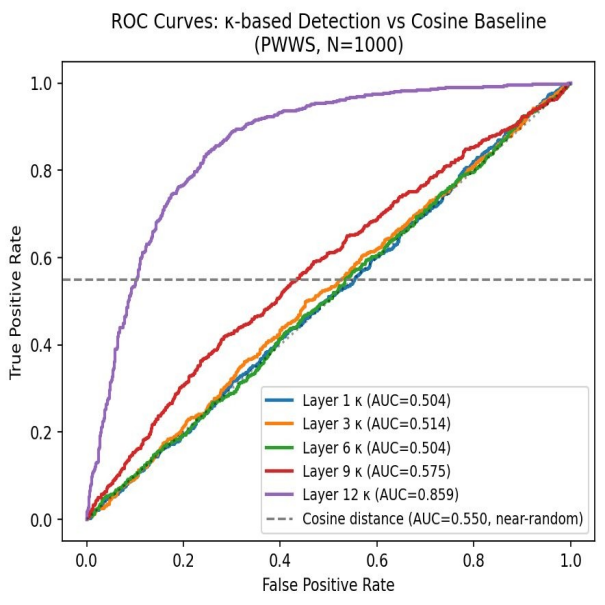
Attack	Fine-tuned L12 AUC	Base model L12 AUC	Drop
TextFooler	**0.907**	0.517	**−0.390**
PWWS	**0.859**	0.502	**−0.357**
DeepWordBug	**0.847**	0.545	**−0.302**

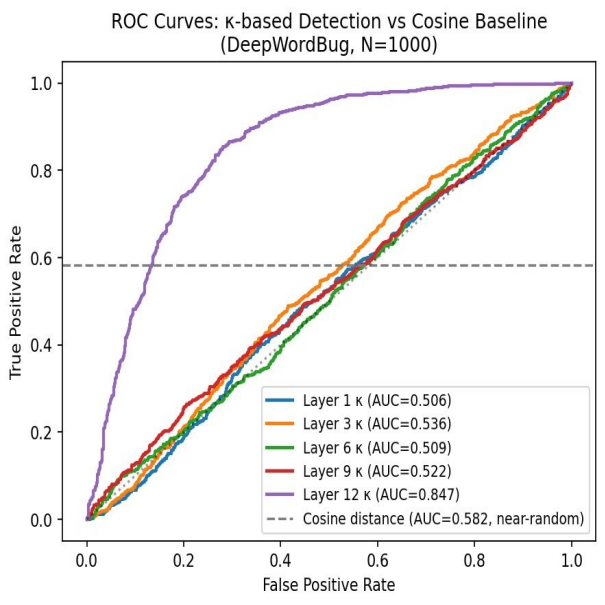
Layer	TF (fine-tuned)	TF (base)	PWWS (ft)	PWWS (base)	DWB (ft)	DWB (base)
L6	0.517	0.529	0.504	0.538	0.509	0.556
L9	0.544	0.520	0.575	0.520	0.522	0.546
L12	**0.907**	0.517	**0.859**	0.502	**0.847**	0.545

L1–L9: all near-random (AUC 0.50–0.58) for both models. L12: fine-tuned AUC jumps to 0.85–0.91 while base stays at 0.50. Per-pair Δκ correlation r = 0.024 (p = 0.44). The signal is created by training, not by the attack.

Detection: κ at Layer 12 vs Cosine Distance







Cosine distance between CLS representations is near-random (AUC 0.46–0.58). **Jacobian geometry at Layer 12 provides strong, attack-agnostic detection (AUC 0.85–0.91).**

Summary

Geometric signature emerges from fine-tuning

κ at Layer 12 reliably separates clean and adversarial inputs (AUC 0.85–0.91) across three attack families. The signal is absent at earlier layers and absent in the base model.

Attacks exploit this geometry unknowingly

Adversarial attacks are optimized against softmax — they have no knowledge of the Jacobian. Yet they land in ill-conditioned regions because crossing the classification boundary is unavoidable.

Practical implications

Monitor Layer 12 κ as a runtime signal. It requires no knowledge of the attack type, is orthogonal to softmax confidence, and can serve as a CI regression test for model updates.

? **Open question:** Can attacks be modified to avoid high-κ regions at Layer 12 while still crossing the decision boundary? This may be geometrically incompatible.

github.com/sapenov/adversarial-conditioning-paradox