POSTER

#Safety Alignment #Adversarial Robustness
#Representation Engineering #Jailbreak Transfer

Paper

LinkedIn

TECHNION Israel Institute of Technology

AI ATLAS LAB

"Shared Geometry ⟺ Shared Jailbreaks"

AnchorRep: Defending LLMs Against Cross-Model Adversarial Transfer via Representation Repulsion

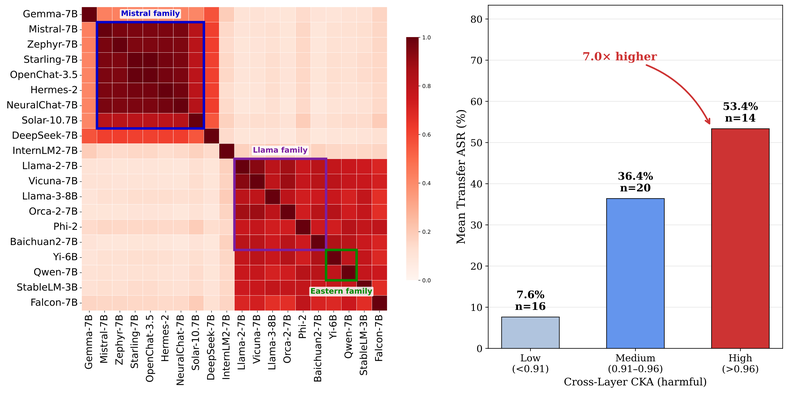
Gal Wertheizer* Rom Himelstein Tomer Peretz Avi Mendelson – Technion, Israel Institute of Technology

The Problem: Transfer Through Shared Geometry

An attack built on one **open-weight** model jailbreaks architecturally different ones. Existing defenses target same-model attacks, not this cross-model channel.

Transfer Tracks Representation Geometry

Across 20 LLMs, higher within-family CKA predicts higher GCG transfer ASR. Shared mid-layer geometry is the mediator, and the target.



Left: similarity across 20 LLMs, three families highlighted. Right: within families, higher CKA means higher transfer ASR.

Similarity metric	Same-family ρ	Cross-family ρ
CKA (harmful, mid-layer)	+0.91**	−0.34*
Variance Explained	+0.72**	−0.51**
RSA (harmful)	+0.74**	−0.69**

*Spearman ρ of similarity vs. transfer ASR. **p<0.01, *p<0.05.*

The Insight → The Defense

Independently trained models learn the **same internal geometry** (*Platonic Representation Hypothesis*). As the data above shows, **shared geometry means shared jailbreaks**.

So we break it. A **LoRA** adapter trains the defender to **minimize CKA** to a frozen **cross-family anchor**.
No adversarial examples.

AnchorRep: Repel The Geometry

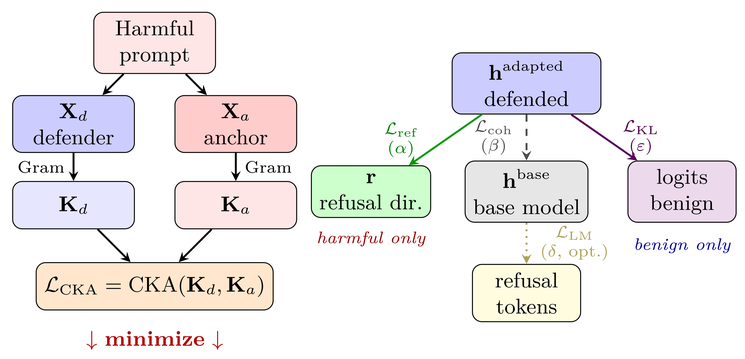
Why CKA

CKA compares *relational* structure, not raw activations: dimension-agnostic (aligns different sizes) and **differentiable** (a trainable loss). Minimizing it repels the defender's harmful geometry from the anchor.

$$\mathrm{CKA}(\mathbf{X}_1, \mathbf{X}_2) = \frac{\mathrm{HSIC}(\tilde{\mathbf{K}}_1, \tilde{\mathbf{K}}_2)}{\sqrt{\mathrm{HSIC}(\tilde{\mathbf{K}}_1, \tilde{\mathbf{K}}_1)\,\mathrm{HSIC}(\tilde{\mathbf{K}}_2, \tilde{\mathbf{K}}_2)}}, \qquad \mathrm{HSIC}(\mathbf{X}, \mathbf{Y}) = \mathrm{tr}(\mathbf{X}^\top \mathbf{Y})$$

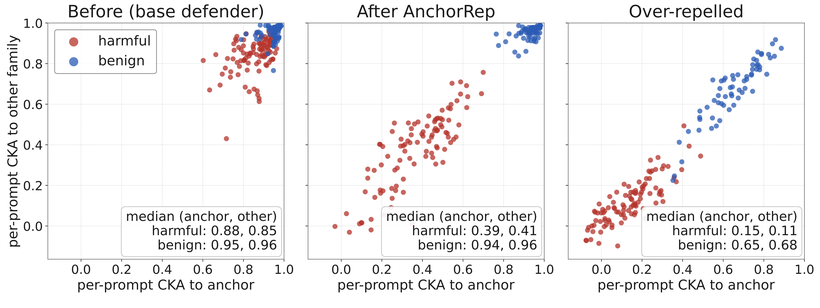
$\tilde{\mathbf{K}}_i$*: centered Gram matrices; HSIC; tr: trace.*

(a) CKA Repulsion **(b) Auxiliary Losses**



(a) CKA repulsion between defender and anchor on harmful prompts suppresses transfer. (b) Four auxiliary losses preserve utility.

$$\mathcal{L} = \gamma\mathcal{L}_{\mathrm{CKA}} + \alpha\mathcal{L}_{\mathrm{refusal}} + \beta\mathcal{L}_{\mathrm{coherency}} + \varepsilon\mathcal{L}_{\mathrm{KL}} + \delta\mathcal{L}_{\mathrm{LM}}$$



Results

		Transfer ASR				
Model	Method	Base	Def.	BGR	ΔOR	ΔMMLU
Llama-3	Circuit Breakers	1.2	**0.0**	77.1	−48.6	−1.6
	RepBend	1.2	1.0	0.1	+9.2	0.0
	AnchorRep	1.2	1.1	**0**	−7.7	−0.4
Mistral	Circuit Breakers	36.8	1.0	22.1	−2.4	−0.2
	CRL	36.8	1.8	33.6	+17.6	−32.3
	AnchorRep	36.8	**1.0**	**0**	+4.1	−1.5
Vicuna	**AnchorRep**	5.9	0.0	**0**	+6.1	−0.3
Qwen-1.5-14B	**AnchorRep**	2.9	0.4	**0**	+6.2	−0.1
Phi-3	**AnchorRep**	1.9	0.0	**0**	−1.0	0.0

2,000 GCG attacks (20 sources). ASR, BGR, ΔOR, ΔMMLU in %.

Transfer ASR ≤ 1.1% · 0% BGR · Preserved Utility

New Metric: Benign Garble Rate (BGR)

Representation editing can make benign outputs **fluent-looking but degenerate**; refusal metrics miss it. BGR = fraction of benign responses that are garbled.

Examples: `":://://icerichier"`, `"I'mI'mI'm..."`

Take-Aways

- Jailbreak transfer (cross-model and cross-prompt) is mediated by **shared representation geometry**.
- AnchorRep **repels it** with a LoRA adapter, no adversarial examples: **≤1.1% transfer ASR, 0% garble** across 5 models, 4 families.
- Reduces ASR in **21 of 25** adaptive settings (GCG, PAIR, AutoDAN, PGD, TAP), increasing none; embedding PGD and TAP remain open.
- Competing defenses pay up to 77% garble or +18% over-refusal.

github.com/galwert/AnchorRep **huggingface.co/galwert**