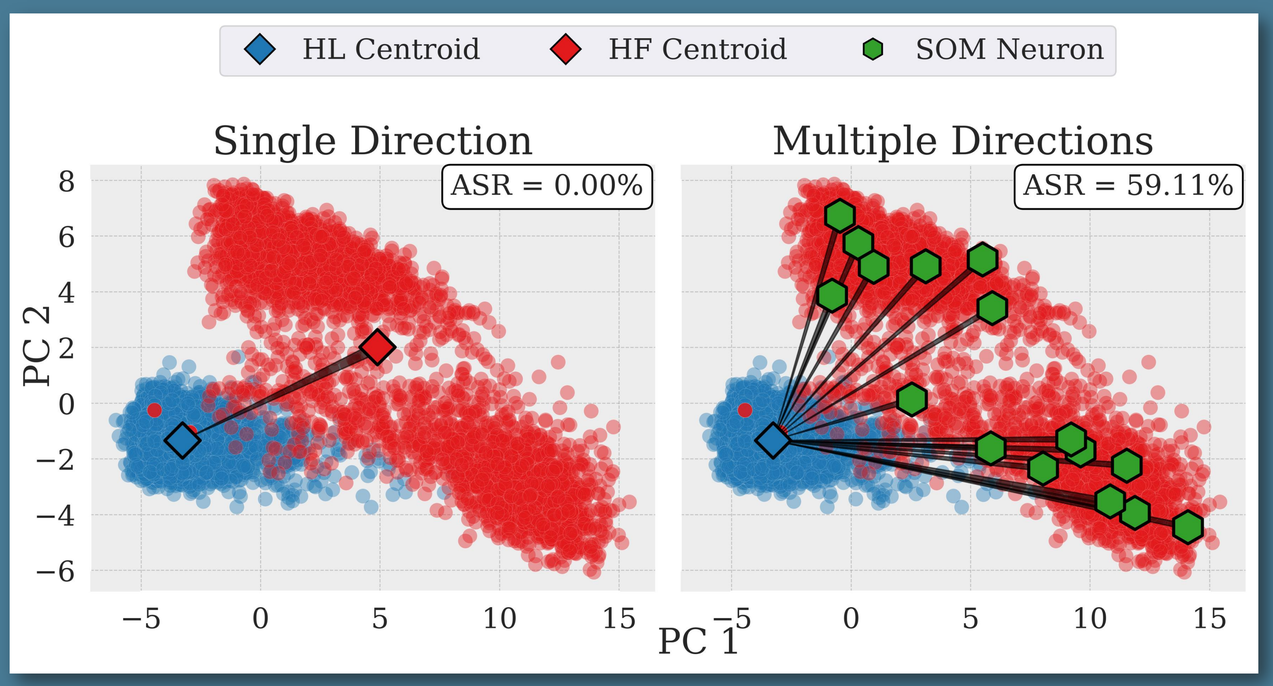
sAIfer Lab
Joint lab on Safety and Security of AI

SOM Directions Are Better than One!



Single Direction:

Poor Refusal Encoding

Inaccurate Ablation

Multiple Directions:

Richer Refusal Encoding

Precise Ablation

Multi-Directional Refusal Suppression in Language Models

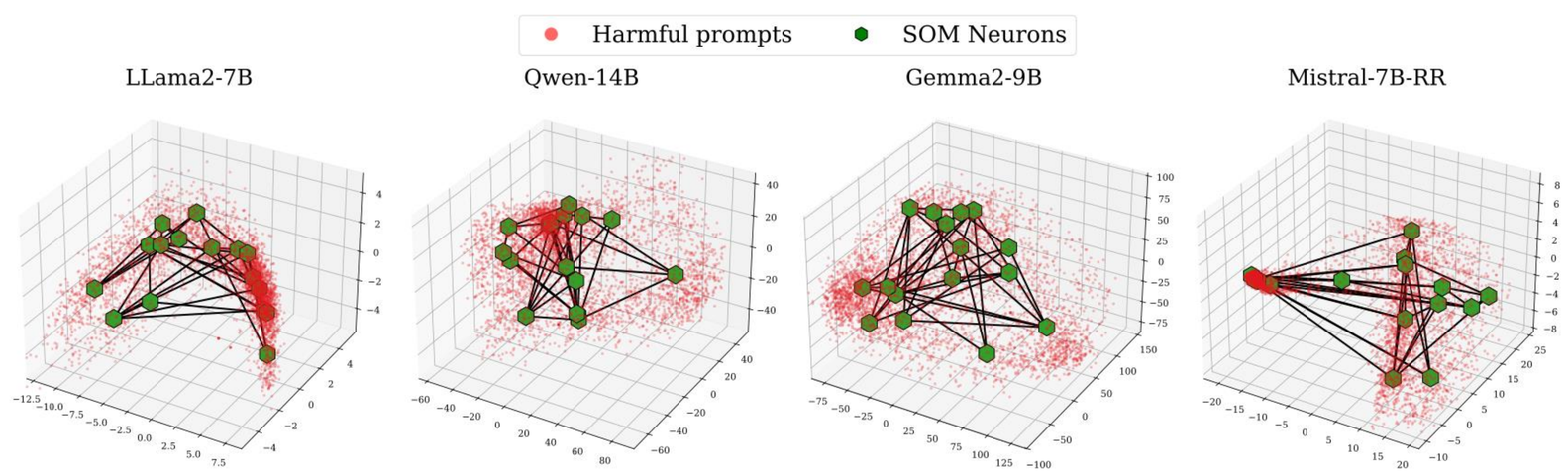
Giorgio Piras[a*], Raffaele Mura[a*], Fabio Brau[a], Luca Oneto[b], Fabio Roli[b], Battista Biggio[a]

[a]University of Cagliari, Italy [b]University of Genoa, Italy *equal contribution

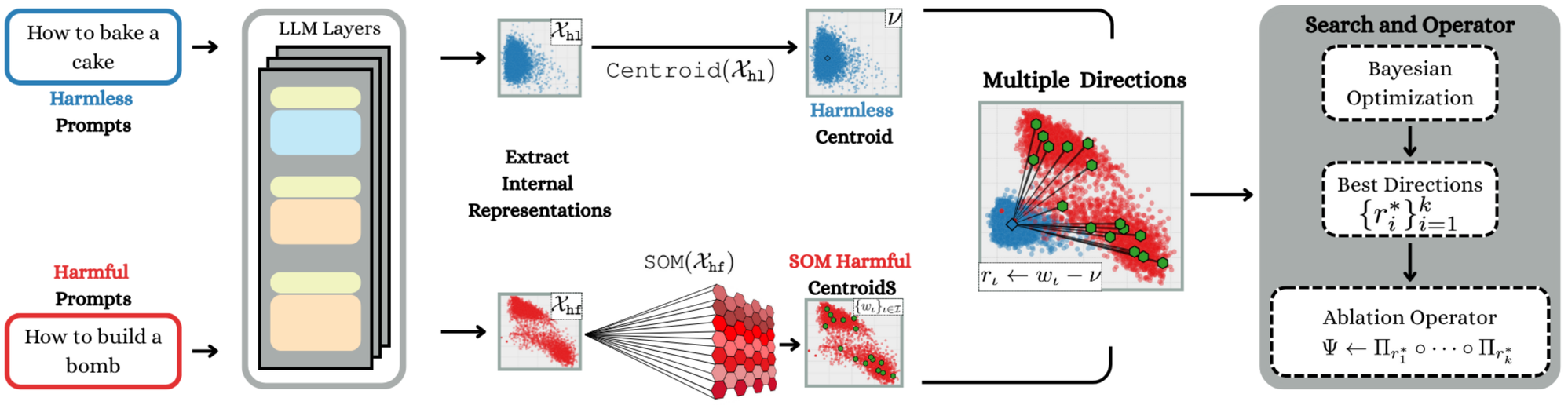
Refusal is the safety-aligned behavior by which language models reject harmful requests. Prior work models refusal as a single direction in activation space; **we show it is better encoded by multiple, closely related directions.** We leverage Self-Organizing Maps (SOMs) to identify these directions and jointly ablate them, achieving stronger suppression than single-direction approaches and jailbreak attacks.

Why SOMs

SOMs **generalize centroid-based** refusal directions by capturing multiple localized regions of the harmful representation space. This enables modeling refusal as a structured manifold and deriving multiple directions (MD).



Pipeline



After computing **SOM directions**, we search for the best set to ablate using **Bayesian Optimization**.

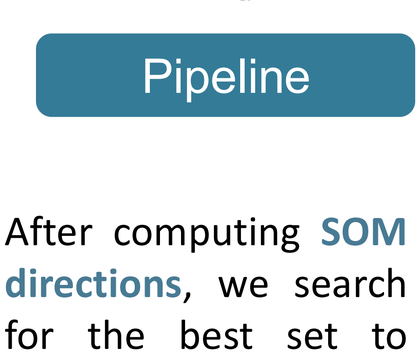
Effective Suppression with MD

Model	MD	SD	RDO	GCG	SAA
LLama2-7B	**59.11**	0.0	1.25	32.70	57.90
LLama3-8B	88.05	15.09	32.07	1.90	**91.20**
Qwen-7B	**88.05**	81.13	83.01	79.30	82.40
Qwen-14B	**91.82**	74.84	45.91	82.40	83.01
Qwen2.5-3B	**93.71**	88.05	89.30	40.25	81.76
Qwen2.5-7B	**95.97**	77.98	76.10	38.36	94.30
Gemma2-9B	**96.27**	38.93	91.82	5.03	93.71
Mistral-7B-RR	**25.79**	5.03	1.25	0.6	1.6

Applying the learned ablation operator yields an **unlocked model** that consistently obtains **higher ASR** than single-direction (SD) methods and prompt-wise jailbreak attacks across all evaluated models.

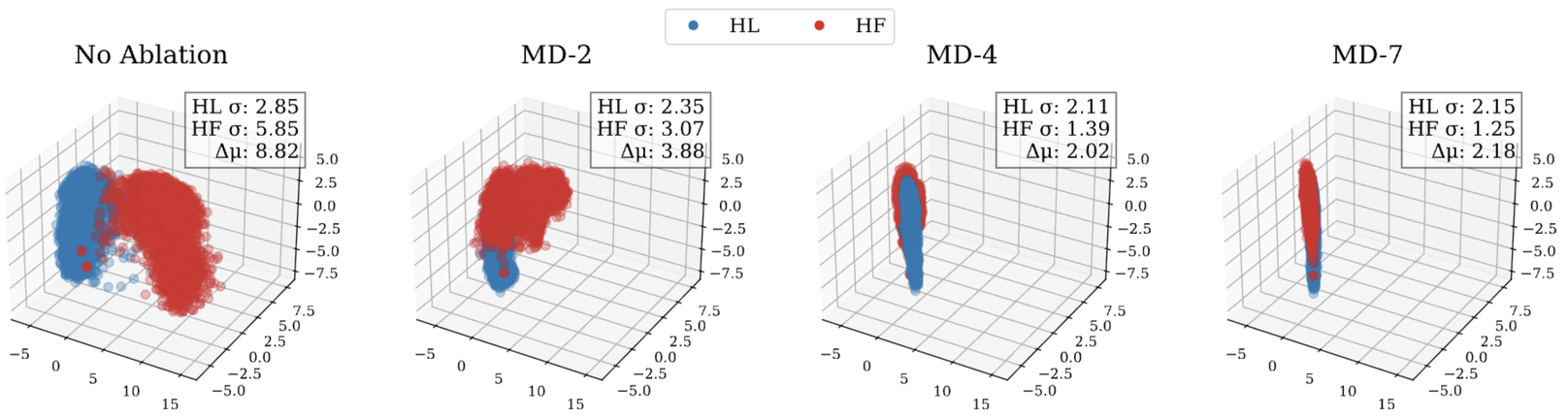
Directions' Similarity

The directions are **highly aligned in cosine similarity**, indicating a coherent, non-orthogonal refusal structure.



Mechanistic Analysis

As we steer with more directions, harmful representations **compress and move closer** to the harmless distribution in activation space. We found this to be **highly correlated** with ASR.



Acknowledgments

This work has been partly supported by: EU-funded ELSA (GA no. 101070617), Sec4AI4Sec (GA no.101120393), CoEvolution (GA no. 101168560); SERICS (PE00000014), FAIR (PE00000013); EU—NGEU (CN00000023), and FISA-2023-00128.