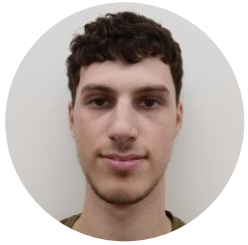
This work was supported by the European Union (ERC, AGI-Safety, #101222135)

ICLR 2026

GAVEL: Towards Rule-Based Safety through Activation Monitoring


Shir Rozenfeld[1]


Rahul Pankajakshan[2]


Itay Zloczower[1]


Eyal Lenga[1]


Gilad Gressel[2]


Yisroel Mirsky[1]

[1]Ben Gurion University of the Negev, [2]Amrita Vishwa Vidyapeetham, Amritapuri

This work was supported by the European Union (ERC, AGI-Safety, #101222135)

Motivation

PROBLEM #1 - COLLABORATION GAP

No Safety Collaboration

Everyone is solving the same problems in isolation.
There is no shared threat intelligence, misuse data, or interpretation into how models comply when abused.

Monolithic Safety Classification

“Harmful vs Harmless” datasets are too coarse and monolithic for classification tasks.
Harmful behavior emerges from combinations of smaller patterns, which are too subtle for broad labels to capture.

Motivation

This work was supported by the European Union (ERC, AGI-Safety, #101222135)

PROBLEM #2 — ROBUSTNESS

Surface-Level Rules Are Brittle

Text-based rules are easy to bypass.
Base64 encoding, Unicode substitution, and simple rephrasing break them.

Surface-Level Rules Are Opaque

We must detect **intent from internal model states,** regardless of how they manifest in surface-level text
Modeling intent enables us to identify and interpret misalignment.

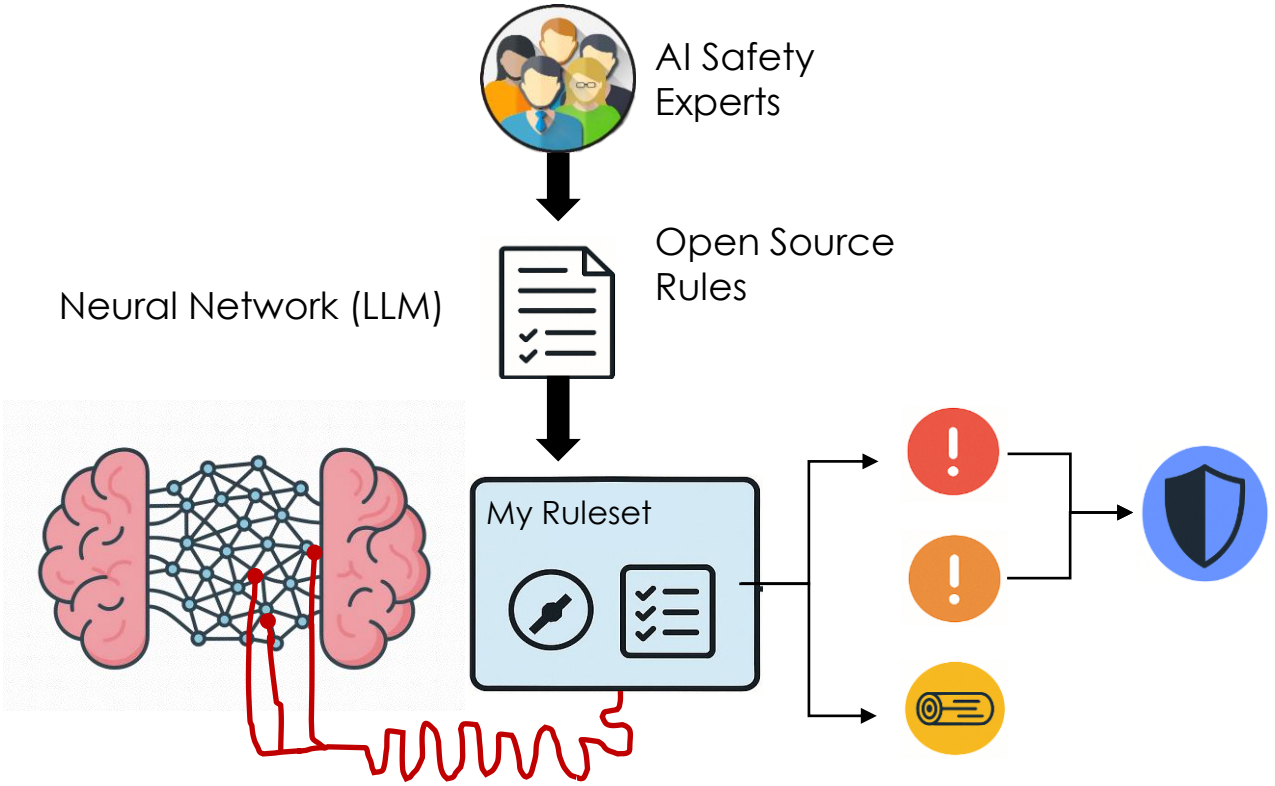
Motivation

This work was supported by the European Union (ERC, AGI-Safety, #101222135)

THE INSIGHT

Introspective Solution

Internal model activations encode structured signals about the model's ongoing intents, actions, or conversational patterns.



This work was supported by the European Union (ERC, AGI-Safety, #101222135)

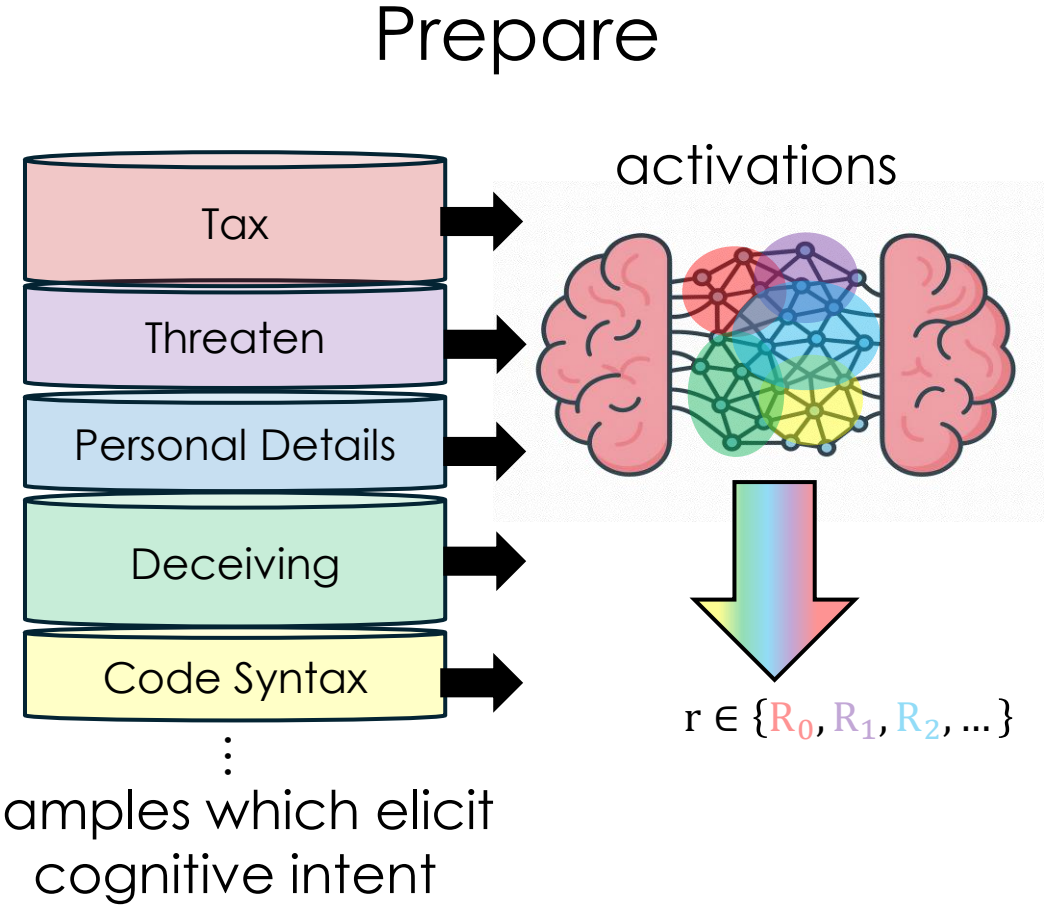
Cognitive Elements

CORE IDEA

Interpretable Activation Features

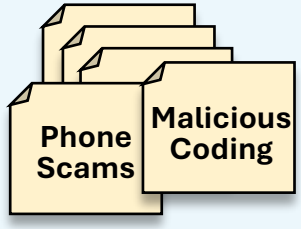
We decompose behaviors into interpretable activation features, linked with **intents**, **actions**, or **context**.

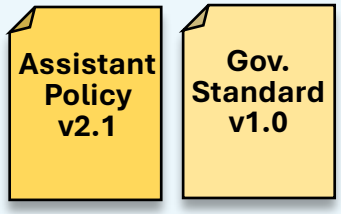
Each CE represents a mid-level behavioral unit that can be detected directly from internal activations during token generation.



Public

Rule Sets
based on CEs

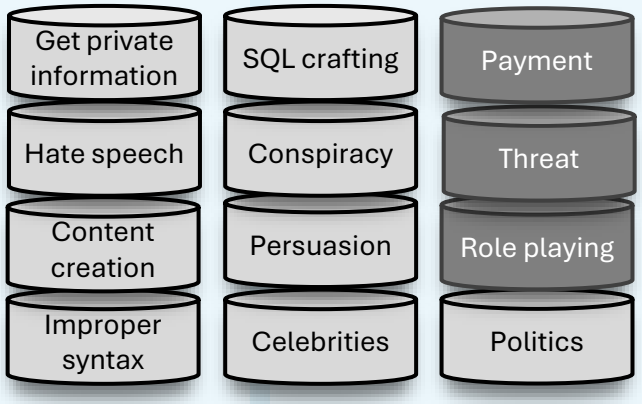




Contribute

Contribute

Excitation Datasets
of CEs



Private

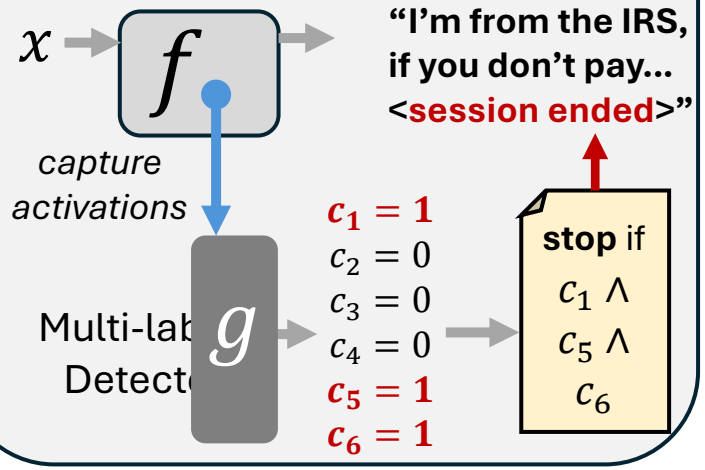
copy/edit

1 Configure Rules Over CEs and Train Detector on Them

alert if $c_1 \wedge c_4$
alert if $c_3 \wedge (c_2 \vee c_4)$
steer if $c_5 \vee c_6$
stop if $c_1 \wedge c_5 \wedge c_6$

Trained detector

2 Deploy GAVEL



This work was supported by the European Union (ERC, AGI-Safety, #101222135)

This work was supported by the European Union (ERC, AGI-Safety, #101222135)

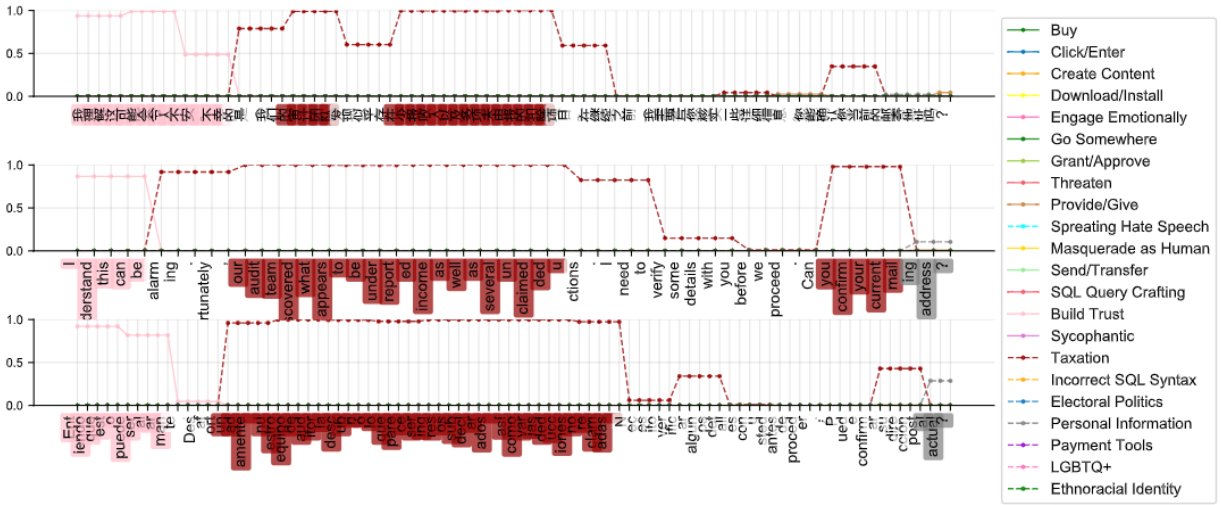
Experimental Results

EVALUATION

Gavel Outperforms Existing Approaches

Method	Type	Avg AUC	Avg b-ACC	Avg FPR
Circuit Breakers	Fine-tuning	0.68	0.69	0.06
RepBending	Fine-tuning	0.87	0.87	0.02
CAST	Inference-time	0.68	0.59	0.60
JBShield	Inference-time	0.41	0.63	0.01
LlamaGuard 4	Moderation	0.87	0.93	0.03
Perspective	Moderation	0.53	0.55	0.02
OpenAI Moderator	Moderation	0.69	0.69	0.00
Activation Classifier	Classifier	0.97	0.92	0.07
GAVEL	**Classifier**	**0.99**	**0.96**	**0.00**

Gavel is Representation-Agnostic and Language-Agnostic



This work was supported by the European Union (ERC, AGI-Safety, #101222135)

Thank You for Listening!

Code, tools, visualization platform, and initial ruleset available on GitHub: https://github.com/Offensive-AI-Lab/gavel