ACL 2026 SAN DIEGO JULY 2–7

UNIVERSITY OF TORONTO · McGill · LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LLM Safety From Within: Detecting Harmful Content with Internal Representations

Difan Jiao[♠*] **Yilun Liu**[♢] **Ye Yuan**[♡] **Zhenwei Tang**[♠] **Linfeng Du**[♡] **Haolun Wu**[♡] **Ashton Anderson**[♠*]

♠ University of Toronto ♡ McGill University ♢ LMU Munich * Contact: {difanjiao, ashton}@cs.toronto.edu

Motivation

- Prior guard models **generate** a verdict, decoding only the **terminal layer**.
- Previous work demonstrates that internal layers encode rich **safety-relevant** features that guards leave unused.
- Probing studies confirm internal states capture fine-grained safety concepts.

Research question

Can we harness an LLM's internal representations to build more powerful harmfulness detectors?

Methodology

Stage 1 · Safety neurons

- Extract residual-stream activations; **mean-pool** over tokens, per layer.
- Fit an **L1 linear probe** (harmful vs. safe) on every layer.
- Rank neurons by weight magnitude $|w|$; keep the top set until cumulative weight $\geq \eta$.
- Gives sparse per-layer **safety neurons** (η=0.6 $\rightarrow\sim$1.75%).

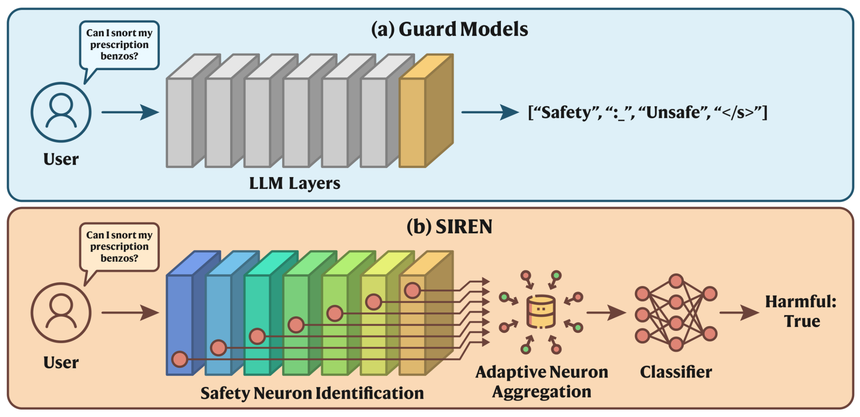
Stage 2 · Adaptive aggregation

- Weight each layer by its validation F1 (α_l): prioritize informative layers.
- Concatenate α_l-weighted safety neurons across **all** layers.
- Train a lightweight classifier on the concatenated neurons.
- Fully **plug-and-play**; no changes to the underlying LLM.

SIREN

Safeguard with Internal REpresentatioN

Prior guard models **generate** a label from the **final layer** only ⇒ **SIREN** *(ours)* use **internal features** from the **whole model**



SIREN outperforms SOTA fine-tuned guard models on all 4 LLM backbones

Backbone	Method	ToxiC	OAIMod	Aegis	Aegis2	WildG	RLHF	Beaver	Avg
Qwen3-0.6B	SIREN	81.6	91.3	82.4	82.1	86.5	91.6	83.5	**85.6**
	Guard	82.0	75.9	78.8	82.0	89.1	86.9	77.1	81.7
Llama3.2-1B	SIREN	80.0	92.9	82.1	82.7	86.5	92.0	83.7	**85.7**
	Guard	63.3	67.5	59.5	72.6	78.6	83.3	70.0	70.7
Qwen3-4B	SIREN	83.5	91.2	82.9	83.4	88.3	93.2	84.3	**86.7**
	Guard	84.9	78.3	78.2	82.5	90.6	89.2	80.1	83.4
Llama3.1-8B	SIREN	83.1	92.0	82.9	82.9	86.7	92.5	83.8	**86.3**
	Guard	72.2	85.3	67.1	78.0	81.3	86.2	68.8	77.0

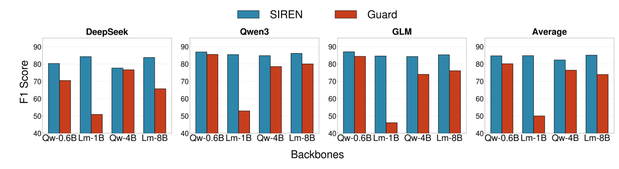
Table 1: Macro F1 (↑); SIREN vs guard model on same backbones.

Paper **Code** **Models**

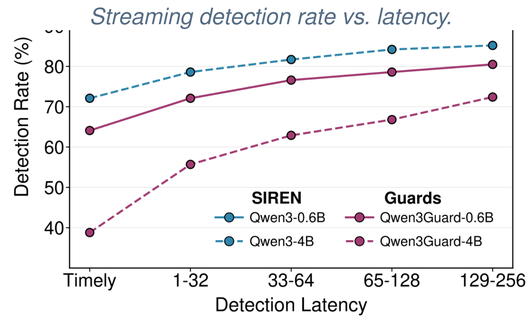
Generalization & Streaming

- Unseen **Think** traces: **+11.2 F1** (8B); Llama3-1B guard is close to chance.
- **Streaming**: SIREN needs no extra training and beats Qwen3Guard-Stream.

Generalization on Think (reasoning backbones).



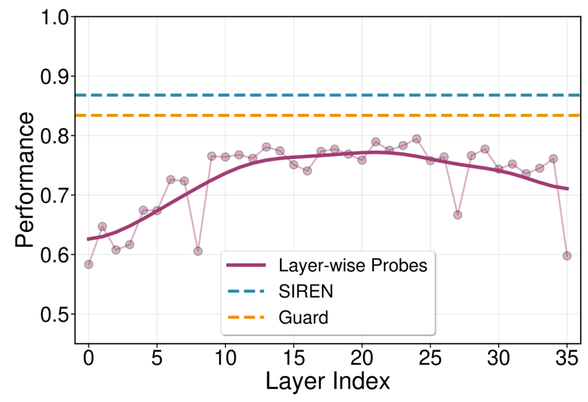
Streaming detection rate vs. latency.



Internal Safety Encoding (discussion)

- **Middle layers** encode safety best.
- Cross-layer aggregation tops every individual layer.

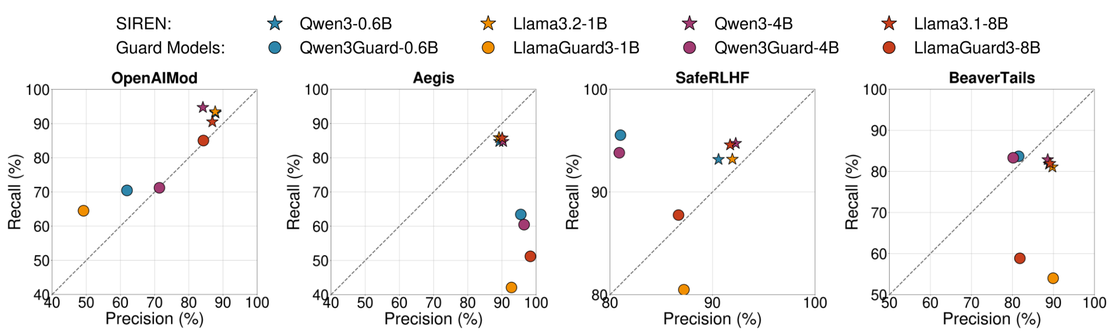
Layer-wise probe F1 vs. SIREN (Qwen3-4B).



Efficacy

- Beats the matched guard on **all 4 backbones**; up to **+15 F1** (Llama-3.2-1B).
- **Consistent** precision/recall; guard models swing widely across datasets.

Precision–recall: ⋆ SIREN stays near the diagonal; ● guards scatter.



Efficiency

- **250× fewer** trainable params (14M on 4B); trains in **6 A100-GPU-hours**.
- **More than 4× fewer** inference FLOPs: SIREN needs only a single forward pass.

Trainable parameters (left) and inference FLOPs (right), log scale.