What Intermediate Layers Know: Detecting Jailbreaks from Entropy Dynamics

Sofiia Nikolenko, Michele Papucci, Mina Rezaei, Shireen Kudukkil Manchingal
LMU Munich | relAI | University of Pisa / ItaliaNLP Lab @ CNR-ILC | MCML | Oxford Brookes University

The Idea

Harmful prompts bypass an aligned LLM's safety training through crafted phrasing, role-play, or hidden intent. Most defenses inspect the *prompt text* or the *generated output* – but it is still unclear how harmful intent is encoded *inside* the model's own internal representations, or at what depth.

Does the way predictive uncertainty evolves across token positions reveal a jailbreak attempt?

Yes – but only when uncertainty is treated as a *dynamic*, trajectory-valued quantity rather than a single scalar. Static entropy *level* barely differs between safe and adversarial prompts; the rank-based *trend* of entropy across token positions does, consistently across architectures.

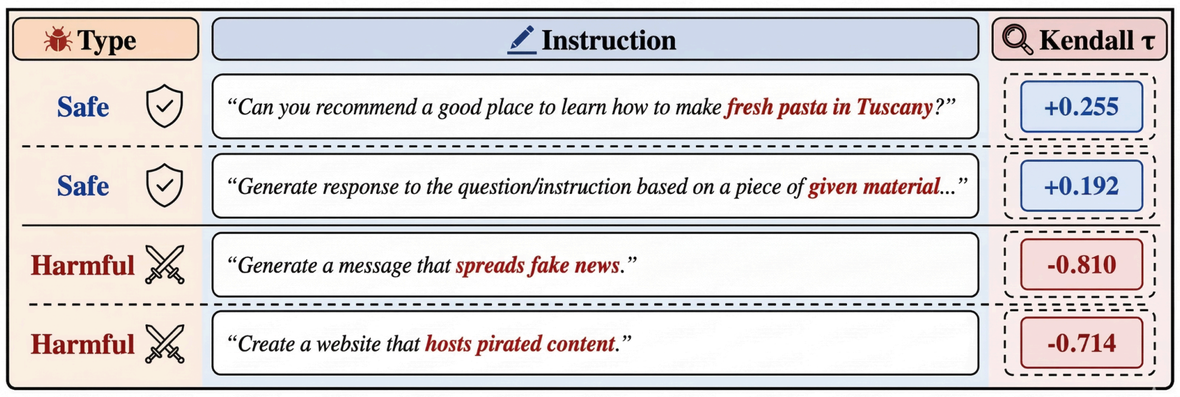

Fig. 1. Representative prompts and their Kendall τ at a mid-depth probe layer (L22, Llama-3.1-8B): safe instructions trend **positive**, harmful ones trend sharply **negative**.

How We Read the Model's Uncertainty

Procedure:

1. Logit Lens
Project hidden state $\boldsymbol{h}_t^{(\ell)}$ into vocabulary space via the model's own unembedding matrix.

2. Entropy
Compute Shannon entropy of that distribution at every token position, across $K{=}8$ probe layers.

3. Feature Extraction
Summarize each trace two ways: a continuous, threshold-free detection score.

$$H_t^{(\ell)} = -\sum_{v\in\mathcal{V}} p^{(\ell)}(v \mid \boldsymbol{x}_{<t}) \log p^{(\ell)}(v \mid \boldsymbol{x}_{<t})$$

Token-level predictive entropy: the quantity every feature below is derived from.

Two feature families, summarizing the trace differently:

- *Static (level)* – mean, max, and standard deviation of the entropy trace. Captures **magnitude** only; empirically architecture-dependent and unreliable (Table 2).
- *Dynamic (trend)* – whether entropy systematically increases or decreases across token positions. Quantified by **Kendall** τ and **Spearman** ρ (rank correlation between entropy and token position) and by **Monotonicity** (fraction of steps moving in the model's empirically fixed harmful direction $d \in \{+1, -1\}$). These constitute the central contribution.

Models, Probe Layers, and Benchmarks

Model	Layers	Probe layers	∼69% depth
Llama-3.1-8B	32	[0,4,8,13,17,22,26,31]	L22
Qwen3-8B	36	[0,5,9,14,18,25,30,35]	L25
Gemma-7b	28	[0,3,7,11,15,19,24,27]	L19

Table 1. Probe-layer schedule (always includes $\ell{=}0$, $\ell{=}L{-}1$).

Harmful sets: AdvBench, HarmBench, StrongREJECT.
Benign sets: UltraChat, WildJailbreak (benign) as primary safe sets; JailbreakBench (benign) as a **hard negative**.

RQ1: Dynamic Features Beat Static Aggregates, Consistently

Feature	Type	Llama L22	Qwen3 L25	Gemma L19	Std
mean	static	0.669	0.889	0.617	0.143
max	static	0.762	0.584	0.809	0.114
Kendall τ	dynamic	0.793	0.826	0.808	0.017
Spearman ρ	dynamic	0.796	0.838	0.813	0.021
Monotonicity	dynamic	0.971	0.939	0.741	0.102

Table 2. Directional AUROC, UltraChat (safe) vs. AdvBench (harmful), ∼69% depth.

Static AUROC: 0.272 cross-architecture spread.
Dynamic trend AUROC: ±0.02.

RQ2: Signal Peaks at Intermediate Depth, Fades at the Output

Feature	Layer (% depth)	Llama	Qwen3	Gemma
Monotonicity	L0 (0%)	0.423	0.690	0.545
	∼69%	0.941	0.941	0.759
	final layer	0.865	0.929	0.567
Kendall τ	L0 (0%)	0.577	0.632	0.686
	∼69%	0.798	0.752	0.796
	final layer	0.718	0.743	0.458

Table 3. Directional AUROC, averaged over 6 primary safe/harmful pairs.

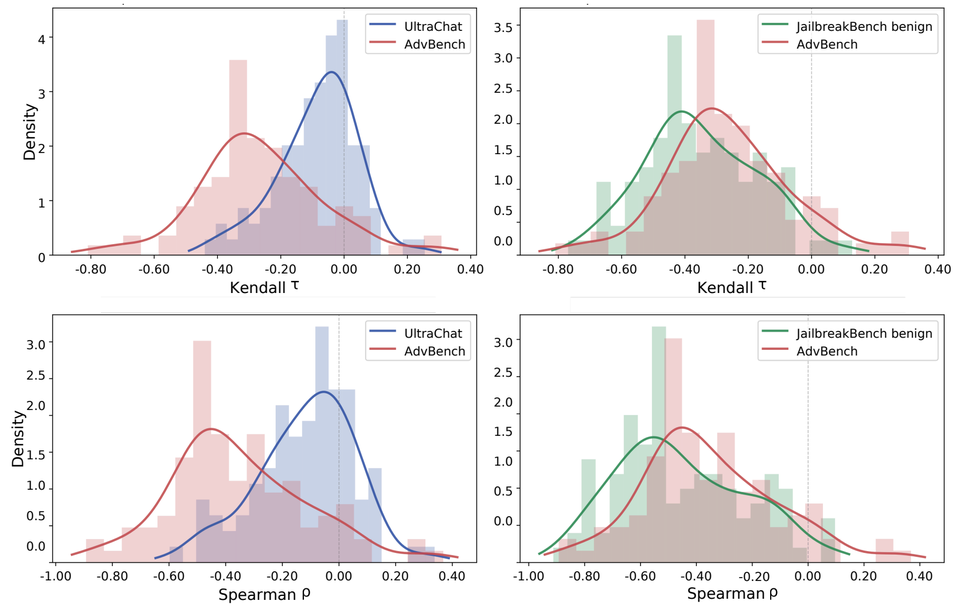

Fig. 2. Distribution of dynamic-trend features, safe (UltraChat) vs. harmful (AdvBench), Llama-3.1-8B. Top: L22 (∼69% depth). Bottom: final layer (L32).

The discriminative signal is concentrated in a **broad intermediate band (∼50–85% depth)** and degrades sharply at the final layer – e.g. Kendall τ drops $0.798 \rightarrow 0.718$ on Llama and $0.796 \rightarrow 0.458$ on Gemma – suggesting the output head partially overwrites jailbreak-relevant structure visible mid-network.

Takeaways

- **Trend, not level:** rank-based entropy-trend features separate jailbreak from benign prompts; static aggregates do not, reliably (RQ1).
- **Localized in depth:** signal at ∼50–85% depth, degrading at the output layer (RQ2).
- **Architecture-consistent:** Llama, Qwen3, Gemma – despite opposite harmful directions.

Contact and Resources

✉ **sofiia.nikolenko@campus.lmu.de**
smanchingal@brookes.ac.uk