Missed jailbreaks differ from benign prompts along directions with low energy in the final-token probe's representational subspace.

Before the Last Token: *Diagnosing Final-Token Safety Probe Failures*

Background: SafeSwitch-style safety probes read one hidden state at the final prompt token to decide whether to trigger a safety intervention. We study why this readout misses jailbreak-wrapped harmful requests across three instruction-tuned LLMs

1. One Readout, One Failure Mode: SafeSwitch-style probes read one hidden state at the final prompt token. High recall on clean harmful prompts. But jailbreak wrappers cause significant misses.

Question: Is the harmful request absent from the internal representations or is it not exposed enough at the final token readout?

2. Geometry - Where the Probe Is Blind: We compute difference-of-means directions in the final-token hidden-state space and measure their projection energy onto the probe's row-space

The missed-jailbreak direction has the lowest energy across all three models (2.4–3.5%).

Let $W \in \mathbb{R}^{w \times D}$ be the probe's first-layer weight matrix. $\Pi_W = V_r V_r^\top$ projects onto $\mathrm{rowspace}(W)$.

$$\mathrm{E}_W(d) = \frac{\|\Pi_W\, d\|_2^2}{\|d\|_2^2} \in [0, 1] \qquad \textit{fraction of } d \textit{ visible to the probe}$$

$d_{\mathrm{harm}} = u\big(\mu(\mathcal{H}_{\mathrm{train}}) - \mu(\mathcal{B}_{\mathrm{train}})\big)$ *training contrast*

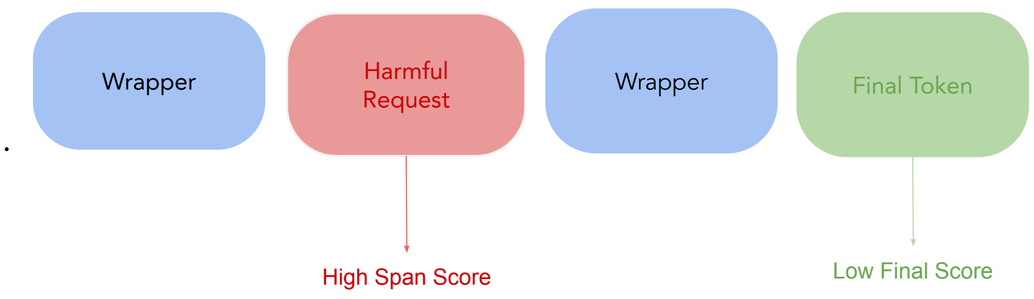
$d_{\mathrm{safe}} = u\big(\mu(\mathcal{S}_{\mathrm{sorry}}) - \mu(\mathcal{X}_{\mathrm{xstest}})\big)$ *safety-contrast proxy*

$d_{\mathrm{miss}} = u\big(\mu(\mathcal{J}_{\mathrm{miss}}) - \mu(\mathcal{B}_{\mathrm{train}})\big)$ *failure direction*

$\Delta_{\mathrm{cm}} = u\big(\mu(\mathcal{J}_{\mathrm{caught}}) - \mu(\mathcal{J}_{\mathrm{miss}})\big)$ *catch vs. miss*

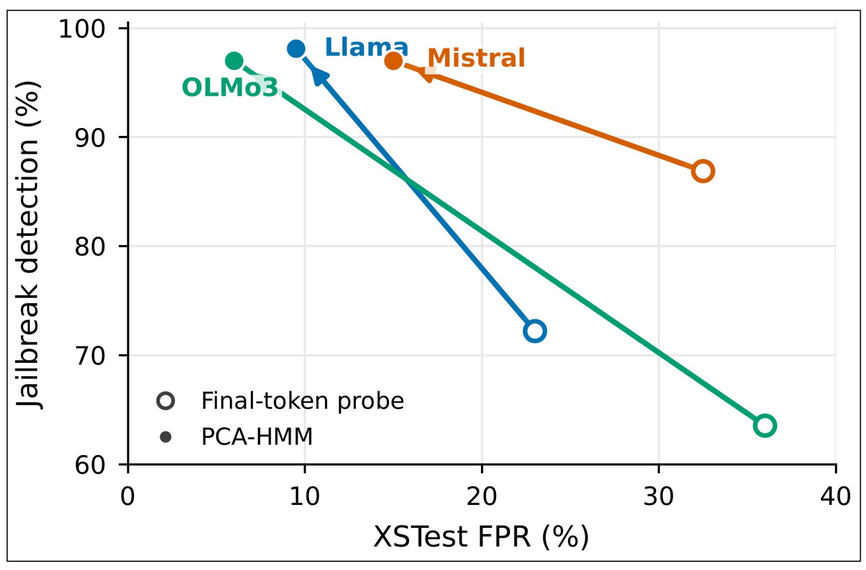
Direction	Llama	Mistral	OLMo3
d_{harm}	**9.0**	**12.3**	**12.6**
d_{safe}	6.7	5.3	7.6
d_{miss}	**3.5**	**2.4**	**3.2**
Δ_{cm}	4.0	5.5	4.5

3: Evidence Before the Final Token: For located spans, the same frozen probe scores high at the harmful-request position but low at the final token



Model	Missed	Located	Goal score	Final score
Llama-3.1-8B	250	87	**0.998**	**0.174**
Mistral-7B	118	65	**0.998**	**0.168**
OLMo3-7B	328	127	**0.946**	**0.121**

4. Trajectories Recover Final-Token Misses: A PCA-HMM trajectory diagnostic trained on clean data recovers 89–95% of final-token misses with lower XSTest false-positive rates.



5. Takeaways

1. Missed jailbreaks differ from benign prompts along directions with low projection energy in the probe visible subspace – a failure of representational exposure, not probe capacity.
2. Safety-relevant features often remain detectable during prefill before collapsing at the final-token readout.
3. When safety-relevant information is exposed may be as important as where it is read out.

Limitation: This is a diagnostic study and not a deployable defense. False positives remain nonzero, and HMM states should not be interpreted as mechanistic safety features.

PAPER

Shravan Doda

Mechanistic Interpretability Workshop, ICML 2026

KIPOAI