HERALD: Harmfulness Propagation Dynamics

Layer-wise Trajectories of Adversarial Intent in Large Language Models

Noor Islam S. Mohammad Uluğ Bayazıt

Department of Computer Science / Faculty of Computer Engineering
Istanbul Technical University, Maslak, Türkiye

Mechanistic Interpretability Workshop, ICML 2026, Seoul, South Korea

Input Moderation: A Spectrum of Trade-offs

Guard models

- Standalone safety classifiers
- Strong accuracy
- Costly: ≈14 GB for a 7B guard
- No access to host-model internals

Latent-based methods

- Use backbone activations
- Lightweight
- Commit to a *single* fixed layer
- Discard cross-depth information

Our question

Is harmful intent a *progressively resolved* property of a prompt — and can the **trajectory** of a representation across layers, not just one snapshot, serve as a discriminative signal?

Prior probing results

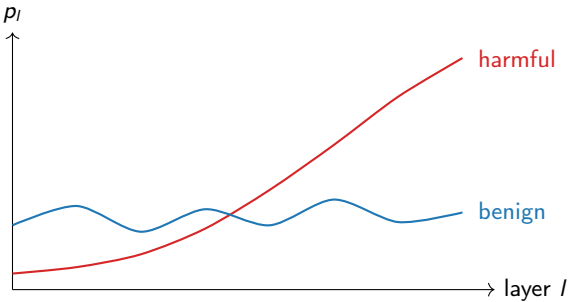
Syntax → early layers
Semantics → middle layers
Pragmatics / intent → late layers

Harmfulness Propagation Dynamics (HPD)

Core observation. Project the last-token hidden state at every layer onto a per-layer harm direction v_l:

$$p_l(x) = \left\langle \frac{h_l}{\|h_l\|}, \frac{v_l}{\|v_l\|} \right\rangle \in [-1, 1]$$

- **Harmful prompts:** near-zero → steady monotonic rise → strongly positive
- **Benign prompts:** flat / oscillatory, mean ≈ 0
- Stable across 4 model families, harm categories, paraphrases



Three phases for harmful prompts: surface form ($l \lesssim 8$) → monotonic rise (middle) → $p_L \gtrsim 0.4$ (late layers)

Not just "late layers are discriminative"

Trajectory *shape* — onset layer, monotonicity, curvature — carries information beyond the terminal value p_L, especially for adversarial jailbreaks.

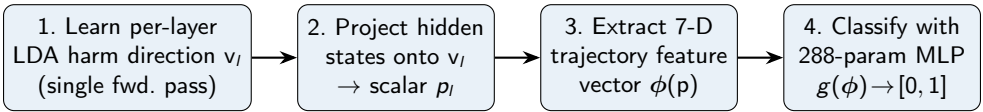
Theoretical Grounding

Proposition (informal)

If (i) harmful intent is a semantic-pragmatic property encoded mainly in later layers, (ii) the LDA harm direction v_l is a consistent estimator of the optimal Fisher discriminant at layer l, and (iii) projections of harmful representations onto v_l increase with l in expectation while benign projections stay bounded — then harmful trajectories are monotonically increasing in expectation, while benign trajectories satisfy $\mathbb{E}[p_l] \approx 0$ for all l.

- Consistent with the layered-computation hypothesis: early layers $\rightarrow$ tokenization / lexical identity; later layers $\rightarrow$ abstract semantic and pragmatic properties.
- Assumption (ii) validated empirically: pairwise cosine similarity of v_l across independent splits > 0.97 at *every* layer and backbone.
- $\Rightarrow$ LDA converges to a stable population direction, not sampling noise.

HERALD: Overview of the Pipeline



262 KB
memory (L=32, d=4096)

2.6×10^{-6}
fraction of prefill FLOPs

Gradient-free
trains in seconds on CPU

Per-Layer Harm Directions via LDA

For each layer l, find the direction maximizing class separation:

$$\mathrm{v}_l = \arg\max_{\|\mathrm{v}\|=1} \frac{\mathrm{v}^\top \mathrm{S}_B^{(l)} \mathrm{v}}{\mathrm{v}^\top \mathrm{S}_W^{(l)} \mathrm{v}}$$

Closed-form solution:

$$\mathrm{v}_l \propto \left(\hat{\mathrm{S}}_W^{(l)}\right)^{-1} \left(\boldsymbol{\mu}_l^{\text{harm}} - \boldsymbol{\mu}_l^{\text{safe}}\right)$$

- $\hat{\mathrm{S}}_W^{(l)}$: Ledoit–Wolf shrinkage estimate (well-conditioned when $d \gg n_{\text{train}}$)
- Each layer retains only $\mathrm{v}_l \in \mathbb{R}^d$ ($\approx$8 KB at d=4096, fp16); scatter matrices discarded after training

Why LDA, not mean-difference?
Class-mean difference $\boldsymbol{\mu}_l^{\text{harm}} - \boldsymbol{\mu}_l^{\text{safe}}$ ignores within-class scatter (length, wording, rhetorical style).

Direction	Llama-8B	Mistral-7B	OLMo2-7B
Random (ctrl)	77.2	76.9	79.9
Mean-diff.	85.8	85.5	88.5
PCA on residuals	85.4	85.1	88.1
LDA (used)	**86.3**	**86.3**	**89.3**

Average F1; LDA best on all backbones.

Trajectory Feature Extraction

Given normalized projections $p_l = \langle \hat{\mathrm{h}}_l, \mathrm{v}_l \rangle$, build $\phi(\mathrm{p}) \in \mathbb{R}^7$:

$$\phi(\mathrm{p}) = \left[\, p_L,\ \bar{p},\ p_L - p_1,\ \Delta^2\mathrm{p},\ \mathrm{mono}(\mathrm{p}),\ p_{\hat{l}^*},\ \hat{l}^* \,\right]$$

Feature	**Symbol**	**Geometric role**
Final value	p_L	Terminal harm alignment
Mean	$\bar{p}$	Global bias across layers
Total rise	$p_L - p_1$	Net directional movement with depth
Curvature	$\Delta^2\mathrm{p}$	Trajectory smoothness / oscillation
Monotonicity	$\mathrm{mono}(\mathrm{p})$	Consistency of directional increase
Onset value	$p_{\hat{l}^*}$	Projection at first salient layer
Onset layer	$\hat{l}^*$	Depth at which harm signal first appears

- $\hat{l}^* = \min\{l : p_l > \tau_{90}\}$ (first layer exceeding 90th-percentile threshold)
- Logistic regression on ϕ already within 1.0 F1 of full MLP $\Rightarrow$ ϕ is nearly linearly separable

Trajectory Classifier & Algorithm

Two-layer MLP, hidden size 32, **288 parameters**:

$$\textsc{Herald}(x) = \sigma\big(\mathrm{W}_2\,\mathrm{ReLU}(\mathrm{W}_1\phi(\mathrm{p}) + \mathrm{b}_1) + b_2\big)$$

- Trains on CPU in seconds
- Larger architectures (hidden 64–128, 2 layers): no significant gain
- Bottleneck is representation quality, not classifier capacity

Training (gradient-free)

1. Single forward pass: collect last-token hidden states $\{\mathrm{h}_l\}$ for all training prompts
2. Per layer: compute class means, Ledoit–Wolf $\hat{\mathrm{S}}_W^{(l)}$, solve for v_l
3. Compute projections p_l; extract ϕ; train MLP g

Inference

Given new prompt x: one forward pass $\rightarrow \{\mathrm{h}_l\} \rightarrow \phi(\mathrm{p}) \rightarrow g(\phi(\mathrm{p})) \geq 0.5 \Rightarrow$ harmful. **No extra forward passes.**

Efficiency: Memory & Compute

Approach	Memory	Relative to Herald
HERALD ($L \times d$, fp16)	262 KB	1×
Full per-layer covariance ($L \times d^2$)	1.07 GB	∼4,100×
7B-parameter guard model	≈14 GB	∼53,000×

Inference overhead

$O(2Ld)$ FLOPs: L dot products + L normalizations.

For L=32, d=4096: ≈262K FLOPs vs. ≈100B prefill FLOPs (100-token prompt)

ratio $\approx 2.6 \times 10^{-6}$

Why it matters

- $O(Ld)$ vs. $O(Ld^2)$: fundamental, not incidental
- Deployable on the *same* hardware as the host model
- No additional accelerators required

Experimental Setup

8 prompt-harmfulness benchmarks

- Aegis, HarmBench, OpenAI Moderation (OAI)
- SimpleSafetyTests (SimpST), ToxicChat (TChat)
- WildGuardMix (WGMix), WildJailbreak (WJB), XSTest

Trained on WildGuardMix; evaluated zero-shot on the rest. Metric: macro-F1.

4 instruction-tuned backbones

- Llama-3.1-8B-Instruct
- Mistral-7B-Instruct
- OLMo2-7B-Instruct
- Qwen3-8B-Instruct

Baselines

- *Latent-based:* Embed. Clf., Act. Delta (same backbone)
- *Guard models:* 4 standalone classifiers (Guard A–D)

All results: mean over 3 seeds; $\geq$0.5 F1 improvements are significant ($p < 0.05$, paired bootstrap).

Main Results: Average F1

Backbone	Embed. Clf.	Act. Delta	Herald (ours)
Llama-3.1-8B	79.0	82.0	**86.3**
Mistral-7B	81.7	81.6	**86.3**
OLMo2-7B	84.7	85.0	**89.3**
Qwen3-8B	80.5	81.9	**84.7**

HERALD outperforms latent baselines on *every* backbone, by 2.3–4.1 F1.

- Best overall: OLMo2-7B, average F1 = **89.3**
- Guard models (standalone, no activation access): best is Guard D at 87.8 avg. F1
- Failure modes: on ToxicChat and OpenAI Moderation, HERALD trails the best guard by 2–4 F1 (discussed later)

Advantage on Adversarial Jailbreaks (WildJailbreak)

	Llama-8B	Mistral-7B	OLMo2-7B	Qwen3-8B
Act. Delta (latent baseline)	90.8	88.4	91.0	87.8
Best Guard model (A–D)		96.9 (cross-backbone)		
HERALD (ours)	**95.8**	**94.3**	**98.4**	**92.1**

Why HERALD excels here

Jailbreaks conceal harmful intent in early tokens and reveal it progressively through multi-step framing $\rightarrow$ exactly the monotonically rising trajectory that HPD captures. A single-layer snapshot reads only the terminal state and misses *how* it got there.

On OLMo2-7B, Herald surpasses all four guard models: 98.4 vs. next-best 96.9 F1.

Data Efficiency & OOD Generalization

Backbone	Method	F1@100	F1@1k	F1@10k	F1@Full
OLMo2-7B	Embed. Clf.	83.1	84.9	85.6	85.7
	Act. Delta	81.0	83.6	85.1	85.9
	HERALD	**83.7**	**86.5**	**88.0**	**87.9**

- HERALD reaches near-plateau performance at **1,000 samples/class**
- Matches baselines' *full-data* F1 using 10× fewer samples (100-sample mark)
- OOD (train on Aegis, test on WildGuardMix): smaller performance drop than both baselines
- Why: 7 geometrically stable scalars act as an implicit regularizer in low-data regimes

Ablation: Trajectory Feature Components

Features added **cumulatively** (WildGuardMix, avg. F1; WJB column = OLMo2-7B):

Feature configuration	Llama-8B	Mistral-7B	OLMo2-7B	WJB (OLMo2)
p_L only	84.7	84.4	87.8	96.3
$+ \bar{p}$	85.0	84.7	88.1	96.6
$+ (p_L - p_1)$	85.5	85.2	88.5	97.1
$+ \mathrm{mono}$	85.9	85.8	88.9	97.6
$+ \Delta^2 \mathrm{p}$	86.1	86.1	89.1	97.9
$+ \hat{l}^*$ (full)	**86.3**	**86.3**	**89.3**	**98.4**

- Onset layer $\hat{l}^*$ gives the largest single gain, especially on WJB ($+1.3$ F1 incrementally)
- Removing *any* feature from the full model reduces performance $\rightarrow$ each contributes independently

Ablation: Direction Learning & Layer Coverage

Direction learning method

Method	Llama-8B	Mistral-7B	OLMo2-7B
Random (control)	77.2	76.9	79.9
Class-mean diff.	85.8	85.5	88.5
PCA on residuals	85.4	85.1	88.1
LDA (used)	**86.3**	**86.3**	**89.3**

LDA's gain over mean-difference (0.5–0.8 F1) quantifies the value of within-class covariance estimation.

Layer coverage strategy

Selection	Llama-8B	Mistral-7B	OLMo2-7B
Single best (oracle)	85.5	85.0	88.5
Early third ($l \leq 10$)	80.4	79.9	83.2
Middle third	84.2	84.7	87.1
Late third ($l \geq 22$)	84.9	83.8	88.0
Top-8 by val. F1	85.9	85.8	89.0
All layers (used)	**86.3**	**86.3**	**89.3**

All-layer aggregation beats the oracle single layer by 0.8–1.4 F1; top-8 recovers most of the gain at 25% storage.

Ablation: Token Position & Suffix-Padding Robustness

Which token to project?

Token position	Llama-8B	Mistral-7B	OLMo2-7B
First token	80.6	79.3	82.4
Mean over tokens	84.1	83.7	87.2
Last token (used)	**86.3**	**86.3**	**89.3**

Last token aggregates context via causal self-attention → concentrates task-relevant information.

Robustness to benign suffix attacks (Llama-8B, WGMix)

Variant	k=**0**	k=**10**	k=**50**	k=**100**	k=**200**
Last token (default)	**86.3**	**85.8**	**85.0**	**83.7**	**81.5**
Max-pool over layers	**85.9**	**85.7**	85.4	85.0	83.8

Degrades gracefully (-1.3 F1 at k=50, -4.8 at k=200). Max-pooling is more robust to long suffixes at -0.4 F1 on clean inputs — recommended when suffix attacks are a realistic threat.

Ablation: Normalization, Classifier Capacity, Shrinkage

Projection normalization

	L-8B	M-7B	O-7B
None	84.5	84.0	87.5
v_l only	85.1	84.8	88.0
Both (used)	86.3	86.3	89.3

Normalizing both h_l, v_l: +1.3–1.8 F1.

MLP capacity

Arch.	Params	O-7B
Log. reg.	8	88.4
Hidden 16	128	88.9
Hidden 32 (used)	288	89.3
Hidden 64	512	89.3
Hidden 128	1024	89.2
2×32 layers	1344	89.4

Plateau at hidden=32; logistic regression within 1.0 F1.

Shrinkage regularization

Inversion	O-7B	F1@100
None (raw inv.)	63.7	52.1
Diagonal approx.	88.2	78.3
Fixed ridge	89.0	81.7
Ledoit–Wolf (used)	89.3	79.6

Without shrinkage, F1 drops up to 27 pts in low-data regimes.

Summary across ablations

Onset layer & LDA direction learning contribute most; token position, layer coverage, normalization each add meaningfully; shrinkage is critical for numerical stability; **classifier capacity matters least**.

Discussion: Failure Modes & Limitations

Why HPD excels on jailbreaks

- Multi-step framing → harmful intent revealed gradually → rising trajectory
- Jailbreaks: earliest onset ($\hat{l}^* \approx 7$), highest monotonicity (0.83)
- Social stereotypes: latest onset ($\hat{l}^* \approx 19$), lowest monotonicity (0.58) — diffuse, less geometrically coherent

Failure modes

- ToxicChat: −4.3 F1 vs. best guard
- OpenAI Moderation: −8.4 F1 vs. best guard
- Hypothesis: implicit/context-dependent harms; WildGuardMix training distribution under-represents their style

Adaptive evasion

- Harder to evade than single-score detectors: attacker must jointly fool terminal value, onset, monotonicity, *and* curvature
- But white-box adversaries with access to learned directions could suppress early-layer projections while keeping a high terminal value

Ethical considerations

- Harm directions derived from (mostly English) WildGuardMix — validate before multilingual / cross-cultural deployment
- Spurious demographic/stylistic correlations → possible false positives; fairness auditing recommended
- Publishing HPD characterizations could inform evasion strategies → continued red-teaming needed

Conclusion

Summary

- **HPD**: harmful prompts show monotonically rising harm-direction projections across layers; benign prompts stay flat/oscillatory
- Per-layer LDA directions are stable (cosine sim. > 0.97) $\rightarrow$ trajectory is a reproducible structured signal
- **Herald**: 7-D trajectory features + 288-param MLP
  - 262 KB memory, 2.6×10^{-6} prefill FLOPs
  - +2.3–4.1 F1 over latent baselines on every backbone (avg. F1 up to 89.3)
  - Surpasses all guard models on adversarial jailbreaks (98.4 vs. 96.9 F1)
- Per-instance trajectories $\rightarrow$ interpretable, machine-readable audit records for LLM governance

Future Work

Multi-directional subspaces for diffuse harms (social stereotypes); multilingual / multimodal extensions; trajectory-aware adversarial training; integration into RAG pipelines.

Thank you — Questions?