Harmfulness Propagation Dynamics: Layer-wise Trajectories of Adversarial Intent in Large Language Models

Noor Islam S. Mohammad [1] Uluğ Bayazıt [2]

[1]Department of Computer Science, Informatics Institute, Istanbul Technical University
[2]Faculty of Computer and Informatics Engineering, Istanbul Technical University

ICML International Conference On Machine Learning

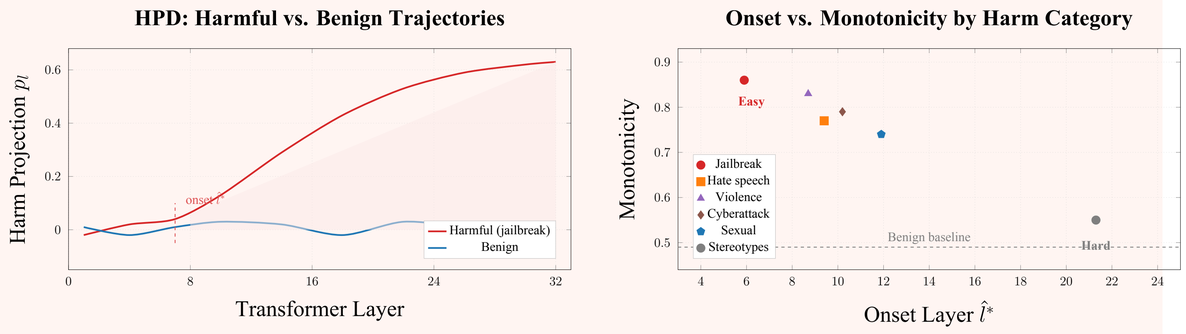
Abstract

We identify **Harmfulness Propagation Dynamics (HPD)**: for harmful prompts, the projection of the last-token hidden state onto a learned harm direction *rises monotonically* with transformer depth, whereas benign prompts remain *flat or oscillatory*. Built on HPD, **HERALD** extracts a **7-dimensional** trajectory feature vector, slope, curvature, monotonicity, and onset layer, and classifies it with a **288-parameter MLP**. HERALD adds only 262 KB of memory and 2.6×10^{-6} of prefill FLOPs, yet surpasses all tested guard models on adversarial jailbreak detection (**98.4 vs 96.9 F1**) and outperforms all latent-based baselines by **2.3–4.1 F1** across eight benchmarks.

Research Objectives

Investigate whether **harmful intent** is encoded as a *progressive*, cross-layer signal in LLM hidden states, rather than concentrated at a single layer. Characterize **Harmfulness Propagation Dynamics (HPD)**: the trajectory along which harmful prompts' projections onto a learned harm direction evolve with transformer depth, versus benign prompts. Establish that per-layer LDA harm directions are **stable and reproducible** across independent data splits. Design HERALD, a **lightweight, gradient-free input moderator** that exploits trajectory shape (slope, curvature, onset, and monotonicity) rather than a single-layer snapshot. Evaluate whether trajectory-based features can **match or exceed** guard models and latent-based baselines on harmfulness detection, particularly adversarial jailbreaks, at a fraction of the memory and compute cost. Provide **interpretable, machine-readable audit records** of per-prompt harm trajectories to support LLM governance and red-teaming workflows.

Harmfulness Propagation Dynamics



Key insight: Two prompts may share the same terminal projection p_L yet have fundamentally different risk profiles. Only the full trajectory, especially the onset layer $\hat{l}^*$ and monotonicity, reveals the difference. Jailbreaks: $\hat{l}^*\approx5.9$, mono 0.86; social stereotypes: $\hat{l}^*\approx21.3$, mono 0.55.

Methodology

Per-layer harm directions via LDA. For each transformer layer $l \in \{1,\dots,L\}$, collect the last-token hidden state $\mathbf{h}_l \in \mathbb{R}^d$ for all training prompts under a single gradient-free forward pass. Apply binary LDA with analytic Ledoit-Wolf shrinkage [1] to find the direction $\mathbf{v}_l$ maximally separating harmful from safe representations:

$$\mathbf{v}_l \propto \left(\hat{\mathbf{S}_W}^{(l)}\right)^{-1}\left(\boldsymbol{\mu}_l^{\text{harm}} - \boldsymbol{\mu}_l^{\text{safe}}\right), \qquad \hat{\mathbf{S}_W}^{(l)} = (1-\alpha)\mathbf{S}_W^{(l)} + \frac{\alpha\,\mathrm{tr}(\mathbf{S}_W^{(l)})}{d}\mathbf{I}$$

Shrinkage (α computed analytically) is essential when $d \gg n_{\text{train}}$: omitting it degrades F1 by up to 27 points. LDA accounts for within-class scatter from length, wording, and rhetorical style variation—outperforming the raw class-mean difference by 0.5–0.8 F1. Only $\mathbf{v}_l \in \mathbb{R}^d$ is retained per layer (≈8 KB fp16); scatter matrices are discarded after fitting, giving $O(Ld)$ total storage.

Cross-layer harm trajectory. At inference, unit-normalize both the hidden state and the stored direction, then compute the cosine projection $p_l = \langle \hat{\mathbf{h}}_l, \hat{\mathbf{v}}_l \rangle \in [-1,1]$ at every layer. The resulting sequence $\{p_l\}_{l=1}^{L}$ is the *harm trajectory*. For harmful prompts—especially jailbreaks—this trajectory rises *monotonically* from near-zero in early layers ($l \lesssim 8$) to strongly positive values ($p_L \gtrsim 0.4$) in late layers, reflecting the progressive consolidation of pragmatic intent with depth. Benign prompts remain flat or oscillatory with a mean near zero. This **Harmfulness Propagation Dynamics (HPD)** pattern is stable across all four model families tested (Llama-3.1-8B, Mistral-7B, OLMo2-7B, Qwen3-8B) and is validated by LDA direction stability: pairwise cosine similarity > 0.97 across five independent splits at every layer.

Trajectory feature extraction. The full L-dimensional trajectory is compressed into a compact 7-D feature vector $\boldsymbol{\phi}(\mathbf{p}) \in \mathbb{R}^7$ capturing complementary geometric properties: terminal harm alignment (p_L), global mean bias ($\bar{p}$), net directional rise ($p_L - p_1$), trajectory smoothness via mean absolute curvature ($\Delta^2\mathbf{p}$), consistency of directional increase via monotonicity ratio ($\mathrm{mono}(\mathbf{p})$), and the depth at which the harm signal first crosses the 90th-percentile threshold ($\hat{l}^*$, $p_{\hat{l}^*}$). The onset layer $\hat{l}^*$ provides the largest single performance gain (+1.3 F1 incrementally on WildJailbreak), as it captures structural differences invisible to any single-layer snapshot: jailbreaks exhibit early onset ($\hat{l}^* \approx 7$, monotonicity 0.83) whereas social stereotypes onset late ($\hat{l}^* \approx 19$, monotonicity 0.58).

Trajectory classifier. A two-layer MLP $g : \mathbb{R}^7 \to [0,1]$ with hidden size 32 (288 parameters total) classifies $\boldsymbol{\phi}(\mathbf{p})$ via a sigmoid output. The classifier trains on CPU in seconds; logistic regression already achieves within 1.0 F1 of the full MLP, confirming the feature space is nearly linearly separable and that gains derive from representation quality rather than classifier capacity.

Inference cost. HERALD performs L dot products and L normalizations per prompt, adding $O(2Ld)$ FLOPs—just 2.6×10^{-6} of the host model's prefill cost for a 100-token prompt at L=32, d=4096. Total storage: 262 KB, versus ≈1.07 GB for full per-layer covariances (650× reduction) and ≈14 GB for a 7B guard model (53,000× reduction).

Key Equations

Harm trajectory. Let $\mathbf{h}_l \in \mathbb{R}^d$ be the last-token hidden state at layer l of an L-layer LLM processing prompt x. The scalar projection onto the per-layer LDA harm direction $\mathbf{v}_l$ yields the harm trajectory value at layer l:

$$p_l(x) = \left\langle \frac{\mathbf{h}_l}{\|\mathbf{h}_l\|}, \frac{\mathbf{v}_l}{\|\mathbf{v}_l\|} \right\rangle \in [-1,1], \qquad l = 1,\dots,L \tag{1}$$

Unit normalization is critical; without it, projections conflate semantic alignment with raw activation magnitude, which varies across layers and prompt lengths, degrading trajectory comparability by 1.3–1.8 F1.

Per-layer LDA direction. Each direction is the closed-form solution to the binary LDA problem, with Ledoit-Wolf shrinkage applied to the within-class scatter matrix $\mathbf{S}_W^{(l)}$ to ensure invertibility when $d \gg n_{\text{train}}$:

$$\mathbf{v}_l \propto \left(\hat{\mathbf{S}_W}^{(l)}\right)^{-1}\left(\boldsymbol{\mu}_l^{\text{harm}} - \boldsymbol{\mu}_l^{\text{safe}}\right), \quad \hat{\mathbf{S}_W}^{(l)} = (1-\alpha)\mathbf{S}_W^{(l)} + \frac{\alpha\,\mathrm{tr}(\mathbf{S}_W^{(l)})}{d}\mathbf{I} \tag{2}$$

where $\alpha \in [0,1)$ is the analytically computed shrinkage coefficient. LDA accounts for within-class covariance structure arising from prompt length, wording, and rhetorical variation—providing a more transferable discrimination axis than the raw mean difference direction.

7-D trajectory feature vector. The sequence $\{p_l\}_{l=1}^{L}$ is compressed into a compact, geometrically interpretable feature vector. The mean absolute curvature $\Delta^2\mathbf{p} = \frac{1}{L-2}\sum_{l=2}^{L-1}|p_{l+1} - 2p_l + p_{l-1}|$ measures trajectory smoothness, the monotonicity ratio $\mathrm{mono}(\mathbf{p}) = \frac{1}{L-1}\sum_{l=1}^{L-1}\mathbf{1}[p_{l+1} > p_l]$ measures consistency of directional increase; the onset layer $\hat{l}^* = \min\{l : p_l > \tau_{90}\}$ marks the first layer exceeding the 90th-percentile projection threshold:

$$\boldsymbol{\phi}(\mathbf{p}) = \left[\, p_L,\ \ \bar{p},\ \ p_L - p_1,\ \ \Delta^2\mathbf{p},\ \ \mathrm{mono}(\mathbf{p}),\ \ p_{\hat{l}^*},\ \ \hat{l}^* \,\right] \in \mathbb{R}^7 \tag{3}$$

MLP classifier. The feature vector is scored by a two-layer MLP with ReLU activation and sigmoid output, producing a final harmfulness probability:

$$\text{HERALD}(x) = \sigma\Big(\mathbf{W}_2\,\mathrm{ReLU}\big(\mathbf{W}_1\boldsymbol{\phi}(\mathbf{p}) + \mathbf{b}_1\big) + b_2\Big) \in [0,1] \tag{4}$$

The MLP has a hidden size of 32 (288 parameters total) and reaches the performance plateau at this capacity—larger architectures yield no statistically significant gain ($p > 0.05$, paired bootstrap), confirming that the bottleneck is feature quality rather than classifier expressivity. A prompt is flagged as harmful if $\text{HERALD}(x) \geq 0.5$.

Results

State-of-the-art average F1: HERALD achieves 89.3 F1 on OLMo2-7B across eight harmfulness benchmarks, outperforming latent-based baselines (Embed. Clf., Act. Delta) by 2.3–4.1 F1 on every backbone [2]. **Best-in-class jailbreak detection:** On WildJailbreak, HERALD reaches 98.4 F1, surpassing all four guard models (next-best 96.9). **Trajectory > terminal value:** Restricting to p_L alone drops performance by 1.5 F1 overall and 2.1 F1 on jailbreaks—the onset layer and monotonicity features capture signals invisible to single-layer methods. **Stable harm directions:** Per-layer LDA directions show pairwise cosine similarity > 0.97 across independent splits, confirming HPD reflects genuine geometric structure. **Data-efficient:** Plateaus near 1,000 samples/class; leads latent baselines by > 2 F1 at just 100 samples/class. **Negligible overhead:** 262 KB memory and 2.6×10^{-6} of prefill FLOPs, ∼53,000× smaller than a 7B guard model. **Limitations:** Trails' best guard models are on ToxicChat (−4.3 F1) and OpenAI Moderation (−8.4 F1), where harms are more implicit/diffuse.

Average F1 Across 8 Benchmarks

Method	BB	Avg F1
Latent-based		
Embed. Clf.	Llama	79.0
Act. Delta	Llama	82.0
HERALD	Llama	**86.3**
Embed. Clf.	OLMo2	84.7
Act. Delta	OLMo2	85.0
HERALD	OLMo2	**89.3**
Guard models		
Guard C	—	84.4
Guard D	—	87.8

HERALD surpasses all latent baselines on every backbone and rivals or beats guard models.

WildJailbreak F1 & Data Efficiency

Method	WJB	F1@100
Embed. Clf.	94.2	83.1
Act. Delta	91.0	81.0
HERALD	**98.4**	**83.7**
Guard C	96.9	—
Guard D	96.4	—
HD-WJB subset		
HERALD	**95.7**	—
Guard C	94.2	—
Guard D	94.8	—

Best jailbreak detection; HD-WJB confirms trajectory mechanism, not overlap, drives gains.

Trajectory Feature Ablation (OLMo2-7B)

Features	Avg	WJB
p_L only	87.8	96.3
$+\bar{p}$	88.1	96.6
+ rise	88.5	97.1
+ mono	88.9	97.6
+ curvature	89.1	97.9
$+\hat{l}^*$ **(full)**	**89.3**	**98.4**
Full $-\hat{l}^*$	88.4	97.1
Full − curvature	89.0	98.0
Full $-\bar{p}$	88.9	97.9

Onset layer $\hat{l}^*$ yields the largest single gain, especially on adversarial jailbreaks.

Memory & Computational Overhead

Method	Mem.	FLOPs
Full cov.	1.07 GB	high
Guard (7B)	14 GB	high
HERALD	**262 KB**	2.6×10^{-6}
Direction stability (pairwise cosine sim.)		
Llama-8B L8	0.991	
Llama-8B L32	0.989	
OLMo2-7B L8	0.988	
OLMo2-7B L32	0.991	

∼53,000× less memory than a 7B guard; LDA directions stable across splits (> 0.97).

Results and Plots

Main Results Across Backbones: HERALD consistently outperforms latent-based baselines across all four backbones, reaching its best average F1 of 89.3 on OLMo2-7B and surpassing guard models on adversarial jailbreaks.

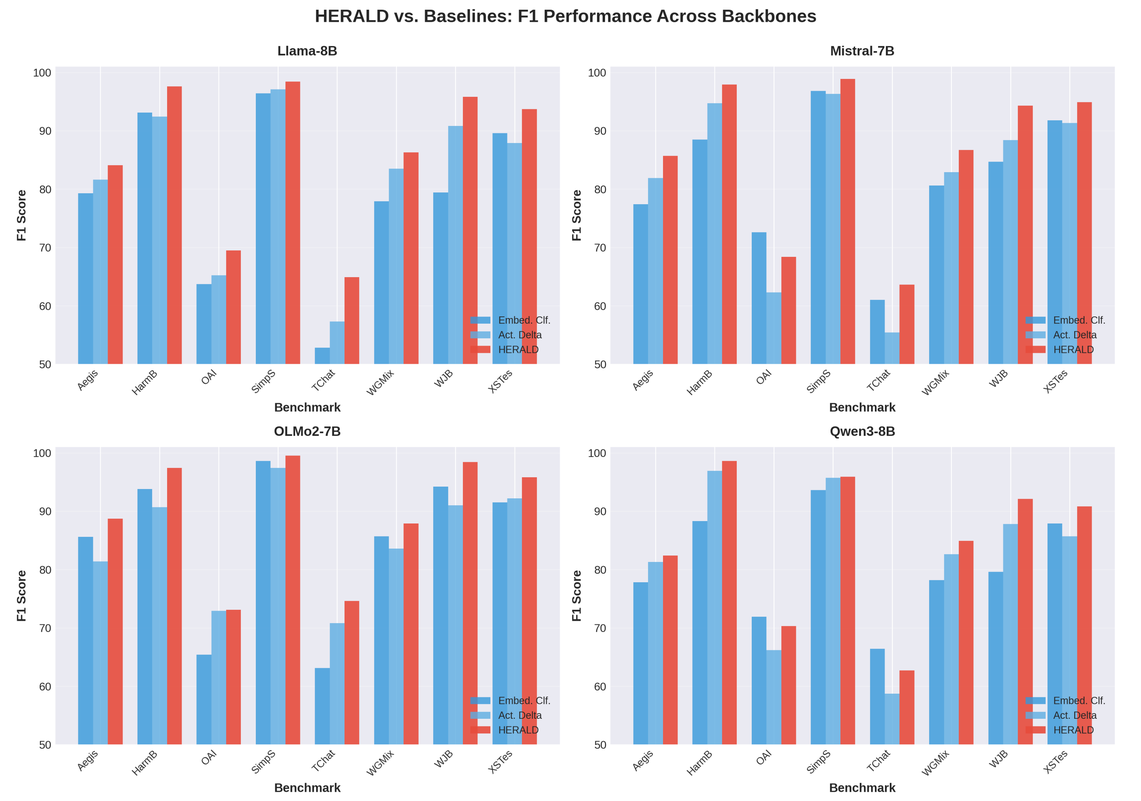


Figure 1. F1 across all eight benchmarks and four backbones.

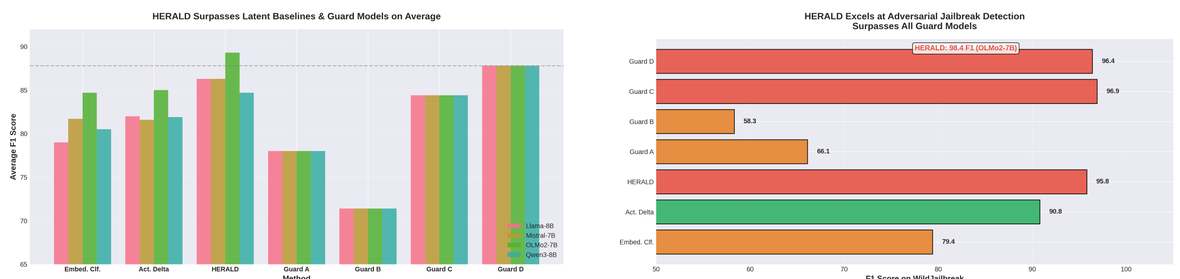
Method	Llama-8B	Mistral-7B	OLMo2-7B	Qwen3-8B
Embed. Clf.	79.0	81.7	84.7	80.5
Act. Delta	82.0	81.6	85.0	81.9
HERALD (ours)	86.3	86.3	89.3	84.7

Table 1. Average F1 over eight benchmarks. HERALD improves over latent-space baselines across all backbones.

Configuration	Llama-8B	Mistral-7B	OLMo2-7B
Random direction (control)	77.2	76.9	79.9
p_L only (no trajectory)	84.7	84.4	87.8
LDA + full trajectory (used)	86.3	86.3	89.3

Table 2. Ablation study. Both the LDA direction and trajectory features contribute to HERALD's performance.

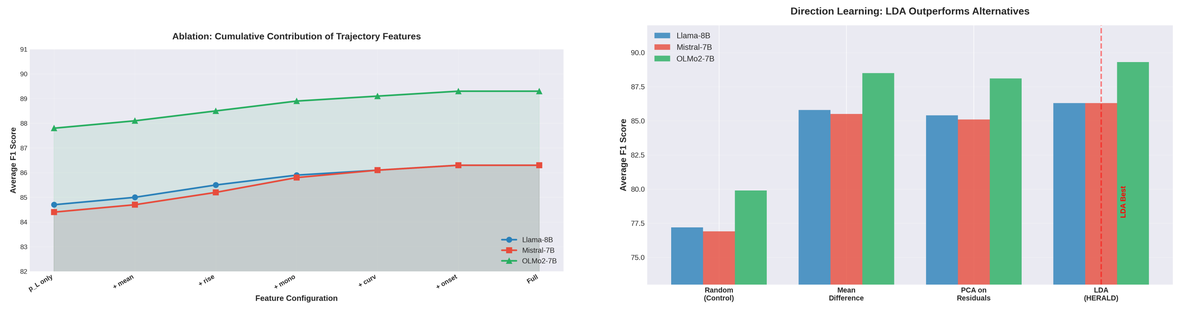
Ablation: Sources of the Gain: Trajectory shape and LDA-based direction learning drive most of the improvement; removing either causes the largest drops, while classifier capacity matters least.



(a) Average F1 score across all benchmarks.

(b) F1 score on the WildJailbreak benchmark.

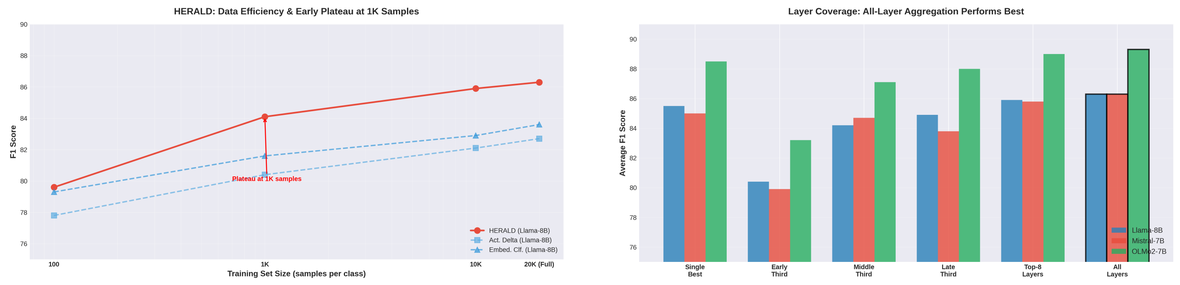
Figure 2. HERALD achieves state-of-the-art jailbreak detection while requiring only a fraction of the computational cost of large guard models. The structured trajectory representation is particularly effective when harmful intent emerges progressively over multiple turns, a setting in which single-layer detection methods are most limited.



(a) Trajectory feature ablation.

(b) Direction learning comparison.

Figure 3. Feature and direction-learning ablations. Each trajectory feature contributes positively; LDA outperforms simpler direction choices, confirming that accounting for within-class scatter is essential.



(a) Data efficiency.

(b) Layer coverage.

Figure 4. HERALD is data-efficient (plateau near 1,000 samples/class) and benefits from full-layer aggregation. The top-8 selection offers a compelling storage-accuracy trade-off.

Conclusions

We identified **Harmfulness Propagation Dynamics (HPD)**: harmful prompts exhibit a monotonically rising projection onto a learned harm direction across transformer depth, while benign prompts remain flat or oscillatory. **HERALD** operationalizes HPD via per-layer LDA directions, a compact 7-D trajectory feature vector, and a 288-parameter MLP—requiring no gradients and adding negligible overhead (262 KB, 2.6×10^{-6} of prefill FLOPs). Across four backbones and eight benchmarks, HERALD outperforms latent-based baselines by 2.3–4.1 F1 and surpasses all tested guard models on adversarial jailbreak detection (98.4 vs. 96.9 F1). Trajectory **shape**—onset layer, monotonicity, and curvature—carries information beyond the terminal value, especially for jailbreaks where harmful intent is revealed gradually. **Future work:** multi-directional subspaces for diffuse harms (e.g., social stereotypes), multilingual/multimodal extensions, and integration into RAG pipelines for flagging structurally ambiguous content.

References

[1] Olivier Ledoit and Michael Wolf.
A well-conditioned estimator for large-dimensional covariance matrices.
Journal of Multivariate Analysis, 88(2):365–411, 2004.

[2] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng.
A holistic approach to undesired content detection in the real world.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018, 2023.