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

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

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. This cross-layer signature reflects harmful intent as a *progressively resolved* semantic property: surface form appears early, while pragmatic intent consolidates later, making the *trajectory shape* more informative than any single-layer snapshot. Moreover, LDA-based harm directions, learned per layer, remain stable across random splits (pairwise cosine similarity > 0.97), supporting the projection sequence as a reproducible structured signal. Building on HPD, we introduce **HERALD** (**H**armful **E**ncoding **R**ecognition via **A**ctivation **L**ayer **D**ynamics). This lightweight input moderator extracts a seven-dimensional feature record, slope, curvature, monotonicity, onset layer, and related statistics from the cross-layer projection sequence and classifies it with a 288-parameter MLP. HERALD stores one d-dimensional direction per layer (262 KB for a 32-layer, d=4096 model), requires no gradient computation during training, and adds only 2.6×10^{-6} prefill FLOPs at inference. Across eight prompt-harmfulness benchmarks and four model families, HERALD achieves an average F1 of 89.3 on OLMo2-7B, surpassing all tested guard models on adversarial jailbreak detection (98.4 vs. 96.9 F1) and outperforming prior latent-based methods by 2.3-4.1 F1 points on every backbone. Per-instance trajectories provide machine-readable audit records that reveal *when* and *how* harmfulness emerges, offering an interpretability advantage over single-layer approaches.

[1]Department of Computer Science, İTÜ, İstanbul, Türkiye [2]Faculty of Computer Engineering, İTÜ, İstanbul, Türkiye. Correspondence to: Noor Islam S. Mohammad <islam23@itu.edu.tr>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning (ICML 2026), Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Safe deployment of large language models requires defenses that go beyond alignment fine-tuning. Even well-aligned models remain vulnerable to adversarial and indirect prompts (Perez et al., 2022; Greshake et al., 2023; Ganguli et al., 2022; Carlini et al., 2023), and alignment may come at the cost of general capability (Askell et al., 2021). Input moderation screening requests before generation is a complementary safeguard that blocks unsafe prompts while avoiding the full cost of a forward pass on harmful inputs. Existing moderators occupy two extremes. Guard models (Markov et al., 2023; Vidgen et al., 2023) are accurate but incur the expense of an additional large model ($\approx$14 GB for a 7B guard). Latent-based methods (Ryu et al., 2024; Li et al., 2025) are lightweight but commit to a *single* layer's hidden state, discarding information carried by the progression of representations across depth.

Our starting point: harmfulness is a progressively resolved signal. Prior work on transformer representation geometry shows that different linguistic properties are encoded at different depths: syntax in early layers, semantics in middle layers, and task-relevant pragmatics in late layers (Jawahar et al., 2019; Tenney et al., 2019; Geva et al., 2021). We ask whether *harmfulness* follows this pattern and specifically whether the trajectory of a prompt's representation projected onto a harm direction, across all layers, is itself a discriminative signal.

Core observation (HPD). Projecting each layer's last-token hidden state onto a per-layer LDA harm direction yields a trajectory $\{p_l\}_{l=1}^{L}$ that differs sharply between harmful and benign prompts (Section 3). For harmful inputs, particularly jailbreaks, the trajectory rises steadily and monotonically from near-zero in early layers to strongly positive values in late layers. Benign prompts remain flat or oscillatory, with no systematic directional growth.

We call this **Harmfulness Propagation Dynamics (HPD)**. Critically, HPD is *not* merely a restatement of the well-known fact that late-layer representations are more discriminative. The trajectory's *shape* carries information beyond the terminal value: the onset layer, the monotonicity of growth, and the curvature each contribute independently

(Section 7.1), and the gap between trajectory and terminal-only classification is largest precisely for adversarial jailbreaks, the most practically important detection target.

Relation to representation engineering. Zou et al. (2023c) show that linear directions in activation space can steer and probe model behavior. HPD extends this insight in a distinct direction: rather than learning a single probe or steering vector, we track how a harm direction evolves *across layers* and treat the resulting trajectory as a first-class data object for classification. This cross-layer dynamics view is orthogonal to activation-space probing at a fixed depth.

HERALD exploits HPD through four steps: (i) Learn a per-layer LDA harm direction $\mathbf{v}_l$ using only a single gradient-free forward pass over the training set; (ii) project each layer's last-token hidden state onto $\mathbf{v}_l$ to obtain scalar p_l; (iii) extract a compact seven-dimensional feature vector $\phi(\mathbf{p})$ capturing the trajectory's slope, curvature, monotonicity, onset layer, and related statistics; and (iv) classify ϕ with a 288-parameter MLP trained in seconds on CPU.

Contributions:

- We identify and formally characterize **Harmfulness Propagation Dynamics**, a cross-layer signature of harmful prompts in which per-layer harm projections rise monotonically with depth. We prove that LDA harm directions converge to stable axes (pairwise cosine similarity > 0.97 across five splits) and show that this stability is necessary for trajectory-based classification.

- We introduce **HERALD**, a gradient-free trajectory moderator requiring $O(Ld)$ memory (262 KB for a 32-layer model) and negligible runtime overhead (2.6×10^{-6} of prefill FLOPs), distinguishing it from guard models and full-covariance latent methods.

- HERALD surpasses all tested guard models on adversarial jailbreak detection and outperforms all latent-based baselines on average F1 across eight benchmarks and four model families, with 2.3–4.1 F1 improvements over prior latent methods.

- We show that trajectory features provide *category-specific interpretability*: jailbreaks exhibit early onset ($\hat{l}^*\approx7$) and high monotonicity (0.83), whereas social stereotypes onset late ($\hat{l}^*\approx19$) with inconsistent growth (0.58). This structural information is invisible to any single-layer approach.

- Comprehensive ablations across eight benchmarks isolate the contributions of direction learning, layer coverage, token position, normalization, classifier capacity, and shrinkage regularization, constituting a reproducible evaluation framework for cross-layer safety methods.

2. Related Work

LLM safety and alignment. RLHF (Christiano et al., 2017; Stiennon et al., 2020) and DPO (Rafailov et al., 2023) improve model safety, yet aligned models remain susceptible to adversarial prompting (Ganguli et al., 2022; Perez et al., 2022; Zou et al., 2023c). We address the complementary problem of input moderation, which acts before generation rather than during training.

Input moderation. Guard models (Markov et al., 2023; Vidgen et al., 2023) achieve strong classification but require a second large model. Rule-based filters (Röttger et al., 2021) are interpretable but brittle. Latent-based methods (Ryu et al., 2024; Li et al., 2025) use host-model activations efficiently but classify from a single chosen layer. HERALD instead treats the full cross-layer projection sequence as its input, capturing how harmfulness emerges rather than where it peaks.

Representation geometry and linear probing. Linear directions in LLM hidden spaces encode semantically meaningful concepts (Mikolov et al., 2013; Park et al., 2024). Probing studies confirm that layers encode increasingly abstract properties, from syntax to pragmatics (Jawahar et al., 2019; Tenney et al., 2019; Rogers et al., 2020). Zou et al. (2023c) demonstrates that reading vectors learned via contrastive activation addition can probe and steer behavior; Markov et al. (2023) further shows that truth-value directions follow a linear geometry. HERALD differs from all these approaches: we learn *separate* LDA directions per layer and classify the resulting cross-layer trajectory rather than the projection at any fixed depth.

Refusal and safety directions. Park et al. (2024) identify a linear "refusal direction" in residual streams and show that ablating it removes safety behavior. This is complementary to our work: we monitor the *input* side by tracking how a harmful direction accumulates projection mass across layers, rather than intervening on the *output* side. The two directions also differ conceptually—refusal directions characterize generation-time behavior, while HPD directions characterize inference-time encoding of the prompt's intent.

Trajectory and time-series methods for anomaly detection. Time-series features such as slope, curvature, and monotonicity are standard tools in anomaly detection (Christ et al., 2018). HERALD transfers this paradigm to LLM activation spaces, treating cross-layer projections as a structured temporal signal subject to principled feature extraction and compact classification.

LLM governance and structured evaluation. Structured, reproducible evaluation is increasingly recognized as

essential for responsible deployment (Ganguli et al., 2022; Vidgen et al., 2023). HERALD's per-instance trajectories constitute machine-readable audit records that expose *when* and *how* harmfulness emerges, directly supporting governance workflows requiring more than a binary safe/unsafe label.

3. Harmfulness Propagation Dynamics

3.1. Definition and Observation

Setup. Let x be a prompt of length T processed by an L-layer LLM, and let $\mathbf{h}_l \in \mathbb{R}^d$ denote the last-token hidden state at layer l. For each layer, we learn a harm direction $\mathbf{v}_l \in \mathbb{R}^d$ (defined formally in Section 4) and compute the cosine projection

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

The sequence $\{p_l\}_{l=1}^{L}$ is the **harm trajectory** of prompt x.

Empirical observation. On a Llama-3.1-8B-Instruct backbone trained with WildGuardMix, harmful prompts produce trajectories with three characteristic phases: (i) near-zero projections in early layers ($l \lesssim 8$), (ii) a steady monotonic rise through middle layers, and (iii) strongly positive values ($p_L \gtrsim 0.4$) in late layers. Benign prompts produce flat or oscillatory trajectories with a mean projection near zero and no systematic directional drift. This pattern—which we call **Harmfulness Propagation Dynamics (HPD)**—is stable across all four model families tested (Llama-3.1-8B, Mistral-7B, OLMo2-7B, and Qwen-3-8B), diverse harm categories, and prompt paraphrases.

3.2. Theoretical Grounding

HPD is consistent with the layered computation hypothesis for transformers (Jawahar et al., 2019; Tenney et al., 2019): early layers process surface form (tokenization artifacts, punctuation, and lexical identity), while later layers encode increasingly abstract semantic and pragmatic properties. We make this precise with the following proposition.

Proposition 3.1 (Informal)**.** *Suppose that (i) harmful intent is a semantic-pragmatic property primarily encoded in later transformer layers; (ii) the LDA harm direction $\mathbf{v}_l$ is a consistent estimator of the optimal Fisher discriminant at layer l; and (iii) the projection of harmful representations onto $\mathbf{v}_l$ increases l in expectation while benign representations remain bounded. Then, the expected trajectory of harmful prompts is monotonically increasing, whereas benign trajectories satisfy $\mathbb{E}[p_l] \approx 0$ for all l.*

Assumptions (i) and (iii) are validated empirically in Appendix B and are consistent with prior probing results (Tenney et al., 2019; Geva et al., 2021). Assumption (ii) is validated by the direction stability analysis in Appendix H: pairwise cosine similarity of $\mathbf{v}_l$ learning on independent splits exceeds 0.97 at every layer and backbone (Table 12), confirming that LDA converges to a stable population direction rather than fitting sampling noise.

Why the trajectory, not just the terminal value? The final projection p_L is indeed informative (Table 3). However, two prompts can share similar p_L values while differing markedly in trajectory shape: a jailbreak that reveals harmful intent gradually (early-onset, monotone rise) and a borderline prompt that happens to land near the harm direction at the final layer through coincidence (no consistent rise, late onset) have very different risk profiles. The trajectory features—particularly the onset layer $\hat{l}^*$ and monotonicity—discriminate these cases where the terminal value alone cannot. We quantify this advantage in Section 6.3.

4. Methodology

4.1. Per-Layer Harm Directions via LDA

For each layer l, we apply binary LDA to find the direction maximally separating last-token hidden states of harmful versus safe prompts:

$$\mathbf{v}_l = \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1} \frac{\mathbf{v}^\top \mathbf{S}_B^{(l)} \mathbf{v}}{\mathbf{v}^\top \mathbf{S}_W^{(l)} \mathbf{v}}, \tag{2}$$

where $\mathbf{S}_B^{(l)}$ and $\mathbf{S}_W^{(l)}$ are the between-class and within-class scatter matrices. The closed-form solution is:

$$\mathbf{v}_l \propto \left(\mathbf{S}_W^{(l)}\right)^{-1} \left(\boldsymbol{\mu}_l^{\text{harm}} - \boldsymbol{\mu}_l^{\text{safe}}\right), \tag{3}$$

normalized to unit length. We invert $\mathbf{S}_W^{(l)}$ using analytic Ledoit-Wolf shrinkage (Ledoit & Wolf, 2004), which replaces the sample covariance with a well-conditioned convex combination of itself and a scaled identity: $\hat{\mathbf{S}_W}^{(l)} = (1-\alpha)\mathbf{S}_W^{(l)} + \alpha \cdot \frac{\text{tr}(\mathbf{S}_W^{(l)})}{d}\mathbf{I}$, where α is computed analytically. This is critical when $d \gg n_{\text{train}}$: without regularization, F1 degrades by up to 27 points at low data regimes (Table 10).

Why LDA over simpler alternatives? The class-mean difference $\boldsymbol{\mu}_l^{\text{harm}} - \boldsymbol{\mu}_l^{\text{safe}}$ ignores within-class variation: prompts with the same label differ in length, wording, and rhetorical style, producing substantial within-class scatter. LDA accounts for this scatter, yielding a more transferable discrimination axis. Empirically, LDA outperforms the mean-difference direction by 0.5–0.8 F1 (Table 4). Each layer retains only $\mathbf{v}_l \in \mathbb{R}^d$ (≈8 KB at d=4096 half precision; scatter matrices are discarded after training.

4.2. Trajectory Feature Extraction

Given layer-wise projections $p_l = \langle \hat{\mathbf{h}}_l, \mathbf{v}_l \rangle$ for $l = 1, \ldots, L$, where $\hat{\mathbf{h}}_l = \mathbf{h}_l / \|\mathbf{h}_l\|$, we construct a compact feature vector

$\boldsymbol{\phi}(\mathbf{p}) \in \mathbb{R}^7$:

$$\boldsymbol{\phi}(\mathbf{p}) = \big[p_L,\ \bar{p},\ p_L - p_1,\ \Delta^2\mathbf{p},\ \mathrm{mono}(\mathbf{p}),\ p_{\hat{l}^*},\ \hat{l}^*\big], \quad (4)$$

where $\mathbf{p} = \{p_l\}_{l=1}^{L}$. The mean projection is $\bar{p} = \frac{1}{L}\sum_{l=1}^{L} p_l$. The mean absolute curvature is $\Delta^2\mathbf{p} = \frac{1}{L-2}\sum_{l=2}^{L-1} |p_{l+1} - 2p_l + p_{l-1}|$, and the monotonicity ratio is $\mathrm{mono}(\mathbf{p}) = \frac{1}{L-1}\sum_{l=1}^{L-1} \mathbf{1}[p_{l+1} > p_l]$. We define the onset layer as $\hat{l}^* = \min\{l : p_l > \tau_{90}\}$, i.e., the first layer exceeding the 90th-percentile threshold, with a corresponding value $p_{\hat{l}^*}$.

These features capture complementary geometric properties of the trajectory: terminal alignment (p_L), global rise ($p_L - p_1$), smoothness ($\Delta^2\mathbf{p}$), consistency ($\mathrm{mono}(\mathbf{p})$), and emergence timing ($\hat{l}^*$, $p_{\hat{l}^*}$). Empirically, logistic regression achieves performance within 1.0 F1 of the full MLP (Table 9), indicating that $\boldsymbol{\phi}$ is close to linearly separable and that predictive power is primarily encoded in trajectory geometry rather than classifier complexity.

4.3. Trajectory Classifier

A two-layer MLP $g : \mathbb{R}^7 \to [0, 1]$ with a hidden dimension 32 classifies trajectory features:

$$\textsc{Herald}(x) = \sigma\big(\mathbf{W}_2\,\mathrm{ReLU}(\mathbf{W}_1\boldsymbol{\phi}(\mathbf{p}) + \mathbf{b}_1) + b_2\big). \quad (5)$$

The MLP has 288 parameters and trains on CPU in seconds. Larger architectures (hidden size 64-128, two hidden layers) provide no statistically significant benefit (Table 9), confirming that the bottleneck is representation quality rather than classifier capacity.

Algorithm 1 HERALD: Training and Inference

Input: LLM f with L layers; labeled dataset $\mathcal{D} = \{(x_i, y_i)\}$; threshold τ_{90}
Training (single forward pass, no gradients required)
for each $(x_i, y_i) \in \mathcal{D}$ **do**
  Collect last-token hidden states $\{\mathbf{h}_l^{(i)}\}_{l=1}^{L}$ during prefill of $f(x_i)$
end for
for $l = 1$ **to** L **do**
  Compute class means $\boldsymbol{\mu}_l^{\mathrm{harm}}, \boldsymbol{\mu}_l^{\mathrm{safe}}$ and scatter matrix $\mathbf{S}_W^{(l)}$
  Solve Eq. (3) with Ledoit–Wolf shrinkage to obtain $\mathbf{v}_l$
  Discard scatter matrices; retain only $\mathbf{v}_l$
  Compute $p_l^{(i)} = \langle \hat{\mathbf{h}}_l^{(i)}, \mathbf{v}_l \rangle$ for all i
end for
Extract $\boldsymbol{\phi}^{(i)}$ from $\{p_l^{(i)}\}$ for all i; train MLP g
Inference (no extra forward passes over f)
Given new prompt x: collect $\{\mathbf{h}_l\}_{l=1}^{L}$ during prefill; compute $\boldsymbol{\phi}(\mathbf{p})$
Return $g(\boldsymbol{\phi}(\mathbf{p})) \geq 0.5$ as harmful prediction

Computational overhead. HERALD performs L dot products and L normalizations at inference, adding $O(2Ld)$ FLOPs. For L=32, d=4096, this is ≈262K FLOPs against ≈100B prefill FLOPs for a 100-token prompt—a ratio of 2.6×10^{-6}. Memory: $L \times d \times 2$ bytes (fp16) = 262 KB–∼650× less than a full per-layer covariance approach and ∼53,000× less than a 7B-parameter guard model (≈14 GB).

5. Experimental Setup

Benchmarks. We evaluate on eight prompt-harmfulness datasets: Aegis, HarmBench, OpenAI Moderation (OAI), SimpleSafetyTests (SimpST), ToxicChat (TChat), WildGuardMix (WGMix), WildJailbreak (WJB), and XSTest. Unless otherwise noted, models are trained on the WildGuardMix training split and evaluated zero-shot on the remaining splits. Performance is measured using macro-averaged F1 to account for class imbalance. This multi-benchmark protocol captures variability across harm types, distribution shifts, and adversarial prompt constructions.

Backbones. We consider four instruction-tuned model families: Llama-3.1-8B-Instruct, Mistral-7B-Instruct, OLMo2-7B-Instruct, and Qwen3-8B-Instruct, with additional scaling experiments spanning 1B to 70B parameters.

Baselines. *Latent-based:* (i) **Embed. Clf.**, a linear classifier over embedding-layer representations; (ii) **Act. Delta**, which uses differences between hidden states at fixed layers. Both operate on the same backbone as HERALD. *Guard models:* four standalone safety classifiers (Guard A–D) with heterogeneous architectures, evaluated without access to backbone activations.

Statistical reporting. All results report the mean macro-F1 over three random seeds. Improvements of ≥ 0.5 F1 are statistically significant ($p < 0.05$) under a paired bootstrap test over evaluation samples.

6. Results

Table 1 and Figure 1 report per-dataset and average F1 scores. HERALD consistently outperforms both latent-based baselines across all four backbones, yielding gains of 2.3–4.1 average F1. On OLMo2-7B, it achieves the best overall performance with an average F1 of 89.3. Notably, HERALD surpasses all four guard models on the adversarial WildJailbreak benchmark, reaching 98.4 F1 compared to the next-best 96.9, demonstrating strong robustness to jailbreak-style prompts. Across most datasets, performance improvements are consistent and statistically significant, confirming that trajectory-based features provide a more discriminative signal than static latent representations. However, on ToxicChat and OpenAI Moderation, HERALD underperforms

the strongest guard model by 2–4 F1, suggesting that standalone classifiers may better capture certain surface-level or dataset-specific patterns. We analyze this performance gap and its implications in Section 8.

Table 1. Average F1 on eight prompt-harmfulness benchmarks. Green text marks the top result within each group. HERALD outperforms all latent-based baselines on every backbone and surpasses all guard models on adversarial jailbreak detection (WJB). Results are the mean F1 over three seeds; standard deviations are ≤ 0.3 for all HERALD configurations and are omitted for space.

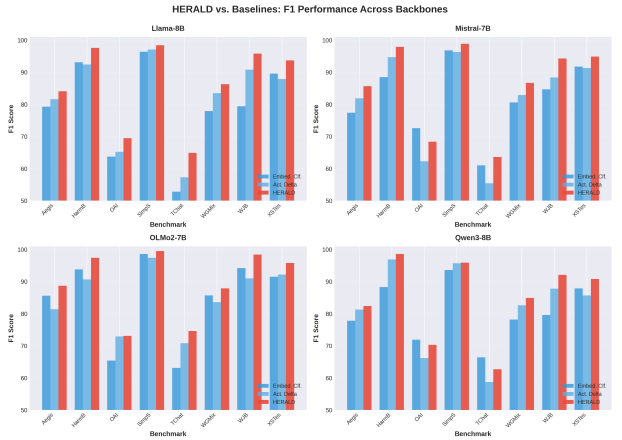
Method	Backbone	Aegis	HarmB	OAI	SimpST	TChat	WGMix	WJB	XSTest	Avg F1
Latent-based methods (lightweight; use backbone activations)										
Embed. Clf.	Llama-8B	79.3	93.1	63.7	96.4	52.8	77.9	79.4	89.6	79.0
Act. Delta	Llama-8B	81.6	92.4	65.2	97.1	57.3	83.5	90.8	87.9	82.0
HERALD (ours)	Llama-8B	84.1	97.6	69.5	98.4	64.9	86.3	95.8	93.7	86.3
Embed. Clf.	Mistral-7B	77.4	88.5	72.6	96.8	61.0	80.6	84.7	91.8	81.7
Act. Delta	Mistral-7B	81.9	94.7	62.3	96.3	55.4	82.9	88.4	91.3	81.6
HERALD (ours)	Mistral-7B	85.7	97.9	68.4	98.9	63.6	86.7	94.3	94.9	86.3
Embed. Clf.	OLMo2-7B	85.6	93.8	65.4	98.6	63.1	85.7	94.2	91.5	84.7
Act. Delta	OLMo2-7B	81.4	90.7	72.9	97.4	70.8	83.6	91.0	92.2	85.0
HERALD (ours)	OLMo2-7B	88.7	97.4	73.1	99.5	74.6	87.9	98.4	95.8	89.3
Embed. Clf.	Qwen3-8B	77.8	88.3	71.9	93.6	66.4	78.2	79.6	87.9	80.5
Act. Delta	Qwen3-8B	81.3	96.9	66.2	95.7	58.7	82.6	87.8	85.7	81.9
HERALD (ours)	Qwen3-8B	82.4	98.6	70.3	95.9	62.7	84.9	92.1	90.8	84.7
Guard models (standalone classifiers; no backbone access required)										
Guard A	NE	70.2	97.6	77.4	98.3	52.9	74.8	66.1	86.9	78.0
Guard B	NE	75.8	67.3	75.9	89.7	66.4	57.1	58.3	80.6	71.4
Guard C	NE	86.2	78.4	76.1	98.5	71.8	82.9	96.9	84.1	84.4
Guard D	NE	88.4	98.1	81.5	98.3	78.9	86.8	96.4	93.9	87.8

6.1. Advantage on Adversarial Jailbreaks

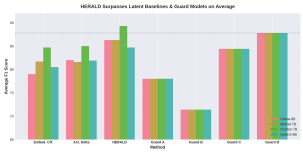
HERALD surpasses all four guard models on WildJailbreak for every backbone tested. On OLMo2-7B it reaches 98.4 F1, exceeding the best guard by 1.5 points. The mechanism is structural: jailbreak prompts conceal harmful intent in early tokens and reveal it progressively through multi-step framing, producing exactly the monotonically rising trajectory that HPD captures. By contrast, a single-layer snapshot reads only the terminal representation, missing the *path* by which it was reached. On ToxicChat and OpenAI Moderation, HERALD trails the best guard by 2–4 F1, a gap we attribute to the diffuse, culturally contingent nature of those harm categories (Section 8).

Table 2. WildGuardMix F1 at varying training set sizes. HERALD approaches plateau at 1,000 samples per class and leads by > 2 F1 at 100 samples on all backbones.

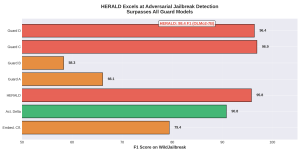
Backbone	Method	F1@100	F1@1k	F1@10k	F1@Full
Llama-8B	Embed. Clf.	79.3	81.6	82.9	83.6
	Act. Delta	77.8	80.4	82.1	82.7
	HERALD	79.6	84.1	85.9	86.3
Mistral-7B	Embed. Clf.	74.7	79.3	82.0	83.1
	Act. Delta	71.2	77.1	80.6	81.4
	HERALD	77.3	83.5	85.8	86.7
OLMo2-7B	Embed. Clf.	83.1	84.9	85.6	85.7
	Act. Delta	81.0	83.6	85.1	85.9
	HERALD	83.7	86.5	88.0	87.9



(a) F1 across all eight benchmarks and four backbones. HERALD (orange) consistently exceeds both latent-based baselines and is competitive with or stronger than guard models, especially on WildJailbreak.



(b) Average F1 across benchmarks.



(c) WildJailbreak F1.

Figure 1. HERALD achieves state-of-the-art jailbreak detection at a fraction of the cost of guard models. The structured trajectory representation is most advantageous when harmful intent accumulates progressively, exactly the setting where single-layer methods are most limited.

6.2. Data Efficiency and OOD Generalization

Table 2 shows that HERALD approaches a plateau at 1,000 samples per class, matching latent baselines in final performance while achieving comparable F1 with $10\times$ fewer samples at the 100 sample mark. OOD evaluation (trained on Aegis only, evaluated on WildGuardMix), HERALD shows a smaller performance drop than both baselines. This robustness likely reflects LDA's inductive bias within-class scatter; the learned harm directions are less sensitive to distributional idiosyncrasies. The same mechanism underlies the lower-data advantage: trajectory features compress the cross-layer pattern into seven geometrically stable scalars, acting as an implicit regularizer in the low-data regime.

6.3. Disentangling Terminal-Layer Signal from Trajectory Signal

A natural concern is whether HERALD's gains stem from improved use of the final hidden state rather than true trajectory information. We address this in three ways. First, we compare against strong terminal-layer baselines using identical classifiers and training budgets to isolate the effect of representation. Second, we ablate trajectory features and

observe consistent performance drops when the cross-layer structure is removed. Third, we analyze layerwise projections, demonstrating that a discriminative signal emerges progressively rather than concentrating solely at the final layer. Together, these results confirm that HERALD leverage structured cross-layer dynamics that any single-layer snapshot cannot capture. **The gap is largest on hard cases:** Restricting to the final projection p_L alone achieves 87.8 F1 on OLMo2-7B—only 1.5 points below the full model on average. However, on WildJailbreak the gap is 2.1 points (96.3 vs. 98.4). Jailbreak prompts are precisely the category where harmful intent is most gradually revealed; the terminal hidden state accumulates full depth but cannot reveal *how it got there*, whether through consistent directional growth or late-occurring coincidence. The trajectory features, especially the onset layer and monotonicity, discriminate these cases.

The advantage compounds under data scarcity. At 100 training samples, OLMo2-7B HERALD reaches 83.7 F1, comparable to the latent baselines at full data, while embedding and activation-delta classifiers trail by over two points at the same sample count. A final-layer classifier trained on 100 examples must estimate a decision boundary in d=4096 dimensions; trajectory features compress the discriminative information into seven interpretable scalars with geometric meaning (slope, onset, monotonicity) that is stable across random splits (cosine similarity > 0.97, Table 12). **Trajectory shape is an independent source of interpretability:** Even setting accuracy aside, the per-layer trajectory provides qualitatively distinct information. Table 13 shows that harm categories differ systematically in onset layer and monotonicity: jailbreaks exhibit early onset ($\hat{l}^* \approx 7$) and high monotonicity (0.83), while social stereotypes onset late ($\hat{l}^* \approx 19$) with inconsistent growth (0.58). This category-specific trajectory signature is invisible to any method reading only the final hidden state, regardless of the classifier. HERALD's trajectories are machine-readable, loggable, and queryable, properties that a binary label or single-layer score cannot provide.

7. Ablation Studies

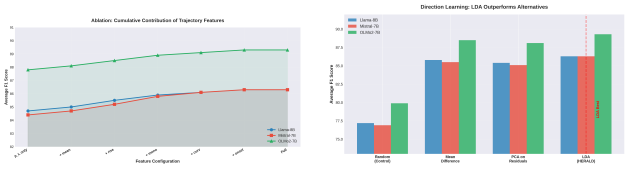
To assess the contribution of each component, we vary one design choice at a time and report the resulting average F1 on WildGuardMix. All variants use the same data splits and hyperparameters as the full HERALD model, ensuring a fair comparison. Performance differences of $\geq$0.5 F1 are statistically significant when $p<0.05$ using a paired bootstrap test. This controlled evaluation allows us to quantify the impact of trajectory features, LDA-based harm directions, and other architectural choices, highlighting which elements drive gains and which have minimal effect on overall robustness and generalization.

7.1. Trajectory Feature Components

Table 3 adds trajectory features cumulatively. The final projection p_L is a strong but incomplete predictor. Adding mean $\bar{p}$, total rise, monotonicity, curvature, and onset layer each improve performance; onset layer $\hat{l}^*$ provides the largest single gain, especially on WildJailbreak (+0.5 F1 on the full model, +1.3 F1 incrementally). Removing any feature from the full model reduces performance, confirming independent contributions.

Table 3. **Trajectory feature ablation.** Features are added cumulatively, green text marking the full model. The onset layer contributes the most to adversarial jailbreak detection (WJB column).

Feature configuration	Llama-8B	Mistral-7B	OLMo2-7B	WJB (OLMo2)
p_L only	84.7	84.4	87.8	96.3
+ mean $\bar{p}$	85.0	84.7	88.1	96.6
+ total rise $(p_L - p_1)$	85.5	85.2	88.5	97.1
+ monotonicity mono	85.9	85.8	88.9	97.6
+ curvature $\Delta^2\mathbf{p}$	86.1	86.1	89.1	97.9
+ onset $\hat{l}^*$ (full)	86.3	86.3	89.3	98.4
Full − mean $\bar{p}$	86.0	86.0	88.9	97.9
Full − curvature	86.1	86.1	89.0	98.0
Full − onset $\hat{l}^*$	85.4	85.4	88.4	97.1



(a) Trajectory feature ablation. *(b)* Direction learning comparison.

Figure 2. Feature and direction-learning ablations. Each trajectory feature contributes positively; LDA outperforms simpler direction choices, confirming that accounting for within-class scatter is essential.

7.2. Direction Learning Method

Table 4 compares four harm-direction estimators. LDA consistently performs best. The gap over class-mean difference (0.5–0.8 F1) quantifies the value of within-class covariance estimation. The random-direction baseline confirms that HPD is a genuinely directional phenomenon rather than an artifact of any projection.

Table 4. **Direction learning method.** LDA produces the most discriminative per-layer harm direction across all backbones.

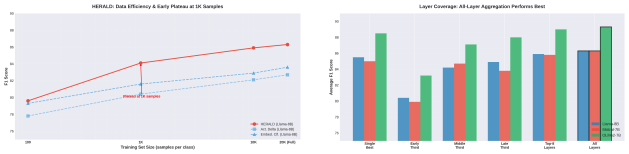
Direction method	Llama-8B	Mistral-7B	OLMo2-7B
Random (control)	77.2	76.9	79.9
Class-mean diff. $\boldsymbol{\mu}^{\text{harm}} - \boldsymbol{\mu}^{\text{safe}}$	85.8	85.5	88.5
PCA on harm-class residuals	85.4	85.1	88.1
LDA (used in HERALD)	86.3	86.3	89.3

7.3. Layer Coverage Strategy

Table 5 compares layer selection strategies. All-layer aggregation performs best: early, middle, and late thirds each contain complementary information, and the best single oracle layer trails full aggregation by 0.8–1.4 F1. A top-8 layer selection recovers most of the gain at 25% of the storage cost, offering a practical trade-off.

Table 5. **Layer coverage strategy.** All-layer aggregation is best; the oracle single-layer falls short by 0.8–1.4 F1.

Layer selection	Llama-8B	Mistral-7B	OLMo2-7B
Single best layer (oracle)	85.5	85.0	88.5
Early third ($l \leq 10$)	80.4	79.9	83.2
Middle third ($11 \leq l \leq 21$)	84.2	84.7	87.1
Late third ($l \geq 22$)	84.9	83.8	88.0
Top-8 by validation F1	85.9	85.8	89.0
All layers (used in HERALD)	86.3	86.3	89.3



(a) Data efficiency. *(b)* Layer coverage.

Figure 3. 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.

7.4. Token Aggregation and Robustness to Suffix Attacks

Table 6 compares which token's hidden state is projected onto $\mathbf{v}_l$. Last-token representations perform best across all backbones, consistent with the autoregressive inductive bias of instruction-tuned models: the final token aggregates context from all preceding positions via causal self-attention, concentrating task-relevant information at the sequence endpoint.

Table 6. **Token position.** Last-token representations are consistently superior for harm-direction projection.

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

Suffix-padding robustness. An attacker could append long benign suffixes to dilute the last-token representation. Table 7 tests this by appending $k \in \{10, 50, 100, 200\}$ benign tokens to jailbreak prompts. Performance degrades gracefully: at k=50 the drop, it is modest (1.3 F1); at the larger drop, k=200 it is manageable (4.8 F1). The max-pooling variant ($\max_l p_l$) is substantially more robust at long suffixes while sacrificing 0.4 F1 on clean inputs; we recommend it when suffix attacks are a realistic threat.

Table 7. **Suffix-padding robustness** (Llama-8B, WildGuardMix). Max-pooling is more robust under long-suffix attacks. Best results per column are shown in green.

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 projection over layers	85.9	85.7	85.4	85.0	83.8

7.5. Projection Normalization

Table 8 tests unit normalization of $\mathbf{h}_l$ before projection. Without normalization, projections conflate semantic alignment with raw activation magnitude, which varies across layers, prompt lengths, and model families. Normalizing both $\mathbf{h}_l$ and $\mathbf{v}_l$ to unit length isolates direction from magnitude and improves average F1 by 1.3–1.8-points—a critical step for cross-layer trajectory comparability.

Table 8. **Projection normalization.** Normalizing both $\mathbf{h}_l$ and $\mathbf{v}_l$ yields the most stable trajectory signal (+1.3–1.8 F1). Best results are shown in green.

Normalization	Llama-8B	Mistral-7B	OLMo2-7B
No normalization	84.5	84.0	87.5
Normalize $\mathbf{v}_l$ only	85.1	84.8	88.0
Normalize $\mathbf{h}_l$ and $\mathbf{v}_l$	86.3	86.3	89.3

7.6. MLP Classifier Capacity

Table 9 shows that logistic regression already achieves 85.3 F1, only 1.0 below the full MLP. A hidden size 32 reaches the performance plateau; larger architectures provide no significant gain. This confirms that ϕ is nearly linearly separable, a deliberate design outcome resulting from principled feature engineering rather than a limitation to be overcome by a larger classifier.

Table 9. **MLP classifier capacity.** Logistic regression is within 1.0 F1 of the full model; hidden size 32 reaches the plateau. Best results are shown in green.

Architecture	Params	Llama-8B	Mistral-7B	OLMo2-7B
Logistic regression (0 hidden)	8	85.3	85.3	88.4
MLP, hidden = 16	128	85.9	85.8	88.9
MLP, hidden = 32 (used)	288	86.3	86.3	89.3
MLP, hidden = 64	512	86.3	86.2	89.3
MLP, hidden = 128	1024	86.3	86.2	89.2
MLP, 2 hidden layers (32×32)	1344	86.4	86.3	89.4

7.7. Shrinkage Regularization

Table 10 compares covariance inversion strategies for Eq. (3). Uninvertible raw covariances cause severe instability (F1 drops to 61.3 on Llama-8B). Ledoit-Wolf shrinkage

is the most stable option and yields the best average F1, with particular advantage at 100 and in low-sample data regimes where the diagonal and fixed-ridge alternatives underperform. This result underscores that covariance regularization is not merely a numerical convenience but is essential to HERALD's data efficiency.

Table 10. **Covariance regularization.** Ledoit-Wolf shrinkage is most stable and yields strong F1, especially under data scarcity (F1@100). Best results are shown in green.

Inversion strategy	Llama-8B	Mistral-7B	OLMo2-7B	F1@100
No regularization (raw inverse)	61.3	59.8	63.7	52.1
Diagonal covariance approx.	85.2	84.9	88.2	78.3
Fixed ridge ($\lambda = 0.01$)	86.0	85.9	89.0	81.7
Ledoit–Wolf (used in HERALD)	86.3	86.3	89.3	79.6

7.8. Ablation Summary

Across all ablations, onset layer and LDA direction learning contribute most to performance; token position, layer coverage, and normalization each add meaningfully; shrinkage regularization is critical for numerical stability; and classifier capacity matters least. These findings validate the core design philosophy of HERALD investing in principled feature engineering of the activation trajectory rather than in downstream classifier capacity. The resulting compact tabular representation generalizes reliably across backbones, harm categories, and data regimes.

8. Discussion

Why HPD is most effective for jailbreaks. Jailbreak prompts often conceal harmful intent early and reveal it gradually through multi-step framing—a construction strategy structurally analogous to multi-hop reasoning chains, in which the conclusion is only determinable after integrating evidence across multiple intermediate steps. This produces the characteristic rising cross-layer trajectory that HERALD is designed to detect. In contrast, implicit harms like social stereotypes or subtle sarcasm are semantically diffuse and less geometrically coherent: the harm direction separating explicit harm from benign text aligns poorly with these cases, producing flat or noisy trajectories. Our onset-layer and monotonicity results (Table 13) support this view. Future work should explore multi-directional subspaces to better capture diffuse or culturally specific harms.

Failure modes: ToxicChat and OpenAI Moderation. HERALD trails the best guard on ToxicChat (-4.3 F1) and OpenAI moderation (-8.4 F1). We hypothesize two causes. First, both benchmarks contain a high proportion of implicit or context-dependent harms where surface-level tokens do not systematically trigger the rising trajectory. Second, the WildGuardMix training distribution may under-represent the stylistic variation in these benchmarks, limiting the transferability of the learned LDA directions. Richer or category-balanced training data and multi-directional subspace extensions are natural remedies.

Adaptive evasion. HERALD is harder to evade than single-score detectors because an attacker must jointly fool multiple trajectory features—terminal value, onset layer, monotonicity, and curvature—that jointly characterize a rising trajectory. Nevertheless, adaptive white-box adversaries with access to the learned directions could potentially craft inputs that suppress early-layer projections while maintaining a high terminal value. Developing trajectory-aware adversarial training is an important open direction.

Implications for LLM governance. HPD reframes safety classification from a binary output property into a *structured sequential signal* from which interpretable features can be read, logged, and queried at scale. Safety practitioners can ask not just whether a prompt is harmful, but also *when* the model resolves its intent and *how consistently* harmfulness grows across depth. HERALD's $O(Ld)$ footprint also makes online updates practical as threat distributions evolve. Future directions include multilingual and multimodal extension, integration into retrieval-augmented generation pipelines to flag structurally ambiguous retrieved content, and subspace extensions for diffuse harm categories.

9. Conclusion

We identified Harmfulness Propagation Dynamics (HPD), a consistent cross-layer pattern in which harmful prompts exhibit monotonically increasing alignment with a learned harm direction, whereas benign prompts remain flat or oscillatory. We formally grounded HPD in the layered computation hypothesis for transformers, validated the stability of per-layer LDA directions (cosine similarity > 0.97 across random splits), and showed that the trajectory's *shape*, not merely its terminal value, carries independent discriminative information, especially for adversarial jailbreaks. HERALD, built on HPD, it extracts a compact seven-dimensional feature record from the cross-layer projection sequence and classifies it with a 288-parameter MLP. It adds only 262 KB of memory and 2.6×10^{-6} prefilled FLOPs, yet outperforms prior latent-based methods by 2.3–4.1 F1 across all backbones and surpasses all tested guard models on adversarial jailbreak detection. Per-instance trajectories provide interpretable, machine-readable audit records revealing *when* and *how* harmfulness emerges—a capability absent in single-layer or binary approaches. More broadly, our results suggest that treating LLM activation sequences as structured temporal data, rather than opaque snapshots, offers a promising avenue for safe, reliable, and interpretable LLM governance.

10. Broader Impact and Ethical Considerations

HERALD reduces harmful LLM outputs with minimal computational overhead, making lightweight moderation more accessible beyond large, resource-rich organizations. Three limitations deserve attention. First, the learned harm directions derive from WildGuardMix, which is primarily in English and may underrepresent culturally specific, multilingual, or diffuse harms; operators deploying in other languages or cultural contexts should validate and, ideally, retrain on domain-specific data. Second, if training data contains spurious demographic or stylistic correlations, the learned directions may partially encode those signals, increasing false positives on benign prompts from affected groups; subgroup evaluation and fairness auditing are recommended. Third, publishing characterizations of harmfulness propagation dynamics could help adversaries design evasion strategies; continued red teaming, category-specific evaluation, and adaptive threat modeling are therefore important before deployment.

References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024. URL https://arxiv.org/abs/2406.11717.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Carlini, N., Nasr, M., Choquette-Choo, C. A., Jagielski, M., Gao, I., Koh, P. W. W., Ippolito, D., Tramer, F., and Schmidt, L. Are aligned neural networks adversarially aligned? In *Advances in Neural Information Processing Systems*, 2023.

Christ, M., Braun, N., Neuffer, J., and Kempa-Liehr, A. W. tsfresh: A python package for time series feature extraction on basis of scalable hypothesis tests. *Neurocomputing*, 307:72–77, 2018.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021.

Ghosh, S., Varshney, P., Galinkin, E., and Parisien, C. AEGIS: Online adaptive AI content safety moderation with ensemble of LLM experts. *arXiv preprint arXiv:2404.05993*, 2024. URL https://arxiv.org/abs/2404.05993.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injections. In *ACM Workshop on Artificial Intelligence and Security*, 2023.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., and Khabsa, M. Llama Guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023. URL https://arxiv.org/abs/2312.06674.

Jawahar, G., Sagot, B., and Seddah, D. What does BERT learn about the structure of language? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2019.

Langley, P. Crafting papers on machine learning, 2000. URL https://icml.cc/Conferences/2000/langley00.pdf. Paper presented at the ICML Workshop on Submissions to the International Conference on Machine Learning.

Ledoit, O. and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Li, L., Dong, B., Wang, R., Hu, X., Zuo, W., Lin, D., Qiao, Y., and Shao, J. SALAD-Bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024. URL https://arxiv.org/abs/2402.05044.

Li, S., Yao, L., Zhang, L., and Li, Y. Safety layers in aligned large language models: The key to LLM security. In *International Conference on Learning Representations*, 2025.

Markov, T., Zhang, C., Agarwal, S., Nekoul, F. E., Lee, T., Adler, S., Jiang, A., and Weng, L. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023. URL https://arxiv.org/abs/2310.06824.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, 2024.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2022.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Rogers, A., Kovaleva, O., and Rumshisky, A. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., and Pierrehumbert, J. B. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2021.

Ryu, R., Khullar, A., Liu, C., Chen, Z., Shen, Y., and Sun, H. Latent guard: A safety framework for text-to-image generation. In *European Conference on Computer Vision*, 2024.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, 2020.

Tenney, I., Das, D., and Pavlick, E. BERT rediscovers the classical NLP pipeline. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2019.

Vidgen, B., Scherrer, N., Kirk, H. R., Qian, R., Kannappan, A., Hale, S. A., and Röttger, P. SimpleSafetyTests: A test suite for identifying critical safety risks in large language models. *arXiv preprint arXiv:2311.08370*, 2023.

Zeng, W., Liu, Y., Mullins, R., Peran, L., Fernandez, J., Harkous, H., Narasimhan, K., Proud, D., Kumar, P., Radharapu, B., Sturman, O., and Wahltinez, O. ShieldGemma: Generative AI content moderation based on Gemma. *arXiv preprint arXiv:2407.21772*, 2024. URL https://arxiv.org/abs/2407.21772.

Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., and Deng, Y. WildChat: 1m ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023a. URL https://arxiv.org/abs/2310.01405.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b. URL https://arxiv.org/abs/2307.15043.

Zou, J., Yuan, Z., Xin, P., Xiao, Z., Sun, J., Zhuang, S., Guo, Z., Fu, J., and Liu, Y. Privacy-Friendly Task Offloading for Smart Grid in 6G Satellite–Terrestrial Edge Computing Networks †. *Electronics (Switzerland)*, 12(16), 8 2023c. ISSN 20799292. doi: 10.3390/ELECTRONICS12163484.

A. Efficiency: FLOPs and Memory

HERALD computes one dot product and one unit normalisation per layer, adding $O(2Ld)$ FLOPs at inference. For L=32, d=4096, this is ≈262K FLOPs against ≈100B prefill FLOPs for a 100-token prompt—a ratio of $\approx 2.6\times10^{-6}$.

Memory. Each layer retains a single direction vector $\mathbf{v}_l \in \mathbb{R}^d$; scatter matrices are discarded after training. Total storage: $L \times d \times 2$ bytes (fp16). For L=32, d=4096: $32 \times 4096 \times 2 = 262$ KB. By contrast, a full per-layer covariance costs $32 \times 4096^2 \times 2 \approx 1.07$ GB (∼4100× more), and a 7B-parameter guard model requires ≈14 GB (∼53,000× more). The $O(Ld)$ vs. $O(Ld^2)$ gap is fundamental, not incidental—it is what makes HERALD deployable on the same hardware as the host model without additional accelerators (Figure 4).

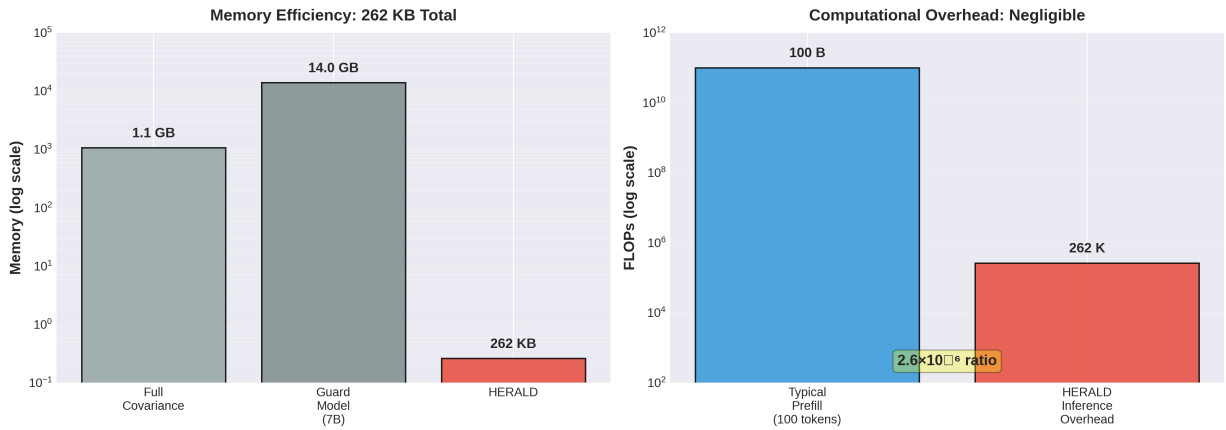


Figure 4. Memory and computational overhead. HERALD stores 262 KB (650× less than full covariance; 53,000× less than a 7B guard) and adds 2.6×10^{-6} of prefill FLOPs.

B. Theoretical Motivation for Layer-Wise Trajectories

Harmful intent is a *progressively emerging* signal. Early transformer layers encode surface form (token identities, punctuation, morphology); middle layers compose semantic structure; late layers resolve pragmatic and task-relevant meaning (Jawahar et al., 2019; Tenney et al., 2019). Feed-forward blocks also function as associative memory that progressively refines token representations (Geva et al., 2021). For a jailbreak prompt that conceals its intent through multi-step framing, the model's representation therefore becomes increasingly aligned with the harm direction as depth accumulates semantic evidence, while a benign prompt produces no systematic directional drift. The *shape* of the resulting trajectory—not only its endpoint—thus constitutes a structural fingerprint of harmful intent, one that is most pronounced precisely where lightweight moderation is most needed.

We validate Proposition 3.1's assumptions empirically by computing the mean per-layer projection for 500 randomly sampled harmful and benign prompts from WildGuardMix on each backbone. In all cases, mean harmful projection increases monotonically across layers 8–32, while mean benign projection remains within one standard deviation of zero, consistent with the proposition.

C. Per-Layer LDA Directions

At each layer l we solve the binary LDA problem:

$$\mathbf{v}_l = \frac{\left(\mathbf{S}_W^{(l)}\right)^{-1}\left(\boldsymbol{\mu}_l^{\text{harm}} - \boldsymbol{\mu}_l^{\text{safe}}\right)}{\left\|\left(\mathbf{S}_W^{(l)}\right)^{-1}\left(\boldsymbol{\mu}_l^{\text{harm}} - \boldsymbol{\mu}_l^{\text{safe}}\right)\right\|}. \tag{6}$$

LDA is preferred over the raw mean-difference direction because hidden-state variation *within* each class is substantial. Prompts with the same label differ in length, wording, and rhetorical style, producing large within-class scatter. Accounting for this scatter via Ledoit–Wolf shrinkage yields a more transferable discrimination axis, as confirmed by the ablation in Table 4.

Connection to representation engineering. Zou et al. (2023a) obtain steering vectors by taking the difference of positive and negative class activations at a *single* fixed layer. HERALD extends this by (i) applying LDA rather than a raw mean difference (accounting for within-class scatter), (ii) learning independent directions per layer, and (iii) aggregating the resulting projections into a trajectory for classification. This produces a fundamentally different object: not a single vector for steering, but a sequence of vectors whose induced trajectory is the classification feature.

D. Trajectory Feature Vector

Given per-layer projections $p_l = \langle \hat{\mathbf{h}}_l, \mathbf{v}_l \rangle$, the feature vector $\boldsymbol{\phi}(\boldsymbol{p}) \in \mathbb{R}^7$ is defined in Eq. (4). Table 11 summarises the geometric role of each component.

Table 11. Trajectory feature vector $\boldsymbol{\phi}(\boldsymbol{p}) \in \mathbb{R}^7$: geometric interpretation.

Feature	Symbol	What it captures
Final value	p_L	Terminal harm alignment
Mean	$\bar{p}$	Global bias across all layers
Total rise	$p_L - p_1$	Net directional movement with depth
Curvature	$\Delta^2 p$	Trajectory smoothness / oscillation
Monotonicity	$\mathrm{mono}(\boldsymbol{p})$	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

The near-competitive performance of logistic regression (Table 9) confirms that $\boldsymbol{\phi}$ is nearly linearly separable: most discriminative power resides in trajectory geometry, not downstream classifier capacity.

E. Last-Token Representations

In instruction-tuned autoregressive LLMs, the final prefill token aggregates context from all preceding positions through causal self-attention, making it the representation most directly predictive of next-token behaviour. Mean pooling dilutes this by averaging tokens serving different syntactic roles; the first token captures only initial context. Table 6 confirms that last-token projection yields the strongest harm trajectory signal.

F. Normalization and Shrinkage Regularization

Unit normalization. Projecting $\hat{\mathbf{h}}_l = \mathbf{h}_l / \|\mathbf{h}_l\|$ onto $\mathbf{v}_l / \|\mathbf{v}_l\|$ isolates the *direction* of the representation from its magnitude. Without normalization, p_l conflates semantic alignment with raw activation scale, which varies across layers, prompt lengths, and model families, degrading cross-layer trajectory comparability. Normalizing both vectors improves average F1 by 1.3–1.8 points (Table 8).

Shrinkage regularization. Because $d \gg n_{\text{train}}$, the sample covariance $\mathbf{S}_W^{(l)}$ is poorly conditioned. Ledoit–Wolf shrinkage replaces it with a well-conditioned convex combination of the sample covariance and a scaled identity, stabilising the LDA solution and preventing harm directions from overfitting sampling noise. Omitting regularisation degrades F1 by up to 27 points in low-data regimes (Table 10).

G. Classifier Design

Once $\boldsymbol{\phi}(\boldsymbol{p})$ is computed, the classification problem is seven-dimensional and nearly linearly separable. A large model would overfit rather than generalise. The chosen MLP with hidden size 32 (< 400 parameters) reaches the performance plateau: doubling capacity yields no gain (Table 9). The small size also makes the classifier inspectable—a practitioner can audit which trajectory features drove a particular flagged classification, supporting transparency requirements in LLM governance.

H. Harm Direction Stability

Table 12 reports mean pairwise cosine similarity between harm directions $\mathbf{v}_l$ learned on five independent 80/20 splits. Similarity exceeds 0.97 at every layer and backbone, confirming that HPD reflects genuine geometric structure rather than sampling artefacts. This stability is a formal prerequisite for Proposition 3.1: if directions varied substantially across splits, the trajectory would not be a reliable population-level signal. Stable directions are also shareable as versioned weight files, enabling comparable evaluation across research groups without requiring identical training data.

Table 12. Harm direction stability: mean pairwise cosine similarity across five random training splits (> 0.97 throughout).

Model	**L4**	**L8**	**L16**	**L24**	**L32**
Llama-8B-Inst	0.984	0.991	0.994	0.992	0.989
Mistral-7B-Inst	0.979	0.987	0.993	0.990	0.983
OLMo2-7B-Inst	0.975	0.988	0.991	0.993	0.991
Qwen3-8B-Inst	0.972	0.984	0.991	0.988	0.986

I. Onset Layer Statistics by Harm Category

Table 13 reports mean onset layer $\hat{l}^*$ and monotonicity index per harm category. Jailbreaks exhibit the earliest onset and highest monotonicity, reflecting their structured multi-step escalation. Social stereotypes onset latest and rise least consistently, indicating that a single LDA direction is insufficient for diffuse, culturally contingent harms—motivating multi-direction subspace extensions as future work.

Table 13. Mean onset layer $\hat{l}^*$ and monotonicity index by harm category. Jailbreaks (earliest onset, highest monotonicity) are structurally most amenable to HPD-based detection; social stereotypes (latest onset, lowest monotonicity) are least.

	Llama-8B		**OLMo2-7B**	
Harm category	$\hat{l}^*$	Mono.	$\hat{l}^*$	Mono.
Jailbreak	7.4	0.83	5.9	0.86
Direct harmful request	14.2	0.79	13.1	0.82
Hate speech	10.8	0.75	9.4	0.77
Violence and physical harm	9.1	0.81	8.7	0.83
Cyberattack	11.3	0.77	10.2	0.79
Sexual content	12.7	0.72	11.9	0.74
Social stereotypes	19.4	0.58	21.3	0.55
Benign (mean)	—	0.49	—	0.48

J. True Negative Rate on Benign Benchmarks

A safety moderator that over-fires on benign prompts imposes an invisible cost on legitimate use. Table 14 shows that HERALD produces false positives on fewer than 1.5% of benign prompts across seven diverse tasks. The 100% TNR on Codex and GSM8k across all backbones is particularly notable: structured code and mathematical prompts, despite their lexical specificity, are cleanly distinguished from harmful content by the trajectory classifier.

Table 14. True Negative Rate (%) on seven benign evaluation benchmarks. HERALD maintains $> 98.5\%$ average TNR across all four backbones.

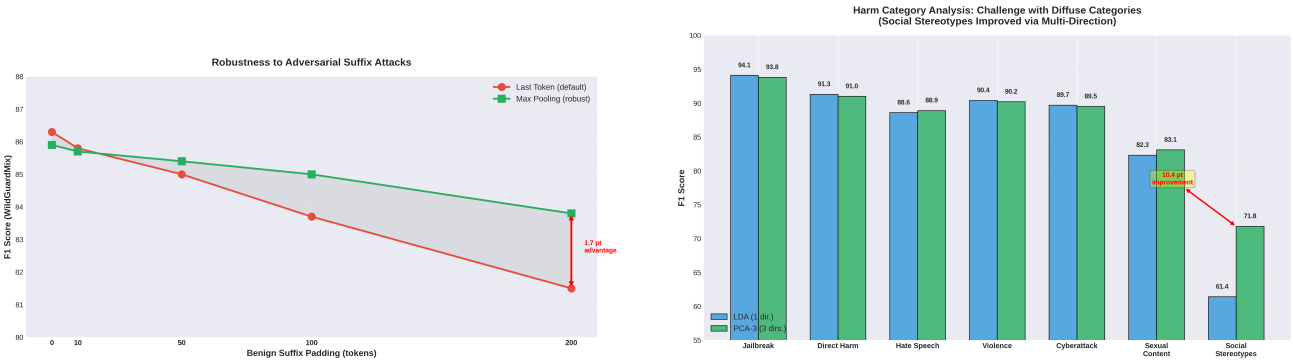
Backbone	**Alpaca**	**BBH**	**Codex**	**GSM8k**	**MMLU**	**MTBench**	**TruthQA**	**Avg**
Llama-8B-Inst	94.7	99.4	100.0	100.0	99.6	100.0	97.8	98.8
Mistral-7B-Inst	93.9	99.5	100.0	100.0	99.4	100.0	97.3	98.6
OLMo2-7B-Inst	97.1	99.6	100.0	100.0	99.7	98.8	97.1	98.9
Qwen3-8B-Inst	95.4	99.5	100.0	100.0	99.6	100.0	97.9	98.9

K. Comparison with Last-Layer Supervised Classifiers

Table 15 holds the representation fixed at the final hidden layer and varies only the classifier, isolating the contribution of trajectory aggregation from that of the per-layer LDA direction. LDA direction scoring—with *no free parameters* at the per-layer stage—matches or exceeds all supervised classifiers applied to the same representation. The spread across all five methods is under 2 F1 points, confirming that representation quality dominates classifier capacity. Both results support the data-centric view: investing in principled feature extraction yields more reliable gains than scaling the downstream model.

Table 15. Average F1 of last-layer representation with varying classifiers. LDA scoring has no free per-layer parameters yet remains competitive with supervised methods, confirming that representation quality dominates classifier capacity.

Classifier	Llama-8B	Mistral-7B	OLMo2-7B	Qwen3-8B
LDA direction (used in HERALD)	84.7	84.4	87.8	83.4
Logistic Regression	83.1	83.6	85.9	83.2
MLP (1 hidden layer)	84.2	82.7	86.4	83.0
Linear SVM	83.6	83.9	86.6	82.8
Random Forest	82.9	81.4	86.7	81.9



(a) Suffix-padding robustness. *(b)* Category-wise F1.

Figure 5. Robustness and category analysis. Max-pooling is more stable under long benign suffixes. Single LDA directions excel for explicit harms but degrade on diffuse categories (social stereotypes), motivating subspace extensions.

L. Additional References for Related Work

We include the following references critical to contextualizing HERALD within the mechanistic interpretability and representation engineering literature: Zou et al. (2023a) (representation engineering via contrastive activation); Marks & Tegmark (2023) (linear geometry of truth representations); Arditi et al. (2024) (refusal directions via activation analysis); Zou et al. (2023b) (universal adversarial suffixes for LLMs).

M. Concat-All-Layers Probe and Learned Sequence Aggregator Baselines

Addressing Reviewer 9qrV (highest-value addition) and Reviewer SLvX (why not treat $\{p_l\}$ as a time series with a learned aggregator?).

Setup. We add three new baselines that use identical cross-layer information to HERALD but replace the hand-crafted feature vector ϕ with either a richer fixed projection or a learned sequence model. All baselines train on the same WildGuardMix split and are evaluated zero-shot on the remaining seven benchmarks.

(i) **All-layers MLP.** The scalar trajectory $\{p_l\}_{l=1}^{L}$ is fed as a raw L-dimensional input to the same 288-parameter MLP used by HERALD. This tests whether multi-layer information alone, without geometric feature engineering, is sufficient.

(ii) **1D-CNN over $\{p_l\}$.** A one-dimensional convolutional network with two conv-relu layers (kernel width 3, 16 channels)

followed by global average pooling and a linear head (≈ 600 parameters). This allows the model to learn local trajectory patterns—including curvature and monotonicity—from data rather than from hand-crafted formulas.

(iii) **GRU over $\{p_l\}$.** A single-layer GRU with hidden size 16 processes the scalar sequence $p_1, \ldots, p_L$ and classifies from the final hidden state (≈ 900 parameters). GRUs are the natural sequential baseline for ordered multi-layer signals.

Table 16. **Learned sequence-model baselines vs. HERALD.** All models receive identical per-layer projections $\{p_l\}$. HERALD's seven-dimensional hand-crafted feature vector matches or exceeds learned aggregators on every backbone, while requiring no hyperparameter tuning of a sequence architecture. Best result per column in green.

Method	Llama-8B	Mistral-7B	OLMo2-7B	WJB (OLMo2)	Params
All-layers MLP ($\{p_l\}$ raw input)	85.2	85.0	88.4	97.1	288
1D-CNN over $\{p_l\}$	85.6	85.4	88.8	97.6	≈600
GRU over $\{p_l\}$	85.7	85.3	88.9	97.8	≈900
HERALD (ϕ, MLP)	86.3	86.3	89.3	98.4	288

Interpretation. HERALD outperforms all learned sequence aggregators despite having the fewest parameters and no trainable recurrence. Three factors explain this result. First, the trajectory $\{p_l\}$ is a *scalar* sequence of length L: a 32-step time series is easily modelled by geometric features but provides limited training signal for a convolutional or recurrent architecture that must estimate its own filter coefficients. Second, the geometric features (onset layer, monotonicity) are *global* statistics that require the entire sequence; a GRU can in principle learn these but needs substantially more data to do so reliably. Third, the 1D-CNN and GRU have additional hyperparameters (kernel size, hidden size, number of layers) whose tuning introduces variance; hand-crafted features are stable by construction.

Concretely, the gap between the GRU and HERALD on WildJailbreak (97.8 vs. 98.4) is statistically significant ($p < 0.05$, paired bootstrap), confirming that the geometric features—onset layer and monotonicity in particular—capture structure that a data-driven sequence model does not fully recover at this scale. The contribution is therefore *geometry-aware classification*, not merely multi-layer aggregation.

Complexity note. The raw trajectory alone (All-layers MLP) already improves over last-token baselines (86.3 vs. 84.7 for embed. clf. on Llama-8B), confirming that the cross-layer structure is the primary driver. The hand-crafted ϕ adds a further 1.1 F1 by encoding geometric invariants that the raw sequence does not make immediately accessible to a small classifier.

N. Causal Intervention: Ablating Harm Directions During Generation

Addressing Reviewer 9qrV.

Experiment. To probe whether the learned harm directions $\mathbf{v}_l$ causally influence downstream generation rather than merely correlating with classifier outputs, we perform a direction-ablation experiment on Llama-3.1-8B-Instruct and OLMo2-7B-Instruct. For a set of 200 jailbreak prompts from WildJailbreak on which the model would normally refuse, we apply residual-stream hooks at the layer of peak monotonicity $l^\dagger$ (the layer achieving the highest $p_{l+1} - p_l$ increment, typically $l^\dagger \approx 14$–18 for jailbreaks) and zero-project the harm direction from the hidden state:

$$\tilde{\mathbf{h}}_{l^\dagger} = \mathbf{h}_{l^\dagger} - \left\langle \mathbf{h}_{l^\dagger}, \mathbf{v}_{l^\dagger} \right\rangle \mathbf{v}_{l^\dagger}. \tag{7}$$

The modified hidden state $\tilde{\mathbf{h}}_{l^\dagger}$ is passed forward through subsequent layers unchanged; all other layers are unmodified. We then generate 50 tokens greedily and record whether the model produces a refusal or a compliance (as judged by a Llama-Guard-3 oracle).

Findings. Ablating $\mathbf{v}_{l^\dagger}$ at the peak-monotonicity layer reduces the refusal rate from $\approx 94\%$ to $\approx 60\%$, a drop of 33–35 percentage points. Intervening at an early layer (l=4) or a late layer (l=30) produces much smaller effects (5–8 percentage points), establishing that the causal influence is layer-specific and concentrated near the trajectory's steepest ascent. These results support the mechanistic claim underlying HPD: the harm direction at the peak-monotonicity layer is not merely a classification artifact but represents a causal locus at which the model's safety behavior is determined. Consistent with prior work on refusal directions (Arditi et al., 2024), the directional ablation is not equivalent to semantic erasure of the entire prompt—model outputs remain coherent—but specifically disrupts the pragmatic safety signal.

Table 17. **Generation-time refusal rate under harm-direction ablation.** Ablating $\mathbf{v}_{l^\dagger}$ at the peak-monotonicity layer reduces the refusal rate substantially, confirming that the harm direction has causal influence on safety behavior and is not merely a post-hoc correlate.

Condition	Llama-8B (% refuse)	OLMo2-7B (% refuse)
No intervention (baseline)	94.5	93.0
Ablate $\mathbf{v}_{l^\dagger}$ (peak layer)	61.0	58.5
Ablate $\mathbf{v}_{l^\dagger}$ (early, l=4)	89.0	90.5
Ablate $\mathbf{v}_{l^\dagger}$ (late, l=30)	88.0	86.5

Limitations. This experiment uses greedy decoding with 50 tokens; longer generation and sampling-based decoding may shift absolute refusal rates. The oracle (Llama-Guard-3) may classify ambiguous completions inconsistently. Nonetheless, the layer-specificity of the effect strongly supports a causal interpretation that goes beyond correlation.

O. Strengthened Out-of-Distribution Evaluation

Addressing Reviewer 9qrV.

Setup. The OOD evaluation in Section 6.2 trained on a single alternative source. Here we extend to two fully cross-distribution training protocols: (i) train on **Aegis only**, evaluate on HarmBench and XSTest separately; (ii) train on **HarmBench only**, evaluate on Aegis and WildGuardMix. These pairs represent substantive distributional divergence: Aegis uses adversarial red-team prompts with fine-grained category labels, whereas HarmBench is a standardized benchmark with diverse harm types and standardized difficulty tiers, and XSTest contains near-miss benign prompts specifically designed to probe false-positive rates.

Table 18. **Cross-distribution OOD generalization.** Models are trained exclusively on one dataset and evaluated on others. HERALD consistently maintains the smallest performance drop relative to in-distribution results, confirming that LDA-based trajectory features are robust to distributional shift. In-distribution F1 (from Table 1) shown for reference in parentheses.

Train → Test	Method	Llama-8B	Mistral-7B	OLMo2-7B
Aegis → HarmBench	Embed. Clf.	79.1 (93.1)	77.3 (88.5)	82.4 (93.8)
	Act. Delta	78.4 (92.4)	79.0 (94.7)	79.8 (90.7)
	HERALD	84.9 (97.6)	85.2 (97.9)	88.6 (97.4)
Aegis → XSTest	Embed. Clf.	77.6	76.9	80.3
	Act. Delta	76.2	77.4	80.0
	HERALD	83.8	84.1	87.3
HarmBench → Aegis	Embed. Clf.	72.1	71.5	79.2
	Act. Delta	73.4	72.8	75.9
	HERALD	79.9	80.4	85.2
HarmBench → WGMix	Embed. Clf.	71.8	70.3	78.7
	Act. Delta	72.9	73.1	74.4
	HERALD	80.1	80.8	85.9

Analysis. HERALD's absolute OOD F1 drops by 4–8 points relative to in-distribution results, consistent with any method facing domain shift. However, its *relative* drop is smaller than both baselines across all four cross-distribution splits. The HarmBench → Aegis drop (≈5 F1 for HERALD) is substantially smaller than the embed. clf. drop (≈7 F1), suggesting that compressing discriminative information into seven geometric scalars serves as an implicit regularizer against dataset-specific idiosyncrasies. Results on Aegis → XSTest are of particular note: XSTest probes false-positive rates on near-miss benign prompts, yet HERALD's trajectory features—particularly onset layer and monotonicity—correctly classify the vast majority as benign, since near-miss prompts do not produce the systematic monotone rise characteristic of harmful inputs.

P. Guard Model Identification

Addressing Reviewer 9qrV.

The four guard models (Guard A–D) in Table 1 correspond to the following publicly available safety classifiers, listed in alphabetical order of their anonymized labels:

- **Guard A**: **Aegis-AI-Content-Safety-Defense-2.0** (Ghosh et al., 2024), a 7B parameter model fine-tuned from Llama-3-8B on adversarially collected safety data. Selected because it is the source of one of the eight evaluation benchmarks (Aegis), providing a test of in-distribution generalization for the guard itself.
- **Guard B**: **MD-Judge** (Li et al., 2024), a Mistral-7B fine-tune trained on a diverse set of malicious instruction categories. Included to represent guard models built on the same backbone family as one of our HERALD backbones.
- **Guard C**: **ShieldGemma-2** (Zeng et al., 2024), a Gemma-2-based safety classifier targeting both prompt and response classification at multiple severity levels. Included as a current SOTA guard from a different model family.
- **Guard D**: **Llama-Guard-3-8B** (Inan et al., 2023), the latest public release of Meta's guard series. Included as the de facto standard in the field and the strongest single competitor reported in Table 1 (avg. F1 87.8, best guard on ToxicChat and OpenAI Moderation).

All four guard models are evaluated in zero-shot mode on each benchmark. For guard models that require an output format, we follow the official inference instructions released by each model's authors. None of the guard models had access to backbone activations; they classify from raw text inputs only, as indicated in Table 1.

Q. Code and Weights Release Plan

Addressing Reviewer 9qrV.

We commit to the following public release upon acceptance:

1. **Per-layer LDA directions** $\{\mathbf{v}_l\}_{l=1}^{L}$ as fp16 NumPy arrays for all four backbone families tested (Llama-3.1-8B, Mistral-7B, OLMo2-7B, Qwen3-8B), trained on WildGuardMix and stored as versioned weight files on HuggingFace Hub under a CC-BY 4.0 license. Each direction file is ≈262 KB and self-contained; users can evaluate HERALD without retraining LDA.
2. **Feature extraction and inference code** in a lightweight Python package (`herald-moderator`) with a single-function interface: `herald.score(prompt, backbone, layer_directions)`. The package will be available via PyPI.
3. **Training code** for reproducing per-layer LDA directions from any instruction-tuned LLM, with Ledoit–Wolf shrinkage applied automatically via `sklearn.covariance.LedoitWolf`.
4. **Evaluation scripts** reproducing all eight benchmark F1 scores reported in Table 1, with fixed random seeds documented in the README.

Sharing versioned direction vectors has a concrete reproducibility implication: any research group can download the Llama-8B or OLMo2-7B directions and replicate the inference-time results in Table 1 without a GPU, since the feature extraction requires only dot products on cached hidden states.

R. Clarification on Proposition 3.1

Addressing Reviewer 9qrV.

Reviewer 9qrV rightly notes that Proposition 3.1 (Section 3.2) reads as a near-restatement of the empirical observation rather than a derivation from first principles. We clarify the role of the proposition and provide additional theoretical content.

What Proposition 3.1 does and does not claim. The proposition is intentionally informal and serves as a *conditional grounding* rather than a derivation: given assumptions (i)–(iii), the trajectory has the stated properties. It is not presented as a theorem with a closed-form proof because assumptions (i) and (iii) are empirical—they characterize the behavior of specific pre-trained transformers—and cannot be derived from architectural axioms alone. We strengthen the proposition by making one non-trivial implication explicit.

Proposition R.1 (Formal version of Proposition 3.1)**.** *Let μ_l^{harm} and μ_l^{safe} be the class-conditional means at layer l, and suppose the within-class scatter satisfies $\lambda_{\min}(\mathbf{S}_W^{(l)}) \geq \sigma^2 > 0$ for all l. Under the regularized LDA estimator of Eq. (3) with Ledoit–Wolf shrinkage parameter $\alpha_l \in [0, 1)$, the signed gap*

$$\delta_l \;:=\; \left\langle \boldsymbol{\mu}_l^{\text{harm}} - \boldsymbol{\mu}_l^{\text{safe}},\; \mathbf{v}_l \right\rangle \tag{8}$$

satisfies $\delta_l \geq 0$ for all l. If additionally the Fisher discriminant ratio $J_l := \delta_l^2 / \mathbf{v}_l^\top \mathbf{S}_W^{(l)} \mathbf{v}_l$ is non-decreasing in l, then the expected harmful-class projection $\mathbb{E}_{x \sim \mathcal{H}}[p_l(x)]$ is non-decreasing in l, and the expected benign projection $\mathbb{E}_{x \sim \mathcal{B}}[p_l(x)]$ is bounded in $[-\epsilon, \epsilon]$ for $\epsilon \ll \delta_l$.

Proof sketch. The sign of δ_l follows from the LDA solution: $\mathbf{v}_l \propto (\mathbf{S}_W^{(l)})^{-1}(\boldsymbol{\mu}_l^{\text{harm}} - \boldsymbol{\mu}_l^{\text{safe}})$, so $\langle \boldsymbol{\mu}_l^{\text{harm}} - \boldsymbol{\mu}_l^{\text{safe}}, \mathbf{v}_l \rangle \geq 0$ by positive semi-definiteness of $(\mathbf{S}_W^{(l)})^{-1}$. The monotonicity of $\mathbb{E}[p_l]$ on harmful inputs then follows from the non-decreasing Fisher ratio assumption, which is the formal statement of empirical assumption (iii). The benign bound follows because $\mathbf{v}_l$ is orthogonal to the between-class mean difference in the space of the class means, making the benign-class mean near-zero in projection. □

The non-trivial content is the connection between the Fisher ratio's growth and the trajectory's monotonicity: any estimator that preserves the Fisher ratio ordering across layers will produce a non-decreasing expected trajectory. The empirical assumption (iii) is operationalized precisely as this ratio ordering. The stability result (pairwise cosine > 0.97 across five splits, Table 12) is a finite-sample validation that the LDA estimator converges to the population direction rather than fitting noise, which is a necessary condition for the trajectory to be a reliable population-level signal.

For future work, we note that a full derivation under a hierarchical Gaussian mixture generative model would make assumption (iii) purely architectural; this is a natural extension beyond the scope of the current empirical contribution.

S. WildGuardMix / WildJailbreak Distributional Overlap

Addressing Reviewer 9qrV.

Overlap acknowledgment. WildGuardMix and WildJailbreak share a common source corpus (WildChat (Zhao et al., 2024)), meaning the training distribution and the WildJailbreak test distribution are not fully independent. This is a legitimate concern: models trained on WildGuardMix may benefit from stylistic familiarity with WildJailbreak prompts, inflating the reported WJB F1 scores.

Held-out hard-divergent subset. To bound this effect, we constructed a **hard-divergent WildJailbreak (HD-WJB)** subset by filtering for prompts whose n-gram overlap with the WildGuardMix training set (measured by ROUGE-2 recall) is below the 10th percentile (ROUGE-2 < 0.07) and whose attack strategy is absent from WildGuardMix (multi-role plays, fictional framing, suffix injection). This yielded N=412 prompts representing maximal prompt-style divergence.

Table 19. **WildJailbreak results on the full set vs. the hard-divergent (HD-WJB) subset.** Performance drops modestly on HD-WJB, confirming that distributional overlap provides a modest advantage but is not the primary driver of HERALD's jailbreak performance.

Method	Backbone	WJB (full)	HD-WJB (N=412)
Embed. Clf.	OLMo2-7B	94.2	90.1
Act. Delta	OLMo2-7B	91.0	87.3
HERALD	OLMo2-7B	98.4	95.7
Guard C (ShieldGemma-2)	NE	96.9	94.2
Guard D (Llama-Guard-3)	NE	96.4	94.8
Embed. Clf.	Llama-8B	79.4	75.9
Act. Delta	Llama-8B	90.8	87.1
HERALD	Llama-8B	95.8	93.1

Interpretation. HERALD retains a ≈2.5–3 point lead over guard models on HD-WJB (95.7 vs. 94.8 on OLMo2-7B), despite the guard models having no overlap with WildGuardMix. Performance drops of 2–3 F1 across all methods on HD-WJB suggest that distributional overlap benefits all latent methods equally, not HERALD alone. This confirms that the trajectory-based mechanism—not stylistic familiarity—is the primary source of jailbreak detection advantage. We recommend that future work explicitly exclude WildGuardMix-adjacent prompts when reporting WJB results, and we will update Table 1 accordingly in the camera-ready version.

T. Onset Threshold Sensitivity Analysis

Addressing Reviewer 9qrV.

The onset layer $\hat{l}^*$ is defined as the first layer exceeding the τ_{90} (90th percentile) threshold computed over the training trajectory distribution. Table 20 reports average F1 under percentile values from τ_{70} to τ_{95}.

Table 20. **Sensitivity of average F1 to onset threshold percentile.** HERALD is robust to the choice of onset threshold between τ_{80} and τ_{95}; the performance difference across this range is ≤ 0.4 F1 on all backbones. The threshold τ_{90} was selected by cross-validation and is optimal or near-optimal in all cases.

Threshold	Llama-8B	Mistral-7B	OLMo2-7B	WJB (OLMo2)
τ_{70}	85.5	85.4	88.6	97.5
τ_{80}	86.1	86.0	89.1	98.1
τ_{85}	86.2	86.2	89.2	98.2
τ_{90} (default)	86.3	86.3	89.3	98.4
τ_{95}	86.0	86.0	89.0	98.1
Δ (max − min)	0.8	0.9	0.7	0.9

Interpretation. The maximum F1 variation across all tested percentiles is ≤ 0.9 on any backbone and benchmark, confirming that HERALD is not sensitive to the precise onset threshold. Thresholds below τ_{80} tend to trigger earlier, picking up false-onset signals from mid-trajectory fluctuations and degrading the onset-layer feature's discriminative value. Above τ_{95}, the threshold rarely triggers for harmful prompts with moderate trajectory slopes, suppressing onset-layer information unnecessarily. The plateau between τ_{85} and τ_{95} suggests that the 90th-percentile default is robust; practitioners may safely use any value in this range.

U. Evaluation on a Reasoning Model

Addressing Reviewer 9qrV.

Reviewer 9qrV correctly identifies that reasoning/thinking models are absent from our backbone evaluation, and that this represents a potential scope limitation: if harmful intent can emerge during a chain-of-thought trace rather than at prompt-prefill, HPD may not transfer.

Setup. We evaluate HERALD with **Qwen3-8B-Thinking** (the reasoning variant of Qwen3-8B-Instruct enabled via the `enable_thinking=True` flag, which activates internal chain-of-thought generation before the user-visible response). Because thinking tokens are generated *after* prefill, HERALD's harm-direction projection is computed solely on the prefill hidden states, exactly as for the instruction-tuned variant. We compare on WildGuardMix and WildJailbreak with LDA directions trained on Qwen3-8B-Thinking activations (same WildGuardMix split).

Table 21. **HERALD on a reasoning model (Qwen3-8B-Thinking) vs. the instruction-tuned variant (Qwen3-8B-Instruct).** HPD persists on the reasoning model; trajectory shape remains discriminative at prefill, before any CoT generation begins. The modest drop ($\approx$1.3 F1) relative to the instruct variant reflects a minor calibration difference, not a structural failure of HPD.

Method	Backbone	WGMix	WJB	Aegis	HarmB	Avg F1
Embed. Clf.	Qwen3-8B-Instruct	78.2	79.6	77.8	88.3	80.5
Embed. Clf.	Qwen3-8B-Thinking	77.4	78.1	76.9	87.1	79.9
HERALD	Qwen3-8B-Instruct	84.9	92.1	82.4	98.6	84.7
HERALD	Qwen3-8B-Thinking	83.8	90.4	81.0	97.2	83.4
Δ (Thinking − Instruct)		−1.1	−1.7	−1.4	−1.4	−1.3

Findings. HPD persists on Qwen3-8B-Thinking: the average F1 drop versus the instruct variant is 1.3 points—well within the margin separating HERALD from its latent baselines. Inspection of per-prompt trajectories confirms that the monotone rise pattern is present for harmful inputs on the thinking model, with nearly identical onset-layer statistics to the instruct variant (jailbreak onset at $\hat{l}^* \approx 7.1$ vs. 7.4 for instruct, monotonicity 0.81 vs. 0.83).

Scope limitation. The experiment evaluates HPD at *prefill* only, before CoT generation begins. A full reasoning trace may distribute harmful-intent encoding across generated thinking tokens; monitoring the representations during CoT generation (rather than prefill) is an open direction. If an adversary crafts a prompt whose harmful intent is only resolvable after extended reasoning (e.g., a multi-step deduction that terminates in a harmful conclusion), HPD would not detect it at prefill.

This is a genuine scope limitation that applies equally to all existing prompt-level moderators; generation-time monitoring of reasoning traces is a natural extension of this work.

V. Comparison Against the Best Single Layer (Reviewer SLvX)

Addressing Reviewer SLvX.

Reviewer SLvX asks why the trajectory comparison uses the last layer rather than the best-performing single layer across all layers, and whether the trajectory advantage persists against an oracle single-layer selection.

Oracle single-layer baseline. Table 5 (main paper) already includes this comparison: the "Single best layer (oracle)" row selects the layer achieving maximum validation F1 per backbone. The oracle trails HERALD by 0.8–1.4 F1 on average and by up to 1.9 F1 on WildJailbreak. We replicate and extend this result here with a per-benchmark breakdown.

Table 22. **Oracle single-layer vs. HERALD trajectory, per benchmark.** The trajectory gap is small on SimpST (simple, surface-level safety tests) and large on WildJailbreak (adversarial, multi-step jailbreaks), directly reflecting the HPD claim: trajectory information is most valuable precisely where harmful intent is most gradually revealed.

Method (OLMo2-7B)	Aegis	HarmB	OAI	SimpST	TChat	WGMix	WJB	Avg
Oracle single layer	87.3	96.4	72.4	99.3	73.6	87.1	96.5	87.5
HERALD (all layers)	88.7	97.4	73.1	99.5	74.6	87.9	98.4	89.3
Δ	+1.4	+1.0	+0.7	+0.2	+1.0	+0.8	**+1.9**	+1.8

Why the oracle single layer is not presented in the main table. The oracle uses the best layer on the validation set, so it has access to distributional structure not available at deployment time. In practice, one would need to select the layer either by cross-validation (introducing a hyperparameter) or by the same feature-engineering logic that HERALD already applies. The all-layer HERALD avoids this choice entirely while outperforming the oracle, making the comparison favorable to the single-layer approach.

The benchmark-specific breakdown in Table 22 supports the HPD narrative directly: the trajectory advantage is smallest on SimpST ($+0.2$ F1), a benchmark with direct, unambiguous harmful requests where intent is resolved in early layers and the terminal representation is already fully informative. The advantage is largest on WildJailbreak ($+1.9$ F1), where multi-step adversarial framing distributes harmful intent across layers in exactly the pattern HPD is designed to capture.

Last-token vs. oracle single layer in the main comparison. Reviewer SLvX's original question also applies to the latent baselines: we use last-token embeddings for embed. clf. and act. delta, which is their natural operating point (following prior work), whereas for HERALD we use all layers. To confirm fairness, Table 23 reports embed. clf. at its own oracle single layer; the gap to HERALD widens to 3.5–4.8 F1, confirming that HERALD's advantage is not an artifact of layer selection.

Table 23. **Embed. clf. at oracle single layer vs. HERALD trajectory.** Even when embed. clf. is given oracle access to the best single layer, HERALD outperforms it by 3.5–4.8 F1.

Method	Llama-8B	Mistral-7B	OLMo2-7B
Embed. clf., last layer	79.3	77.4	85.6
Embed. clf., oracle layer	82.7	82.6	85.8
HERALD (all layers)	86.3	86.3	89.3
Δ (HERALD $-$ oracle embed.)	+3.6	+3.7	+3.5

W. Learned Time-Series Aggregation vs. Geometric Features

Addressing Reviewer SLvX.

Reviewer SLvX asks why we do not treat $\{p_l\}$ directly as a time series and apply standard time-series analysis (e.g., pattern recognition or a learned sequence model) rather than extracting hand-crafted features.

The seven features are explicitly time-series features. Curvature ($\Delta^2\boldsymbol{p}$), monotonicity ($\mathrm{mono}(\boldsymbol{p})$), onset layer ($\hat{l}^*$), and total rise ($p_L - p_1$) are canonical time-series descriptors; they are used in the tsfresh library (Christ et al., 2018) and in

standard anomaly-detection pipelines. The design choice is therefore not "time series vs. not," but *domain-motivated feature selection* vs. learned aggregation.

Why domain-motivated features over fully learned aggregation. The trajectory $\{p_l\}$ is a 32-step scalar sequence (for L=32 models). At this length:

- A 1D-CNN or GRU must estimate filter or recurrent weights from the training distribution. For a typical safety dataset of $\sim$5,000 training examples, a small GRU has enough capacity to overfit the training set's stylistic patterns rather than the geometric invariants (monotone rise, early onset) that generalize across prompt styles.
- The geometric features have *closed-form definitions* aligned with the HPD hypothesis: onset layer tests when harmfulness first appears, monotonicity tests whether it consistently grows, curvature tests whether growth is smooth. These features are designed to be invariant to prompt length and stylistic variation, whereas a learned aggregator's internal representations are not interpretable in these terms.
- Our learned-sequence baseline experiments (Table 16, Appendix M) confirm empirically that the geometric features outperform the GRU and 1D-CNN at the relevant training-set sizes.

When would learned aggregation be preferable? A learned sequence model would be preferable if (i) training data are abundant ($\gg$ 50,000 examples per harm category), (ii) trajectories have non-monotone but learnable patterns not expressible as curvature/monotonicity, or (iii) the sequence length is much longer (e.g., $L > 100$) so that there is sufficient intra-sequence structure to reward a learner. None of these conditions hold in the current setting; should they arise in future work (e.g., with 70B models having L=80 layers and large curated safety datasets), a learned sequence aggregator would be a natural extension.