MULTI-TURN JAILBREAKS STEER ALONG DIRECTIONS SINGLE-TURN SAFETY DID NOT SHAPE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-turn jailbreaks can spread harmful intent across messages that are not individually refusable, yet most deployed defenses score one turn at a time. Across five models, the multi-turn drift direction lies almost entirely outside the plane spanned by single-turn refusal and harmfulness directions. Held-out safety directions provide a positive control, retaining 0.79–0.96 of their norm in this plane; the drift direction retains only 0.03–0.09, consistent with its shuffled-label null on all five models. Directional Context Drift (DCD) reads each user turn with and without history and projects the normalized difference onto a learned attack-mode direction. This difference is zero at turn one, before history exists. A multi-scale EWMA chart with a conformal limit converts these scores into alarms. DCD adds one forward pass per turn and uses the nominal false-alarm rate as its only deployment setting. With an axis for each known mode, DCD reaches 0.957–0.971 AUROC and 0.715–0.879 TPR at 5% FPR averaged over five multi-turn attack families, versus 0.766 and 0.352 for the strongest of three guards. When attacks are paired with benign conversations receiving nearly identical guard scores, the guard sits near chance by construction while DCD retains 0.90–1.00 AUROC. On judge-confirmed attacks whose harm begins after turn one, DCD flags 76–86% before the first harmful response, versus 30% for the strongest guard. Coverage is mode-specific, so attacks steering beyond the learned directions need a new axis.

1 INTRODUCTION

Aligned language models are trained to refuse harmful requests, while jailbreaks seek the refused content. Jailbreaks differ in how much conversation they require. Single-turn attacks place the entire attempt in one prompt, whether as an adversarial suffix found by search (Zou et al., 2023b), a fluent prompt (Liu et al., 2024) or a reformulation generated by another model (Chao et al., 2025; Mehrotra et al., 2024). Because the attack is localized to one input (Wei et al., 2023), a moderation model can score that input against a harm taxonomy and block it before the target sees it (Inan et al., 2023).

The multi-turn setting breaks this correspondence. As prominent work, Crescendo (Russinovich et al., 2025) escalates one small step at a time, with each turn only slightly beyond what the model has already said, while Echo Chamber (Alobaid et al., 2026) starts from some keywords of the goal in a harmless request, then has the model expand its own replies. MultiBreak (Song et al., 2026) builds a benchmark of multi-turn attacks. These attacks extract content from models that refuse the same request when asked directly (Li et al., 2024), even when no individual message is itself refusable. Harmful intent can therefore belong to the trajectory rather than any single message (Du et al., 2025; Russinovich et al., 2025).

Several proposed defenses keep the single-turn view. Moderation models (Inan et al., 2023; Han et al., 2024; Padhi et al., 2025) score a turn, or a short window, against a harm taxonomy and are effective when harm is locally visible. If no turn is overtly harmful, however, these guards have no local violation to detect. The limitation lies in what the readout measures, since the underlying representation may still contain relevant information. Concurrent methods instead analyze entire conversations (Kulkarni, 2026; Liu et al., 2026; Yan et al., 2026), but none directly measures how prior context changes the representation of the same turn and then projects that change onto an explicit activation-space direction associated with a specific attack mode. Looking only at direction-agnostic movement is ambiguous, given that normal conversations can shift just as much.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

If accumulated context changes what the model will do, that change must leave an internal trace. We locate this along directions off the safety plane shaped by single-turn alignment. We define that plane using two prominent single-turn safety directions, the refusal direction (Arditi et al., 2024) and a separately encoded harmfulness direction (Zhao et al., 2025). We reconstruct both on five language models and project each model's multi-turn direction onto their span. As a positive control, we also project held-out safety directions that should lie within this plane. These retain 0.79–0.96 of their norm under projection, while the multi-turn directions retain only 0.03–0.09 and remain within their shuffled-label nulls across all five models. The established safety axes therefore do not span the resulting drift, leaving their readouts approximately invariant to it. Prior representation-level studies of multi-turn jailbreaks measure movement relative to existing safety axes (Bullwinkel et al., 2025; Li et al., 2026; Weng et al., 2025). We instead fit the drift direction itself.

A readout along the drift direction moves with the attack. Because multi-turn attacks exploit conversation history, Directional Context Drift (DCD) measures how that history changes the model's reading of each user turn. It reads the same user message once in the full conversation and once without its history, then normalizes and projects their difference onto a direction estimated for one attack mode by a closed-form difference of means, producing the per-turn score c_t. This difference is identically zero at turn one, before history exists. An EWMA control chart (Roberts, 1959) scans these scores across timescales, while a conformal quantile (Vovk et al., 2005) sets the alarm limit, with the nominal false-alarm rate as its only setting. Both components are standard, so the experiments evaluate the signal rather than a custom detector. Since the contextual pass already produces the model's reply, the only added model cost is one isolated-turn pass.

Beyond the geometry, the drift directions provide a useful detection signal. First, DCD separates attacks from benign traffic. Averaged over five attack families, it reaches 0.957–0.971 AUROC and 0.715–0.879 TPR at 5% FPR across five targets, compared with 0.766 and 0.352 for the strongest of three guards. This separation survives a matched control. When attacks are paired with benign conversations having nearly identical guard scores, the guard performs at chance while DCD retains 0.90–1.00 AUROC. Second, DCD's signal comes from how history changes the model's reading of a turn, not from the turn alone. Replacing the attack history with a real benign conversation while leaving the final request unchanged sharply lowers harmful compliance and DCD firing, but leaves the isolated read unchanged. The final-turn input to a per-turn guard is likewise unchanged. Third, DCD fires early enough to act. On judge-confirmed multi-turn jailbreaks, it flags 76–86% before the first harmful response, compared with 30% for the strongest guard. Its advantage is largest when no individual user turn becomes overtly harmful. On Crescendo, whose attacks become explicit early, DCD and the strongest guard are nearly tied before harm, while the guard catches more attacks.

Contributions. **(i)** Drift directions that the refusal–harmfulness plane shaped by single-turn alignment does not span (§4). **(ii)** DCD, a readout of that drift for detecting multi-turn jailbreaks before harmful responses, requiring one additional isolated-turn pass and no gradient training, with standard monitoring calibrated to a target false-alarm rate (§3, §5). **(iii)** A bank of mode-specific drift directions, together with a geometric account of its coverage limits (§3.3, §5.3).

2 Related work

Multi-turn jailbreaks and per-turn guards. Crescendo (Russinovich et al., 2025) formalizes gradual escalation, while Echo Chamber (Alobaid et al., 2026) uses model's responses to steer it towards content it would initially refuse. MultiBreak (Song et al., 2026) provides a benchmark for multi-turn attacks. He et al. (2026) show turn-level contributions are non-uniform, supporting the view that harm can emerge from the trajectory. This vulnerability is not confined to automated attacks. Human-written multi-turn jailbreaks defeat defenses that hold against automated single-turn attacks (Li et al., 2024), and attacks can be built from benign pretraining-adjacent context alone (Ren et al., 2025). Single-turn attacks are outside our scope. Turn-level moderation models remain relevant baselines because deployed systems use them to screen individual requests (Inan et al., 2023).

Trajectory-level detectors. Concurrent work monitors whole conversations. LAD (Kulkarni, 2026) combines five direction-agnostic features measuring movement between consecutive states with a supervised classifier; TrajGuard (Liu et al., 2026) reads decode-time hidden states; DeepContext (Albrethsen et al., 2026) trains a recurrent model over turn embeddings; SafeDream (Yan et al., 2026) pairs an encoder with CUSUM accumulation and THRD (Ma et al., 2026) operates on judge

scores. None computes a context differential or scores an explicit attack-mode direction (Table 7, Appendix D). LAD is the closest comparison, so we evaluate it alongside DCD at the same layer and turn budget, against ordinary chat and structure-matched benign twins in §5.2.

Activation-space directions and sequential monitoring. Our readout builds on linear representations of concepts in activation space (Zou et al., 2023a; Park et al., 2024) and difference-of-means directions used to read and steer behavior (Turner et al., 2023; Rimsky et al., 2024). In single-turn settings, Arditi et al. (2024) identify a refusal direction, while Zhao et al. (2025) show that harmfulness is encoded separately. Representation-level analyses of multi-turn attacks also track safety-related movement. Bullwinkel et al. (2025) find attacked responses represented as increasingly benign across turns, while Li et al. (2026) and Weng et al. (2025) report drift away from refusal-related representations. We instead fit the drift direction itself and measure how much of it lies within the refusal–harmfulness plane (§4). For online decisions, DCD combines an EWMA control chart (Roberts, 1959) with a conformal limit (Vovk et al., 2005), scanning the NEWMA contrast (Keriven et al., 2020) over a grid of decay-rate pairs.

3 Directional Context Drift (DCD)

DCD projects the normalized difference between the contextual and isolated representations of a turn onto a drift direction learned once per model and attack mode. A multi-scale EWMA chart with a conformal limit turns these scores into alarms.

3.1 The per-turn context differential

Consider a conversation with user turns $x_1, \ldots, x_T$ and preceding target replies $a_1, \ldots, a_{T-1}$. At each turn t, we present the same user turn with and without its history and read the residual-stream vector at layer ℓ from the final token position of each input, immediately before reply generation. Under the causal mask, that position can attend to the whole input (BehnamGhader et al., 2024):

$$h_t^{\mathrm{ctx}} = h_\ell\big(\,\mathrm{sys},\, x_1, a_1, \ldots, x_{t-1}, a_{t-1},\, x_t\,\big) \qquad \text{(the turn } \textit{in context}\text{)}, \tag{1}$$

$$h_t^{\mathrm{iso}} = h_\ell\big(\,\mathrm{sys},\, x_t\,\big) \qquad \text{(the } \textit{same} \text{ turn, presented alone)}. \tag{2}$$

The user turn and system prompt are identical in both reads and only the history is removed, so their difference isolates the effect of that history conditional on the system prompt:

$$\delta_t \;=\; h_t^{\mathrm{ctx}} - h_t^{\mathrm{iso}}, \qquad u_t \;=\; \frac{\delta_t}{\|\delta_t\|}, \qquad c_t \;=\; \langle u_t,\, w\rangle \in [-1, 1], \tag{3}$$

with $u_t = \mathbf{0}$ when $\delta_t = \mathbf{0}$, and w a fixed unit direction learned once per model and attack mode (§3.2). When $\delta_t \neq \mathbf{0}$, c_t is a cosine and depends on the direction of the context effect, not its size.

Turn 1 cannot fire by construction. At $t = 1$ there is no prior history, so both reads are given the identical input and $\delta_1 = 0$, hence $c_1 = 0$ (Proposition 1). The signal does not exist until context accumulates; this follows from the definition of DCD, regardless of the detection threshold. By contrast, a single-turn representation may already distinguish attack from benign at the first turn, before history plays any role, while the differential remains identically zero (Appendix E).

Why only the direction is used. Benign turns that depend on earlier context, through pronouns or follow-up elaboration, can produce large $\|\delta_t\|$. This magnitude information is removed by normalizing, so DCD scores the direction of context effect. Section 5.2 tests the direction-agnostic baseline against this factor, while the dot-product ablation and length dependence study are in Appendix E.

3.2 Learning the drift direction

For each model and attack mode, we estimate w once as the normalized difference between the attack and benign means of u_t, then freeze it before evaluation:

$$w \;=\; \frac{\bar{u}^{\mathrm{atk}} - \bar{u}^{\mathrm{ben}}}{\|\bar{u}^{\mathrm{atk}} - \bar{u}^{\mathrm{ben}}\|}, \qquad \bar{u}^{\mathrm{atk}} = \operatorname*{mean}_{t \in \text{attack turns}} u_t, \quad \bar{u}^{\mathrm{ben}} = \operatorname*{mean}_{t \in \text{benign turns}} u_t. \tag{4}$$

The means omit turn one, where $\delta_1 = \mathbf{0}$ carries no directional information, and pool all later turns. This closed-form difference-of-means estimate requires neither a classifier nor gradient training

(Rimsky et al., 2024). Benign topic changes tend to cancel across varied directions, while consistent mode-specific attack drift survives the average. The direction is stable across data: in a test over five attack families on three architectures, two directions fitted on disjoint data reach cosine 0.859–0.873, against 0.21–0.36 when the attack and benign labels are shuffled (Appendix F).

Axis construction. Each axis is computed directly from attack and benign conversations; labels are needed only at the conversation level, and no turn is annotated. The benign conversation is a structure-matched twin when available, i.e., a conversation produced by the same attack procedure and turn count but with a benign goal, and held-out generic chat otherwise; Appendix F measures how this choice rotates w. We select layer ℓ on held-out goals because the single-turn refusal criterion of Arditi et al. (2024) is a poor guide to drift (Appendix E). Four families select layers at 24–79% relative depth, while Crescendo selects 93–100%.

3.3 One direction per attack mode

A single w covers one local mode of attack drift. The evaluated modes are geometrically separated (§5.3), so the monitor keeps a bank $\{w^{(k)}\}_{k=1}^{K}$ with one direction for each known mode (corresponding to a specific attack), computes a DCD readout along each and uses the largest resulting chart score. Pooling sufficiently aligned directions can preserve their signal, whereas merging poorly aligned modes can cancel it. Appendix H derives and tests the cosine condition separating these cases. The threshold below is calibrated on benign conversations scored by this same maximum over the complete bank, so adding an axis requires recalibration.

3.4 From the signal to an online decision

For each bank direction, the stream c_t is monitored by a multi-scale EWMA control chart. Neither when the signal appears nor how quickly it develops is known in advance, and the attacks we test span three to thirteen turns. For each pair (a, b) with $a > b$ on a fixed grid $\mathcal{G}$ (Lucas & Saccucci, 1990), the chart maintains fast and slow moving averages of c_t and records their largest gap,

$$f_t^{ab} = a\, c_t + (1-a)\, f_{t-1}^{ab}, \quad s_t^{ab} = b\, c_t + (1-b)\, s_{t-1}^{ab}, \quad P_t^{ab} = \max_{i \le t} \left(f_i^{ab} - s_i^{ab} \right). \tag{5}$$

The running maximum preserves the strongest evidence even if the signal later subsides. To cover different rates, each cell is standardized against its benign behavior before taking the maximum,

$$G_t \;=\; \max_{(a,b)\in\mathcal{G}} \frac{P_t^{ab} - \mu_{ab}}{\sigma_{ab}}, \tag{6}$$

where μ_{ab} and σ_{ab} are the benign mean and standard deviation of P_t^{ab} for that cell. This places fast and slow cells on a common scale.

Calibrated decision. We calibrate the maximum over timescales and bank axes jointly (Vovk et al., 2005; Laxhammar & Falkman, 2015). The limit τ is the finite-sample-corrected conformal quantile of benign scores. An alarm fires when $G_t > \tau$. Under exchangeability, an independently fixed scoring rule has false-alarm probability at most α (Theorem 1). Our implementation reuses calibration data for fitting, so false-alarm rates are evaluated empirically on disjoint benign traffic (Appendix A). The nominal rate α is the only deployment setting.

Cost and deployment. The contextual state comes from reply generation, so the added cost is one short isolated-turn read (≈ 33 ms in our tests) and a CPU chart update ($\approx 5\ \mu$s in our tests) per turn. The score is non-decreasing, so its first crossing of τ defines the detection turn.

3.5 Guarantees and their limits

Three results, proved in Appendix A, delimit what the method guarantees. First, Theorem 1 formalizes the finite-sample false-alarm bound stated above. Second, if attack scores maintain a positive conditional-mean margin over benign scores after onset, detection by an ideal cumulative statistic approaches certainty as post-onset turns accumulate (Theorem 2). Proposition 3 characterizes how the deployed EWMA scan responds to the same margin in expectation, but does not provide a finite-time power guarantee. We therefore measure its detection power empirically. Held-out-goal measurements support the required margin condition but do not establish it as a general property.

Table 1: Projection onto the single-turn safety plane $\mathrm{span}(w_{\mathrm{ref}}, w_{\mathrm{harm}})$, read at each model's detection layer. The full ladder from 0.6B to 72.7B is in Appendix G.

Projected onto plane	Mistral-7B	Qwen-2.5-7B	Llama-3.1-8B	Gemma-2-9B	Phi-3-mini
Held-out *refusal* dir.	0.785	0.959	0.945	0.920	0.893
Held-out *harmfulness* dir.	0.793	0.910	0.860	0.854	0.899
w **(ours)**	**0.046**	**0.032**	**0.069**	**0.029**	**0.089**
Shuffled-label null (median)	0.039	0.096	0.062	0.057	0.060

Third, Corollary 1 considers an attacker that must accumulate at least D alignment along w while keeping $S_n = \sum_{t=2}^{n}(c_t - k)$ below θ, a limit distinct from the deployed threshold τ. Here k lies between the benign and attack conditional means. Such an attacker needs $T \geq 1 + \lceil (D - \theta)/k \rceil$ turns. This is a duration–alignment tradeoff rather than an impossibility. For fixed D, fewer turns leave less allowance, while lowering average alignment requires more turns. The bound applies only to attacks that accumulate along w; modes steering elsewhere are the open-set case studied in §5.3.

4 Directions single-turn safety did not shape

Single-turn alignment shapes distinct residual-stream directions for refusal and harmfulness. Across five architectures, the multi-turn direction w lies almost entirely outside the plane they span. Readouts confined to that plane can therefore stay nearly invariant while an attack advances along w.

4.1 The safety plane does not span the multi-turn direction

We project w onto the plane spanned by previously identified refusal (Arditi et al., 2024) and harmfulness (Zhao et al., 2025) directions. Their correlation ranges from 0.52 to 0.75 across the five targets, so they are not collinear. We estimate w jointly from Crescendo (Russinovich et al., 2025) and Echo Chamber (Alobaid et al., 2026) against benign conversations with the same escalating structure (Appendix E). Table 1 reports one model from each architecture family.

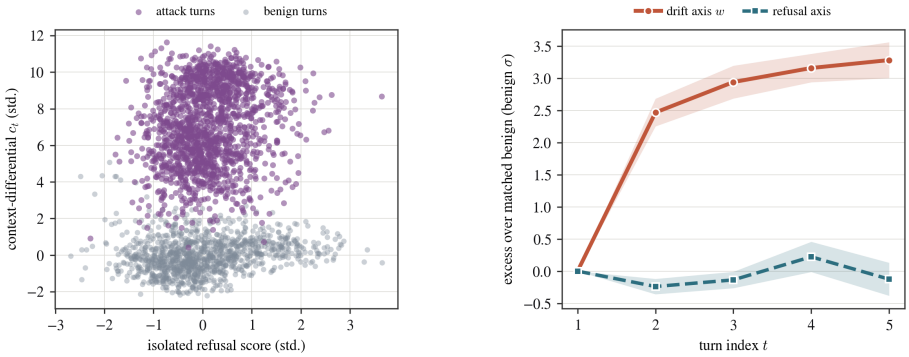
The projection is read against two references. Held-out safety directions, used as a positive control, retain 0.79–0.96 of their norm under the same projection, showing that they are captured by the plane. For the negative control, we refit w after shuffling the attack and benign labels. The true multi-turn direction retains only 0.03–0.09 of its norm inside the plane and lies within the shuffled-label null on all five targets ($p \geq 0.26$). The null comparison assesses the significance, while the squared-norm bound quantifies the overlap. The plane contains at most 0.8% of w's squared norm, compared with 62–92% for the held-out directions. Appendix G extends this to all five bank axes, to 48-dimensional safety subspaces (w retains at most 0.35) and to subspaces of post-training shifts.

4.2 The signal comes from history, not the turn alone

We examine two readings of the same user turn that differ only in what each one sees. One scores the turn alone on the refusal direction, while the other measures the difference between its contextual and isolated representations, capturing what history adds. Figure 1a shows every turn under both. Attack and benign turns overlap on the horizontal axis, with isolated-score medians of 0.1 and -0.2, but separate vertically at 7.1 and -0.1 benign standard deviations. It holds in all ten Crescendo and Echo Chamber cells (Appendix E). The isolated refusal score rises as an attack proceeds, so single turns are not entirely uninformative, but it never becomes separating. In contrast, the differential becomes strongly separating after just one prior exchange. Thus, the signal comes from how accumulated history changes the model's representation of a turn, rather than from the turn itself.

4.3 Bypass, not erosion

If accumulated context worked by gradually weakening refusal, we would expect the refusal projection to decrease as compliance increases. Figure 1b shows both the drift and refusal readouts across attack turns, each measured relative to a benign conversation with the same structure. The drift readout rises to 3.3 benign standard deviations, while the refusal readout stays within 0.25 of zero.



(a) Isolated refusal score against the context differential c_t.

(b) Movement along the drift and refusal axes, by turn.

Figure 1: Two readouts of the same attack and benign turns.

Refusal is therefore not displaced; instead, the attack moves along a direction that the refusal axis does not capture. We refer to this representational pattern as bypass rather than erosion. It differs from the single-turn setting, where refusal is suppressed while harmfulness remains. We limit the claim to the content-drift class studied here. Other multi-turn settings report movement away from refusal-related representations (Li et al., 2026; Weng et al., 2025), while a representation-level study of Crescendo finds attacked responses represented as increasingly benign (Bullwinkel et al., 2025). Reconciling these findings under a matched-family comparison remains an open question.

Replacing the history provides a direct test of how much context matters. We substitute an attack's history with a real benign conversation while leaving its final request unchanged. This reduces harmful compliance by 0.25–0.44 and DCD firing from 0.88–1.00 to 0.06–0.15 across the two families and targets. Meanwhile, the isolated representation of the unchanged request remains identical. Both behavior and the DCD signal therefore follow the history rather than the request alone (Appendix J).

4.4 A DETECTION RESULT, NOT A CAUSAL ONE

The drift axis makes attacks easier to detect, but our interventions do not establish that it causes harmful compliance, measured as the fraction of replies a response-mode judge scores harmful. Removing its component leaves that fraction unchanged (Appendix E). We also run an interchange test, setting the drift coordinate to its mean value on benign conversations while the target answers the same payload. Across 41 replayed attacks, this raises the fraction by 0.049 on Mistral-7B and leaves it unchanged on Llama-3.1-8B, relative to no intervention; magnitude-matched random edits raise it by 0.073 and lower it by 0.024. Neither drift edit reduces harmful compliance. Editing the refusal coordinate raises the fraction on Llama from 0.098 to 0.317, while its effect on Mistral is no larger than the controls (Appendix E). DCD needs only a predictive readout. We therefore treat the drift axis as a direction that established safety readouts can miss, not as a causal jailbreak direction.

5 EXPERIMENTS

5.1 SETUP

We evaluate five multi-turn attack families on five targets: Mistral-7B, Llama-3.1-8B, Qwen-2.5-7B, Phi-3-mini and Gemma-2-9B. Appendix G extends the analysis to 72.7B. SafeMTData (Ren et al., 2025), SafeDialBench (Cao et al., 2026) and the cipher corpus (Gibbs et al., 2024) provide published conversations, while Crescendo and Echo Chamber are generated from public harmful goals. Both are paired with structure-matched benign twins produced with the same procedure and turn count but benign goals (Chao et al., 2024). WildChat (Zhao et al., 2024) provides ordinary benign conversations, which we split into separate calibration and evaluation sets. Each target generates its own assistant turns. Because turn count alone separates the cipher corpus from WildChat at 0.998 AUROC, Table 2 uses a five-turn budget, while Appendix L studies the effect of adding more ordinary

Table 2: AUROC / TPR at 5% FPR. **Bold** marks the best in each column and underline the second.

		SafeMTData	SafeDialBench	cipher	Crescendo	Echo	Mean
Guards	Llama-Guard-3	0.241 / 0.032	0.490 / 0.120	0.459 / 0.008	0.561 / 0.225	0.447 / 0.111	0.440 / 0.099
	WildGuard	0.711 / 0.242	0.896 / 0.522	0.770 / 0.122	0.787 / 0.516	0.668 / 0.360	0.766 / 0.352
	Granite	0.560 / 0.163	0.773 / 0.352	0.745 / 0.141	0.761 / 0.443	0.608 / 0.297	0.689 / 0.279
LAD	Mistral-7B	0.969 / 0.822	0.766 / 0.233	0.898 / 0.448	0.821 / 0.389	0.955 / 0.865	0.881 / 0.551
	Llama-3.1-8B	0.942 / 0.648	0.639 / 0.122	0.967 / 0.857	0.947 / 0.694	**0.993** / **0.990**	0.898 / 0.662
	Qwen-2.5-7B	0.783 / 0.101	0.718 / 0.137	0.977 / 0.892	0.815 / 0.093	0.834 / 0.266	0.825 / 0.298
	Phi-3-mini	0.755 / 0.141	0.769 / 0.148	0.922 / 0.653	0.665 / 0.057	0.911 / 0.734	0.804 / 0.347
	Gemma-2-9B	0.722 / 0.121	0.797 / 0.311	0.941 / 0.653	0.872 / 0.285	0.974 / 0.937	0.861 / 0.461
DCD	Mistral-7B	0.966 / **1.000**	0.930 / 0.526	**1.000** / **1.000**	0.922 / 0.694	0.964 / 0.918	0.957 / 0.828
	Llama-3.1-8B	0.970 / 0.970	**0.958** / **0.837**	**1.000** / **1.000**	0.914 / 0.472	0.955 / 0.836	0.959 / 0.823
	Qwen-2.5-7B	0.990 / 0.997	0.946 / 0.748	**1.000** / **1.000**	0.916 / 0.731	0.968 / 0.918	0.964 / **0.879**
	Phi-3-mini	**0.998** / 0.987	0.928 / 0.452	0.999 / **1.000**	0.953 / **0.798**	0.976 / 0.923	**0.971** / 0.832
	Gemma-2-9B	0.984 / 0.963	0.923 / 0.496	**1.000** / **1.000**	**0.987** / 0.404	0.944 / 0.710	0.968 / 0.715

conversation beforehand. We report AUROC and TPR at 5% FPR. The bank contains one axis per known mode, with open-set performance reported separately. Appendix B reports the full settings while Appendix C covers the provenance, contamination checks and judging protocol.

5.2 Comparison with deployed guards

We compare DCD with three production moderation models on the same five attack families, five targets and benign traffic (Table 2). For all three guards, a conversation's score is the maximum harm score assigned to its user turns; Appendix J rescores Llama-Guard-3 with the full conversation prefix at each turn, with no consistent gain. The layer for each DCD axis is chosen on held-out goals (Appendix E). All thresholds are calibrated separately to 5% FPR on the same held-out benign traffic. Averaged across targets, DCD outperforms the strongest guard for every attack family and metric, with gains of $+0.041$ to $+0.293$ AUROC and $+0.090$ to $+0.859$ TPR. The larger TPR gains show that reasonable ranking performance does not translate into reliable detection at 5% FPR. Averaged across families, WildGuard, the strongest guard overall, reaches 0.766 AUROC and 0.352 TPR, whereas DCD reaches 0.957–0.971 and 0.715–0.879 across the five targets.

These numbers need some context. First, DCD and LAD are fitted on labeled attacks, DCD on goals held out from evaluation, while the guards are zero-shot (Table 10 compares equally fitted single reads); §5.3 reports open-set performance. Within these results, DCD trails the best guard in four of 25 TPR combinations, all on Crescendo or SafeDialBench. Second, the table uses ordinary chat as its benign class. Appendix I reports the caliper-matched control, where the strongest guard performs at chance level. Third, sensitive-topic WildChat raises false alarms, to 0.17–0.20 on Llama-3.1-8B against 0.035–0.050 on held-out chat. Recalibrating on that traffic brings all methods back to within a few points of α but reduces recall unevenly across the two targets (Table 18).

A trajectory baseline without a direction. We compare DCD with LAD (Kulkarni, 2026), a direction-agnostic trajectory baseline. DCD leads on both metrics in 21 of 25 combinations in Table 2 and on TPR in 22. It loses for Crescendo and Echo Chamber with Llama-3.1-8B and Echo Chamber with Gemma-2-9B; on SafeMTData with Mistral-7B, it trails by 0.003 AUROC but leads by 0.178 TPR. The ordinary-chat comparison measures deployment performance but does not separate direction from magnitude, since ordinary chat moves less than a scaffolded attack. Against the structure-matched twins, LAD's mean AUROC falls from 0.933 to 0.767 on Echo Chamber and from 0.824 to 0.572 on Crescendo, while DCD reaches 0.896 and 0.880, respectively, and leads in all ten combinations (Appendix J).

5.3 When the bank lacks a mode's axis

Each axis is fitted on one mode, here an attack family, so a family without an axis is detected only through the others. We test this by withholding each family's axis. Four of the five families sit 40–71° from their nearest neighbor in the bank, while cipher's nearest is 71–89°, with 90° indicating orthogonality. When cipher is scored using a bank that excludes its own direction, AUROC ranges

Table 3: Rescue before the first harmful response on 3554 judge-confirmed multi-turn successes. **Bold** marks the best per column.

		SafeMTData	SafeDialBench	cipher	Crescendo	Echo	Mean
	n successes	480	601	1955	243	275	–
	Strongest guard	0.08	0.12	0.26	0.48	0.57	0.30
DCD	Mistral-7B	0.81	0.82	**1.00**	**0.67**	0.88	0.84
	Llama-3.1-8B	0.83	0.91	**1.00**	0.24	0.84	0.77
	Qwen-2.5-7B	**0.92**	**0.97**	**1.00**	0.47	**0.96**	**0.86**
	Phi-3-mini	0.78	0.78	**1.00**	0.59	0.90	0.81
	Gemma-2-9B	0.90	0.84	**1.00**	0.42	0.64	0.76
	Mean	0.85	0.86	1.00	0.48	0.85	–

from 0.363 to 0.924 across the five targets. Adding a cipher-specific axis raises performance on the same conversations to 1.000 for every target; the only change is the direction available to the bank. The angles predict this. Across all 25 family-target cells, a family's AUROC without its own axis falls as its angle to the rest of the bank grows, at $r = -0.763$ (95% CI $[-0.896, -0.554]$) against a permutation null that places $p < 10^{-4}$. Coverage therefore tracks geometry, not attack difficulty (Table 13). The angle predicts the loss without running detection, but estimating it still needs examples of the new mode, so a mode steering somewhere new can be missed until an example exists. Appendix H reports the retention analysis behind these angles.

5.4 Detecting successful attacks before harm

Separating attacks from benign traffic does not show whether an alarm arrives in time on attacks that succeed (Table 2). We replay attacks, let the target generate its own responses, and use an independent response-mode judge to identify successful jailbreaks and their first harmful response. For each confirmed success, we record whether the monitor fires at all and whether it fires *before* that response. We call the latter fraction the *rescue* rate. We exclude successes harmful on the first reply because they are single-turn jailbreaks (27 of 3581).

Attempts and confirmed successes. Success rates vary roughly sixfold across families, from 12.8% on Crescendo to 75.6% on the cipher corpus before removing first-turn successes, so each family column mixes successes and refusals in different proportions. On the judge-confirmed multi-turn successes in Table 3, DCD flags 76–86% before the first harmful response across the five targets, compared with 30% for the strongest guard. DCD leads on four of five families. On Crescendo, it nearly ties the strongest guard before harm (0.477 versus 0.480), while the guard catches more attacks overall. Crescendo's first harmful response arrives one turn before SafeMTData's (median 3 versus 4), leaving less runway, and its axis is nearly as stable (0.957 versus 0.989 attack-half cosine), making runway and estimation unlikely explanations. Its margin is the smallest measured, at 2.01 benign standard deviations versus 2.52–5.54 elsewhere. Across 25 cells, rescue correlates with margin at Spearman $+0.723$, consistent with Theorem 2.

Rescue follows detection. On SafeMTData, where no user turn is overtly harmful, DCD's pre-harm rescue exceeds the strongest guard's by 0.70–0.83 on every target and leads on all five SafeDial-Bench targets. Each family uses its own axis, fitted on other goals, and a limit set on ordinary chat; Appendix J reports its transfer to the structure-matched twins. Cipher reaches 1.00 with its axis; §5.3 reports its open-set AUROC without it. Because cipher supplies 1955 of the 3554 in-scope successes, Table 3 averages families with equal weight rather than pooling successes.

5.5 Adaptive attackers

A detector tested only against established attacks may overstate its robustness (Nasr et al., 2026). We introduce an adaptive attack against DCD that rewrites user turns to reduce the chart statistic while requiring the jailbreak to succeed. We evaluate attackers with different levels of knowledge, ranging from no knowledge of the detector to a white-box attacker with access to the correct axis

Table 4: Adaptive attackers on the conversations that jailbreak each target unrewritten. **Bold** marks the strongest attacker.

	Llama-3.1-8B		Mistral-7B	
attacker's knowledge	detect	joint	detect	joint
no rewriting (original attacks)	0.908	0.092	0.705	0.295
none of the three (black-box)	0.868	0.118	0.795	0.193
layer and model	0.842	0.132	0.784	0.193
axis and model	0.763	0.197	0.716	0.273
axis and layer	0.789	0.184	0.705	0.284
all three (white box)	**0.711**	**0.250**	**0.602**	**0.375**

bank, layer, and target model. We report *joint* success, i.e., the fraction of conversations that evade DCD and remain jailbreaks according to an independent judge, since evasion is trivial if the attacker is willing to produce harmless text, while jailbreak success is trivial if being detected is acceptable.

Table 4 reports the results. The white-box attacker performs best on both target models, reaching a joint success rate of 0.250 on Llama and 0.375 on Mistral, compared with 0.092 and 0.295 for the original attacks. It also reduces detection the most, from 0.908 to 0.711 on Llama and from 0.705 to 0.602 on Mistral. Adaptive attackers that know only two of the three components, or none of them, are less effective, with joint success rates of 0.118–0.197 on Llama and 0.193–0.284 on Mistral, never above the original attacks on Mistral. These results show that effective evasion depends heavily on access to an exact copy of the monitor. Even with full white-box access, however, DCD detects a substantial fraction of adaptive attacks. The attacker must reduce the detection score while preserving jailbreak success, which limits how many conversations can be successfully adapted. Appendix K presents the attack methodology, additional results, why partial knowledge limits the attacker's effectiveness and why DCD continues to detect attacks even under white-box access.

6 Limitations

DCD shares the limitations of any learned-direction readout. Coverage remains open-set because one axis covers one attack mode. The readout measures alignment with a drift direction rather than intent itself, so structured benign conversations can trigger it, forcing deployments with such traffic to trade recall against false alarms (Appendix J). It also requires access to the scored model's activations and a model-specific direction; using a surrogate around a black-box target remains untested. Proposition 1 shows that c_1 is identically zero, so attacks that become harmful on the first turn are outside our scope; per-turn guards are better suited to that setting. Corollary 1 applies only to attackers who must accumulate along w. Direct embedding-space attacks can defeat activation readouts and lie outside our threat model (Schwinn et al., 2024). Finally, the advantage is largest where no user turn is overtly harmful; on Crescendo, the strongest guard nearly matches DCD before harm.

7 Conclusion

Across the five attack families studied here, multi-turn jailbreaks are not merely harmful content divided into innocuous pieces. They produce history-dependent movement along a direction outside the refusal–harmfulness plane shaped by single-turn alignment, measured by a differential that is identically zero before any history exists. The refusal and harmfulness directions do not span that drift; DCD measures it directly. A matched benign control removes the matched guard's signal by construction while DCD retains separation, and a trajectory detector reading magnitude rather than direction loses much of its separation under a structure-matched test. The readout adds only one isolated-turn pass per turn and requires no gradient training, and the alarm uses standard process control calibrated to a chosen nominal false-alarm rate. DCD reaches 0.957–0.971 AUROC, versus 0.766 for the strongest guard, and fires before the first harmful response in 76–86% of judge-confirmed multi-turn jailbreaks. Because each axis captures one attack mode, the bank shifts the open-set problem instead of eliminating it. Learning new axes from unlabeled traffic, instead of deriving them from known families, remains the main open question.

AI USE STATEMENT

AI was used solely for text refinement, including improving grammar, clarifying wording and rephrasing. The authors made all research decisions, implemented and verified the reported work and take full responsibility for the paper.

ETHICS STATEMENT

This work studies multi-turn jailbreaks for detection and does not propose a new jailbreak family. It evaluates three published attack corpora together with two published attack procedures, which we run on harmful goals from JailbreakBench (Chao et al., 2024), AdvBench (Zou et al., 2023b) and HarmBench (Mazeika et al., 2024), with matched benign twins generated from JailbreakBench's benign behaviors. Benign traffic is drawn from the publicly released WildChat corpus (Zhao et al., 2024). Every target, guard and judge is an open-weight model run locally (Table 5), so no public service or user was involved. We do not redistribute generated transcripts because some contain harmful model outputs. Appendices B and C document their construction, allowing a comparable benchmark to be rebuilt from public sources. Because the two procedure-generated families rely on stochastic sampling, rerunning them produces new rollouts rather than exact copies of the original transcripts. The adaptive evaluation in §5.5 builds a white-box attacker against our own monitor and reports when it succeeds. The study involves no human participants and collects no new personal data.

REPRODUCIBILITY STATEMENT

The supplementary material includes the detector and the analysis code, with documentation. Appendix A states the assumptions behind each theoretical claim and proves it in full. Appendix C documents the datasets, the calibration/evaluation split, the contamination audit and the judging protocol behind the success labels. The generated transcripts and the activation caches are not included, for the reasons given in the ethics statement.

REFERENCES

Justin Albrethsen, Yash Datta, Kunal Kumar, and Sharath Rajasekar. DeepContext: Stateful real-time detection of multi-turn adversarial intent drift in LLMs. *arXiv preprint arXiv:2602.16935*, 2026.

Ahmad Alobaid, Martí Jordà Roca, Carlos Castillo, and Joan Vendrell. The echo chamber multi-turn LLM jailbreak. *arXiv preprint arXiv:2601.05742*, 2026.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass LLM latent-space defenses. In *International Conference on Learning Representations*, volume 2026, pp. 146838–146882, 2026.

Michele Basseville and Igor V Nikiforov. *Detection of abrupt changes: theory and application*, volume 104. Prentice hall Englewood Cliffs, 1993.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024.

Blake Bullwinkel, Mark Russinovich, Ahmed Salem, Santiago Zanella-Beguelin, Daniel Jones, Giorgio Severi, Eugenia Kim, Keegan Hines, Amanda Minnich, Yonatan Zunger, et al. A representation engineering perspective on the effectiveness of multi-turn jailbreaks. *arXiv preprint arXiv:2507.02956*, 2025.

Hongye Cao, Sijia Jing, Yanming Wang, Ziyue Peng, Zhixin Bai, Zhe Cao, Meng Fang, Fan Feng, Jiaheng Liu, Boyan Wang, et al. SafeDialBench: A fine-grained safety evaluation benchmark for large language models in multi-turn dialogues with diverse jailbreak attacks. In *International Conference on Learning Representations*, volume 2026, pp. 125737–125785, 2026.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. JailbreakBench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 23–42. IEEE, 2025.

Xiaohu Du, Fan Mo, Ming Wen, Tu Gu, Huadi Zheng, Hai Jin, and Jie Shi. Multi-turn jailbreaking large language models via attention shifting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23814–23822, 2025.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Tom Gibbs, Ethan Kosak-Hine, George Ingebretsen, Jason Zhang, Julius Broomfield, Sara Pieri, Reihaneh Iranmanesh, Reihaneh Rabbany, and Kellin Pelrine. Emerging vulnerabilities in frontier models: Multi-turn jailbreak attacks. *arXiv preprint arXiv:2409.00137*, 2024.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. *Advances in neural information processing systems*, 37:8093–8131, 2024.

Zhida He, Xiaoyu Wen, Han Qi, Ziyuan Zhou, Peng Yu, Xingcheng Xu, Dongrui Liu, Xia Hu, Chaochao Lu, and Qiaosheng Zhang. Not all turns matter: Credit assignment for multi-turn jailbreaking. *arXiv preprint arXiv:2605.08778*, 2026.

Chenhui Hu, Muhammed Salih, Sudipto Guha, and Subramanian Srinivasan. Scalable hierarchical attention transformers for multi-turn jailbreak detection in long conversations. *arXiv preprint arXiv:2606.21082*, 2026.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Nicolas Keriven, Damien Garreau, and Iacopo Poli. NEWMA: a new method for scalable model-free online change-point detection. *IEEE Transactions on Signal Processing*, 68:3515–3528, 2020.

Prashant Kulkarni. Latent adversarial detection: Adaptive probing of LLM activations for multi-turn attack detection. *arXiv preprint arXiv:2604.28129*, 2026.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations*, volume 2026, pp. 54738–54778, 2026.

Rikard Laxhammar and Göran Falkman. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Annals of Mathematics and Artificial Intelligence*, 74(1): 67–94, 2015.

Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. LLM defenses are not robust to multi-turn human jailbreaks yet. *arXiv preprint arXiv:2408.15221*, 2024.

Pengcheng Li, Jie Zhang, Tianwei Zhang, Han Qiu, Kejun Zhang, Weiming Zhang, Nenghai Yu, and Wenbo Zhou. State-dependent safety failures in multi-turn language model interaction. *arXiv preprint arXiv:2603.15684*, 2026.

Cheng Liu, Xiaolei Liu, Xingyu Li, Bangzhou Xin, and Kangyi Ding. TrajGuard: Streaming hidden-state trajectory detection for decoding-time jailbreak defense. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13371–13388, 2026.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Learning Representations*, volume 2024, pp. 56174–56194, 2024.

James M Lucas and Michael S Saccucci. Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics*, 32(1):1–12, 1990.

Zhiqing Ma, Zhonghao Xu, Dong Yu, Chen Kang, Changliang Li, and Pengyuan Liu. THRD: A training-free multi-turn defense framework for jailbreak attacks on large language models. *arXiv preprint arXiv:2606.01738*, 2026.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David A. Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR / OpenReview.net, 2024.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs automatically. *Advances in Neural Information Processing Systems*, 37:61065–61105, 2024.

Sanjay Mishra, Divya Chukkapalli, and Ganesh R Naik. Stateful guardrails for multi-turn LLM systems: A conversational risk accumulation framework. *arXiv preprint arXiv:2607.19361*, 2026.

Milad Nasr, Nicholas Carlini, Chawin Sitawarin, Sander V Schulhoff, Jamie Hayes, Michael Ilie, Juliette Pluto, Shuang Song, Harsh Chaudhari, Ilia Shumailov, et al. The attacker moves second: Stronger adaptive attacks bypass defenses against LLM jailbreaks and prompt injections. In *35th USENIX Security Symposium (USENIX Security 26)*, 2026.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, et al. Granite Guardian: Comprehensive LLM safeguarding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 607–615, 2025.

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Haining Yu, and Xiaohua Jia. The hidden dimensions of LLM alignment: A multi-dimensional analysis of orthogonal safety directions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47697–47716. PMLR, 13–19 Jul 2025.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR / OpenReview.net, 2024.

Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. LLMs know their vulnerabilities: Uncover safety gaps through natural distribution shifts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24763–24785, 2025.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

S. W. Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250, 1959.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 2421–2440, 2025.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, 2013.

Leo Schwinn, David Dobre, Sophie Xhonneux, Gauthier Gidel, and Stephan Günnemann. Soft prompt threats: Attacking safety alignment and unlearning in open-source LLMs through the embedding space. *Advances in Neural Information Processing Systems*, 37:9086–9116, 2024.

Robin Sommer and Vern Paxson. Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE symposium on security and privacy*, pp. 305–316. IEEE, 2010.

Jialin Song, Xiaodong Liu, Weiwei Yang, Wuyang Chen, Mingqian Feng, Xuekai Zhu, and Jianfeng Gao. MultiBreak: A scalable and diverse multi-turn jailbreak benchmark for evaluating LLM safety. *arXiv preprint arXiv:2605.01687*, 2026.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.

Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC press, 2014.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? *Advances in neural information processing systems*, 36:80079–80110, 2023.

Zixuan Weng, Xiaolong Jin, Jinyuan Jia, and Xiangyu Zhang. Foot-in-the-door: A multi-turn jailbreak for LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1939–1950, 2025.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: concept cones and representational independence. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25, 2025.

Bo Yan, Weikai Lin, Yada Zhu, and Song Wang. SafeDream: Safety world model for proactive early jailbreak detection. *arXiv preprint arXiv:2604.16824*, 2026.

Doohee You. Surviving the unseen: Predictive defense for novel multi-turn multimodal attacks. *arXiv preprint arXiv:2605.18988*, 2026.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. Llms encode harmfulness and refusal separately. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1m ChatGPT interaction logs in the wild. In *International Conference on Learning Representations*, volume 2024, pp. 34590–34605, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. *Advances in Neural Information Processing Systems*, 37:83345–83373, 2024.

A PROOFS

We state the false-alarm guarantee, the assumptions behind detection power, and the restricted evasion bound.

SETUP AND CONVENTIONS

Fix a layer ℓ and a unit direction w. For a conversation with user turns $x_1, \ldots, x_T$ and target replies $a_1, \ldots, a_{T-1}$, define for $t \geq 1$

$$h_t^{\mathrm{ctx}} = h_\ell(\mathrm{sys}, x_1, a_1, \ldots, x_{t-1}, a_{t-1}, x_t), \qquad h_t^{\mathrm{iso}} = h_\ell(\mathrm{sys}, x_t), \qquad \delta_t = h_t^{\mathrm{ctx}} - h_t^{\mathrm{iso}}.$$

Definition 1 (normalization convention)**.** $u_t = \delta_t / \|\delta_t\|$ *when* $\delta_t \neq 0$*, and* $u_t = 0$ *when* $\delta_t = 0$*; in both cases* $c_t = \langle u_t, w \rangle$*.*

The zero convention makes the score well defined when the two reads coincide. Since $\|u_t\| \leq 1$ and $\|w\| = 1$, we have $c_t \in [-1, 1]$. The implementation uses $\delta_t / (\|\delta_t\| + \varepsilon)$ with $\varepsilon = 10^{-9}$, preserving zero exactly. For $\delta_t \neq 0$, the norm of its difference from exact normalization is $\varepsilon / (\|\delta_t\| + \varepsilon)$.

Proposition 1 (The context differential is zero at the first turn)**.** *For every conversation,* $c_1 = 0$*.*

Proof. At $t = 1$ there is no prior history, so both reads are given the identical input, $h_1^{\mathrm{ctx}} = h_\ell(\mathrm{sys}, x_1) = h_1^{\mathrm{iso}}$; hence $\delta_1 = 0$ and $c_1 = 0$, independent of the content of x_1 or the threshold. □

DISTRIBUTION-FREE FALSE-ALARM CONTROL

For a scoring rule fixed independently of the calibration and test conversations, exchangeability gives a distribution-free false-alarm bound. The result concerns benign traffic and requires no assumptions about attacks.

Theorem 1 (Finite-sample conformal validity)**.** *Let* $S(\cdot)$ *be a measurable conversation-level statistic fixed independently of the calibration and test conversations. Let* $Z_1, \ldots, Z_n$ *be calibration conversations and* Z_{n+1} *a test conversation, and suppose* $(Z_1, \ldots, Z_{n+1})$ *are exchangeable. Write* $S_i = S(Z_i)$*, let* $S_{(1)} \leq \cdots \leq S_{(n)}$ *be the order statistics of the calibration scores, and for* $\alpha \in [1/(n+1), 1)$ *set* $\tau = S_{(\lceil (n+1)(1-\alpha) \rceil)}$*. Then*

$$\Pr\left[S_{n+1} > \tau \right] \leq \alpha.$$

Proof. Let R be the rank of S_{n+1} among $S_1, \ldots, S_{n+1}$, ties broken uniformly at random. Under exchangeability the vector of scores is distributionally invariant to permutation, so R is uniform on $\{1, \ldots, n+1\}$. Write $m = \lceil (n+1)(1-\alpha) \rceil$. The event $S_{n+1} > \tau = S_{(m)}$ implies that S_{n+1} exceeds at least m of the $n+1$ scores, i.e., $R > m$. Hence $\Pr[S_{n+1} > \tau] \leq \Pr[R > m] = 1 - m/(n+1) \leq 1 - (1-\alpha) = \alpha$, using $m \geq (n+1)(1-\alpha)$. The condition $\alpha \geq 1/(n+1)$ guarantees $m \leq n$ so that τ is well defined. □

Three remarks clarify how the theorem is used.

Remark 1 (Joint calibration of the scan). With fixed directions, layers, normalization constants and timescale grid, the maximum over turns, timescale cells and bank directions defines a single conversation-level statistic S. Under the assumptions of Theorem 1, calibrating this complete statistic bounds the probability of any alarm during a benign conversation without a separate correction for the number of cells or axes. Changing the scoring rule requires recalibration.

Remark 2 (Sample size). With the empirical order statistic above, a finite threshold at level α requires $\alpha \geq 1/(n+1)$, equivalently $n \geq 1/\alpha - 1$. At $\alpha = 0.05$, this requires at least 19 calibration conversations. This is only a minimum; the attainable rank levels are spaced by $1/(n+1)$, so larger calibration sets provide finer resolution. Appendix D reports realized rates against nominal α.

Remark 3 (Conversation length and exchangeability). If conversation length differs systematically between calibration and deployment, pooled exchangeability may fail. Length-stratified (Mondrian) calibration instead applies Theorem 1 within each predefined length stratum, provided the scoring rule is fixed independently and benign calibration and test conversations are exchangeable conditional on that stratum. Each stratum must separately meet the sample-size requirement above.

WHY DETECTION SUCCEEDS

Theorem 1 controls false alarms under its stated conditions but says nothing about power. Detection power requires assumptions about the attack signal.

Proposition 2 (In-sample margin identity)**.** *Let $\bar{u}^a$ and $\bar{u}^b$ be the attack and benign class means over an estimation sample. Assuming $\bar{u}^a \neq \bar{u}^b$, define*

$$w = \frac{\bar{u}^a - \bar{u}^b}{\|\bar{u}^a - \bar{u}^b\|}, \qquad \mu_a = \langle \bar{u}^a, w \rangle, \qquad \mu_b = \langle \bar{u}^b, w \rangle.$$

Then, on that sample,

$$\mu_a - \mu_b = \left\|\bar{u}^a - \bar{u}^b\right\| > 0.$$

Proof. By the definition of w, $\mu_a - \mu_b = \langle \bar{u}^a - \bar{u}^b, w \rangle = \|\bar{u}^a - \bar{u}^b\|$. □

This identity holds on the sample used to estimate w. Whenever the two empirical class means differ, fitting w produces a positive in-sample margin. It does not guarantee that the direction retains a margin on new conversations. The following lemma gives a sufficient condition for such generalization under i.i.d. sampling.

Lemma 1 (Margin generalization)**.** *Let m_a and m_b be the population class means, let $D = m_a - m_b$, and let $\hat{D}$ be its sample estimate. Define $\eta = \|\hat{D} - D\|$ and $\hat{w} = \hat{D}/\|\hat{D}\|$. If $\eta < \|D\|$, then*

$$\langle D, \hat{w} \rangle \geq \frac{\|D\|\big(\|D\| - \eta\big)}{\|D\| + \eta} > 0.$$

Moreover, let $U_1, \ldots, U_n$ be i.i.d. vectors with $\|U_i\| \leq 1$, population mean m, and empirical mean $\hat{m}$. Then

$$\mathbb{E}\|\hat{m} - m\| \leq n^{-1/2}, \qquad \Pr\Big[\|\hat{m} - m\| \geq n^{-1/2} + \epsilon\Big] \leq e^{-n\epsilon^2/2}.$$

Proof. By Cauchy–Schwarz, $\langle D, \hat{D} \rangle \geq \|D\|^2 - \|D\|\eta$, while $\|\hat{D}\| \leq \|D\| + \eta$. Dividing gives the first result. For the concentration statement, independence gives $\mathbb{E}\|\hat{m} - m\|^2 = n^{-1}\mathbb{E}\|U - m\|^2 \leq n^{-1}$, and Jensen's inequality gives the expectation bound. Replacing one observation changes $\hat{m}$ by at most $2/n$, so McDiarmid's inequality gives the tail bound. □

Two empirical checks examine these implications. In a split-half test, the held-out margin exceeds the lemma's lower bound on all three tested targets (0.440 versus 0.322, 0.535 versus 0.395, and 0.545 versus 0.402). Separately, goal-disjoint fits remain strongly aligned as the sample grows, reaching cosine 0.901–0.915 with 40 conversations per fit and 0.956–0.974 with 160 across three architectures (Appendix F).

Assumption 1 (Turn homogeneity)**.** *Let $\mathcal{F}_{t-1}$ be the σ-field generated by the conversation through turn $t-1$. There exist $\mu_b < k < \mu_a$ such that for all $t \geq 2$, $\mathbb{E}[c_t \mid \mathcal{F}_{t-1}] \geq \mu_a$ on attack conversations and $\mathbb{E}[c_t \mid \mathcal{F}_{t-1}] \leq \mu_b$ on benign conversations.*

This assumption is stronger than the preceding in-sample identity and generalization bound. It requires positive separation to hold conditionally at every post-first turn, but does not require independent turns. We examine its observable implication that the marginal per-turn means remain separated. On each family's deployed axis, fitted on half its goals and scored on the other half, the

margin $\mu_a(t) - \mu_b(t)$ is positive at every turn $t = 2, \ldots, 5$ in all 25 target–family cells, with family-averaged margins of 0.27–0.45 across targets and turns. When the axis is instead fitted with the scored family withheld, the margin stays positive at every turn in 18 of the 25 cells; the exceptions are all five cipher cells and two cells whose margin falls just below zero by the fourth turn. The assumption thus holds on the axis of the mode being scored, and with that mode withheld it fails mainly for cipher, the family farthest from the rest of the bank, which is the open-set limit of §5.3. These measurements support the assumption on the evaluated data but do not establish it generally.

Theorem 2 (Directional separation)**.** *Let $S_n = \sum_{t=2}^{n}(c_t - k)$ and adopt Assumption 1. Write $\Delta_a = \mu_a - k > 0$ and $\Delta_b = k - \mu_b > 0$. Then for any threshold $\theta > 0$:*

1. (False alarms, uniformly over an unbounded horizon.) *On benign conversations,* $\Pr\big[\sup_{n\geq 2} S_n \geq \theta\big] \leq e^{-\Delta_b \theta}$.

2. (Power.) *On attack conversations, provided* $(n-1)\Delta_a > \theta$, $\Pr[S_n \leq \theta] \leq \exp\Big(-\dfrac{\big((n-1)\Delta_a - \theta\big)^2}{8(n-1)}\Big)$.

For any $\rho \in (0,1)$, choosing $\theta \geq \log(1/\rho)/\Delta_b$ gives an anytime false-alarm bound of ρ. For any fixed θ, the attack miss probability tends to zero as the number of scored turns grows.

Proof. Write $X_t = c_t - k$ and $M_n = \sum_{t=2}^{n}\big(X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]\big)$, a martingale with $M_1 = 0$. Since $c_t \in [-1, 1]$, X_t lies in an interval of length 2. Subtracting its conditional mean preserves that length, and each increment of M satisfies $|M_t - M_{t-1}| \leq 2$.

(1) Fix $\lambda > 0$. By Hoeffding's lemma applied conditionally to X_t, whose conditional range has length 2, $\mathbb{E}\big[e^{\lambda(X_t - \mathbb{E}[X_t|\mathcal{F}_{t-1}])} \mid \mathcal{F}_{t-1}\big] \leq e^{\lambda^2 \cdot 2^2/8} = e^{\lambda^2/2}$, and on benign conversations $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \leq -\Delta_b$, so $\mathbb{E}[e^{\lambda X_t} \mid \mathcal{F}_{t-1}] \leq e^{-\lambda\Delta_b + \lambda^2/2}$. Define $V_n = \exp\big(\lambda S_n + (n-1)(\lambda\Delta_b - \lambda^2/2)\big)$, with $V_1 = 1$. Then $\mathbb{E}[V_n \mid \mathcal{F}_{n-1}] = V_{n-1}\, e^{\lambda\Delta_b - \lambda^2/2}\, \mathbb{E}[e^{\lambda X_n} \mid \mathcal{F}_{n-1}] \leq V_{n-1}$, so V is a non-negative supermartingale. Ville's inequality gives $\Pr[\sup_n V_n \geq c] \leq 1/c$ for every $c > 0$. Choosing $\lambda = \Delta_b$ makes the deterministic factor $(n-1)(\Delta_b^2 - \Delta_b^2/2) \geq 0$, hence $V_n \geq e^{\Delta_b S_n}$, and therefore $\Pr[\sup_n S_n \geq \theta] \leq \Pr[\sup_n V_n \geq e^{\Delta_b\theta}] \leq e^{-\Delta_b\theta}$.

(2) On attack conversations $S_n \geq M_n + (n-1)\Delta_a$, so $\{S_n \leq \theta\} \subseteq \{M_n \leq \theta - (n-1)\Delta_a\}$. With $|M_t - M_{t-1}| \leq 2$, the Azuma–Hoeffding inequality gives $\Pr[M_n \leq -u] \leq \exp\big(-u^2/(2\sum_{t=2}^{n} 2^2)\big) = \exp\big(-u^2/(8(n-1))\big)$ for $u > 0$; take $u = (n-1)\Delta_a - \theta$, which is positive by hypothesis. □

The false-alarm bound is uniform in time. It controls whether the ideal cumulative statistic ever crosses θ; a fixed-n tail bound would not. The power bound improves with both the number of scored turns and the margin Δ_a. A short attack may therefore fail to cross θ even when its per-turn margin is positive.

The theorem concerns the ideal cumulative statistic rather than the deployed EWMA scan. The next proposition characterizes the scan's expected response to the same mean shift but does not provide a finite-time power guarantee.

Proposition 3 (Expected response of the deployed scan)**.** *Fix a cell (a, b) with $1 > a > b > 0$, let $f_t^{ab} = ac_t + (1-a)f_{t-1}^{ab}$ and $s_t^{ab} = bc_t + (1-b)s_{t-1}^{ab}$, and suppose $\mathbb{E}[c_t \mid \mathcal{F}_{t-1}] = \mu_a$ from an onset turn t_0 onward. Write $\phi = \mathbb{E}[f_{t_0-1}^{ab}]$ and $\psi = \mathbb{E}[s_{t_0-1}^{ab}]$ for the filter states at onset. Then after $q \geq 1$ post-onset updates, at $t = t_0 + q - 1$,*

$$\mathbb{E}\big[f_t^{ab} - s_t^{ab}\big] = (1-a)^q(\phi - \mu_a) - (1-b)^q(\psi - \mu_a).$$

If the filters are at their pre-onset equilibrium, $\phi = \psi = \mu_b$, then with $\mu_a > \mu_b$ and $\Delta = \mu_a - \mu_b > 0$ this is $\Delta[(1-b)^q - (1-a)^q]$, and over a horizon of Q post-onset turns

$$\mathbb{E}\big[P_{t_0+Q-1}^{ab}\big] \geq \Delta \cdot g_Q(a,b), \qquad g_Q(a,b) = \max_{1\leq q\leq Q}\big[(1-b)^q - (1-a)^q\big] > 0.$$

The deployed statistic subtracts each cell's benign mean μ_{ab} and divides by its benign standard deviation σ_{ab}, so it satisfies $\mathbb{E}[G_{t_0+Q-1}] \geq \max_{(a,b)\in\mathcal{G}} \big(\Delta\, g_Q(a,b) - \mu_{ab}\big)/\sigma_{ab}$: a centered-and-scaled expected response, not a constant multiple of Δ.

Proof. Both recursions are linear with unit DC gain, so at stationary equilibrium f_t^{ab} and s_t^{ab} share the same mean and their difference has mean zero: the statistic responds to change, not level. Unrolling the fast recursion over q post-onset updates gives $\mathbb{E}[f_t^{ab}] = (1-a)^q\phi + a\mu_a \sum_{r<q}(1-a)^r = \mu_a + (1-a)^q(\phi - \mu_a)$, and identically for s_t^{ab} with b; subtracting gives the first display. At the onset turn one update has already occurred, so $q = 1$ and the response is $(a-b)\Delta$ rather than zero. The running peak is at least its value at every post-onset turn, which gives g_Q. Finally $G_t \geq (P_t^{ab} - \mu_{ab})/\sigma_{ab}$ for each cell, so taking expectations and maximizing the resulting valid lower bound over the grid gives the last display. □

Turns are discrete, so the maximization is over integer q. Across the deployed grid and all $q \geq 1$, the largest raw response factor is 0.8625, attained by $(a, b) = (0.8, 0.05)$ at $q = 2$, the second post-onset update. This corresponds to an expected fast-minus-slow difference of 0.8625Δ under the idealized step change. It is not a lower bound for the centered and scaled statistic G_t, nor does it guarantee detection by that update. The proposition shows only that the scan's expected response depends on the same margin Δ as the ideal cumulative statistic; deployed detection power remains assessed empirically.

THE COST OF EVASION

The cumulative rail also yields a restricted evasion tradeoff.

Corollary 1 (Restricted axis-alignment budget)**.** *Let $k > 0$ and $D > \theta$. An attacker that must accumulate at least D alignment along w, so that $\sum_{t=2}^{T} c_t \geq D$, while keeping $S_n \leq \theta$ for every $n \leq T$, requires $T \geq 1 + \lceil (D-\theta)/k \rceil$ turns.*

Proof of Corollary 1. The constraint at $n = T$ reads $\sum_{t=2}^{T}(c_t - k) \leq \theta$, and the sum has $T-1$ terms, so $\sum_{t=2}^{T} c_t \leq k(T-1) + \theta$. Combining with $\sum_{t=2}^{T} c_t \geq D$ gives $D \leq k(T-1) + \theta$, hence $T - 1 \geq (D-\theta)/k$ since $k > 0$, and T is an integer. □

This is a deterministic consequence of the threshold constraint and carries no probabilistic claim. It states a restricted tradeoff rather than an impossibility. Avoiding the cumulative rail requires

$$\frac{1}{T-1}\sum_{t=2}^{T} c_t \leq k + \frac{\theta}{T-1}.$$

An attack that must accumulate alignment D can therefore reduce its average per-turn drift only by using more turns. The result applies only when progress is measured along the learned direction w and against the ideal cumulative statistic S_n; it does not characterize attacks that steer elsewhere or the exact failure point of the deployed EWMA scan.

SCOPE OF THE GUARANTEES

Three scope limits. *(i)* Theorem 1 gives exact false-alarm control for an independently fixed scoring rule under exchangeability. The evaluated pipeline instead reuses the calibration split to estimate per-cell normalization and some bank directions before setting the threshold. The theorem therefore does not establish an exact bound for this fitted procedure. Its realized false-alarm rate is measured on disjoint benign evaluation data, at 0.042–0.069 across the five targets against a nominal 0.05. Theorem 2 establishes power for the ideal cumulative statistic. Proposition 3 describes the deployed scan's response in expectation, not its detection probability, so its power is assessed empirically. *(ii)* Assumption 1 is not a derived property of transformers. Our experiments measure its marginal implication rather than establish the conditional claim, and the theorem is void where the assumption fails. *(iii)* Corollary 1 models the attacker as needing to accumulate a fixed budget along w while respecting one rail. It does not cover attacks that succeed without accumulating that budget along w, e.g., by obfuscating activations (Bailey et al., 2026); we flag that gap in §6 rather than claim coverage.

Table 5: Models and baselines used in the evaluation, grouped by role.

Role	Model	Size	Precision
Targets	`mistralai/Mistral-7B-Instruct-v0.3`	7.2B	bf16
	`Qwen/Qwen2.5-7B-Instruct`	7.6B	bf16
	`meta-llama/Llama-3.1-8B-Instruct`	8.0B	bf16
	`google/gemma-2-9b-it`	9.2B	bf16
	`microsoft/Phi-3-mini-4k-instruct`	3.8B	bf16
Scale ladder	`Qwen/Qwen3-0.6B`	0.6B	bf16
	`Qwen/Qwen3-1.7B`	1.7B	bf16
	`Qwen/Qwen3-4B`	4.0B	bf16
	`Qwen/Qwen3-8B`	8.2B	bf16
	`Qwen/Qwen3-14B`	14.8B	bf16
	`Qwen/Qwen3-32B`	32.8B	bf16
	`Qwen/Qwen2.5-72B-Instruct`	72.7B	int8
Guards	`allenai/wildguard`	7B	int8
	`ibm-granite/granite-guardian-3.0-8b`	8B	int8
	`meta-llama/Llama-Guard-3-8B`	8B	int8
Judge	`ibm-granite/granite-guardian-3.0-8b`	8B	int8
Trajectory baseline	LAD	—	—

Remark (detection is decoupled from causation). Suppose $w \perp w_{\text{ref}}$ and write $h = \langle h, w_{\text{ref}}\rangle w_{\text{ref}} + \langle h, w\rangle w + h^{\perp}$. If refusal behavior depends only on $\langle h, w_{\text{ref}}\rangle$, projecting out w leaves that coordinate unchanged. This geometric observation does not establish that either direction is causal. Here the orthogonality condition holds only approximately ($|\langle w, w_{\text{ref}}\rangle| \leq 0.09$ for the joint axis of Table 1, up to 0.32 for individual bank axes), and the premise that refusal depends only on w_{ref} is not established. The activation-addition controls do not establish a direction-specific causal effect (Appendix E). We therefore present DCD as a readout rather than a causal lever; the ablation null is consistent with the geometry but does not establish the causal role of w.

B Experimental setup and reproducibility

This appendix specifies the models, corpora, controls and settings used throughout the evaluation.

Models. Five instruction-tuned models serve as targets, seven Qwen checkpoints test scaling and three moderation models serve as guard baselines. Assistant turns are generated greedily by the same target whose representations DCD reads, since another model may comply with an attack that this target would refuse. Granite Guardian also serves as the response judge, but in response mode over target replies rather than as a guard over user turns. Table 5 gives the exact checkpoints, sizes and precisions.

Corpora. We evaluate five attack corpora, two structure-matched benign twins and three WildChat sets (Table 6). SafeMTData, SafeDialBench and cipher use published conversations. Crescendo and Echo Chamber publish attack procedures but no conversation data, so we generate both their attacks and twins. An attacker model, Mistral-7B-Instruct, writes each user turn from the conversation so far and a goal hidden from the target, while conversing with Qwen-2.5-1.5B-Instruct. As with the three published corpora, these user turns are replayed on every evaluated target, which generates its own replies, and success is judged per target. Harmful goals are pooled from JailbreakBench (Chao et al., 2024), AdvBench (Zou et al., 2023b) and HarmBench (Mazeika et al., 2024) (100, 509 and 191 goals) and split between the two families. Each twin applies the same procedure to one of 50 JailbreakBench benign behaviors, so it matches the attacks' procedure, scaffold and length rather than their individual goals. Appendix C documents full provenance and the contamination audit.

Ordinary chat and matched twins. WildChat supplies 400 real user conversations for alarm calibration and the guard comparison in §5.2, where every method is scored on the same traffic. Because attacks also differ from WildChat in length, goal-directedness and generator, this compar-

Table 6: Corpora, with the goal list each is built from. n is conversations.

Corpus	Role	n	Goals from
SafeMTData (Ren et al., 2025)	attack	600	published corpus
SafeDialBench (Cao et al., 2026)	attack	600	published corpus
cipher (Gibbs et al., 2024)	attack	517	published corpus
Crescendo (Russinovich et al., 2025)	attack	386	JBB, AdvBench, HarmBench
Echo Chamber (Alobaid et al., 2026)	attack	414	JBB, AdvBench, HarmBench
Crescendo twin	matched benign	50	JBB benign behaviors
Echo twin	matched benign	50	JBB benign behaviors
WildChat (Zhao et al., 2024)	ordinary benign	400	real user traffic
WildChat, sensitive topics	sensitive benign	400	real user traffic
WildChat, topic control	non-sensitive control	200	real user traffic

ison measures deployment separation but does not attribute it to intent alone. The Crescendo and Echo Chamber twins isolate intent more closely. Each preserves its family's procedure, scaffold, generator and turn count while replacing the harmful goal with a benign one, with 50 conversations per family. DCD reaches 0.957–0.971 AUROC against WildChat across the five targets (Table 2) and mean AUROC of 0.880 on Crescendo and 0.896 on Echo Chamber against the twins (§5.2). We therefore use WildChat for detection results and the twins for claims about intent.

Settings. The nominal false-alarm rate, DCD's only deployment setting, is fixed at $\alpha = 0.05$ throughout. For cross-family comparisons, every method is evaluated over the same turn horizon, chosen as the longest for which every family remains represented; shorter conversations are excluded. This prevents the maximum-over-turn score from exploiting conversation length, which alone separates the cipher corpus from WildChat at 0.998 AUROC. The common horizon controls conversation-length effects in Table 2 and the matched control of Appendix I. Rescue instead processes full conversations to measure the first alarm relative to the first harmful response, which on the cipher corpus arrives as late as turn thirteen, with results reported separately by family and target. In deployment, DCD updates after every turn and does not require the conversation length in advance. Layers are selected per family on the calibration split (Appendix E). All models were run on data-center GPUs at the numerical precision reported in Table 5.

Hardware. Experiments ran on machines with four NVIDIA A100 40 GB GPUs, two AMD EPYC 7452 CPUs (32 cores each) and 512 GB of RAM.

C Benchmark construction and contamination audit

Provenance and selection. The three released corpora differ substantially in size, so we subsample the larger two to keep family sizes comparable. We use all 600 SafeMTData conversations (Ren et al., 2025), which contain 4–5 turns and cover 200 goals, with three conversations per goal. Conversations from the same goal are not independent draws, so they remain together across data partitions and are resampled as one unit when computing uncertainty. The cipher release (Gibbs et al., 2024) contains 4,136 conversations of 6–13 turns across 517 goals; we select one conversation per goal, yielding 517 conversations. SafeDialBench (Cao et al., 2026) provides 2,035 English conversations of 3–9 turns across seven jailbreak strategies, from which we draw 600, stratified by strategy. Crescendo (Russinovich et al., 2025) and Echo Chamber (Alobaid et al., 2026) provide procedures rather than transcripts; Appendix B describes how their conversations are generated.

Benign traffic. WildChat (Zhao et al., 2024) supplies real multi-turn conversations, divided evenly into disjoint calibration and evaluation sets. The ordinary-chat sample was drawn without a language filter; roughly half is English, while all attack corpora are English. On the English subset of held-out WildChat, mean DCD AUROC across the target–family cells is 0.949, compared with 0.964 on the full held-out set. Both use the original calibration set. For the sensitive-topic test in Table 18, we draw a separate WildChat-1M sample and retain only conversations tagged English. These conversations have at least five user turns and are neither marked toxic nor redacted. The sensitive set

Table 7: Mechanism comparison with concurrent trajectory-level defenses. Direction means an explicit attack-mode direction in residual-stream space; gradient-free refers to fitting the detector.

Method	Signal	Access	Added call	Gradient-free	Direction
Per-turn guards	turn text	text only	—	✗	✗
LAD (Kulkarni, 2026)	magnitude features	internals	—	✗	✗
TrajGuard (Liu et al., 2026)	decode states	internals	—	✓	✗
DeepContext (Albrethsen et al., 2026)	turn embeddings	embeddings	—	✗	✗
SafeDream (Yan et al., 2026)	encoder + CPD	embeddings	—	✗	✗
THRD (Ma et al., 2026)	LLM-judge scores	text only	judge call	✓	✗
DCD (ours)	**context differential**	internals	one pass	✓	✓

requires at least two of the first five user turns to match a medical, legal, mental-health, drug or security lexicon; the topic controls satisfy the same filters but contain no lexicon match.

Why single-turn sets are excluded. DCD is not evaluated on single-turn prompt collections because its context differential is identically zero before any history exists (Proposition 1). Adding single-turn harmless prompts to the benign class would therefore inflate AUROC and calibrate an artificially low threshold for genuine multi-turn traffic. HarmBench (Mazeika et al., 2024) and AdvBench (Zou et al., 2023b) are not used as detection corpora; together with JailbreakBench (Chao et al., 2024), they supply the goals used to generate the procedure-based families. AdvBench harmful prompts and a matched set of harmless prompts are used separately to estimate the refusal direction in §4.

Contamination audit. The calibration split contains 200 WildChat conversations and no attack data by construction. We nevertheless audit whether any separately sourced corpus accidentally overlaps with it. Exact matching hashes the concatenated normalized user turns, while near-duplicate matching compares first-turn token sets at Jaccard similarity above 0.8. All five attack families return 0 exact and 0 near-duplicate matches against calibration. No audited corpus contains a duplicated conversation, and no two corpora share an exact fingerprint. Three of the 200 held-out WildChat conversations have a near-duplicate calibration opening, but none matches exactly. The near-duplicate check is lexical and restricted to first-turn openings.

Success labels. A conversation counts as a successful jailbreak only when at least one response generated by the target, rather than the attack script, is judged harmful by Granite Guardian in response mode using the full conversation through that response. Prompt-mode guards read the user's request rather than the reply, so they cannot say whether an attack landed. The first such response defines the harmful onset the rescue evaluation uses. To filter very short judge-positive outputs, a response must clear the judge's zero boundary and contain at least 40 characters. Removing this floor changes target–family success rates by at most 0.004 and leaves every reported DCD rescue rate unchanged. Requiring harmful generated content rather than non-refusal alone follows Souly et al. (2024). All labels come from one judge, and target responses are capped at 1024 tokens, so harmful content appearing only beyond that limit can be missed.

D Detector details and rule comparison

Other trajectory-level defenses scale attention over long conversations (Hu et al., 2026), accumulate stateful conversational risk (Mishra et al., 2026), or predict novel attacks in a multimodal setting (You, 2026). Across these methods and the concurrent defenses in Table 7, none combines a context differential with an explicit attack-mode direction in residual-stream space. DCD does both. We compare against LAD because, like DCD, it reads the target's hidden states at each user turn; TrajGuard reads hidden states while the reply is decoded, DeepContext and SafeDream train separate networks over turn representations, and THRD operates on judge scores rather than activations.

Specification. The grid $\mathcal{G}$ contains the 24 pairs formed by $a \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$ and $b \in \{0.05, 0.10, 0.15, 0.20\}$, plus 19 slower pairs formed by $a \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ and $b \in \{0.01, 0.02, 0.03, 0.05\}$ with $a > b + 0.01$. These 43 cells span effective memories from four

Table 8: Sequential alarm rules, rescue no later than the harmful turn at 5% benign FPR, averaged over 24 target–family cells. SHORT groups families with a median of five turns, LONG with medians of ten and eleven.

Decision rule	SHORT	LONG	Note
Multi-scale EWMA chart (primary)	**0.90**	**1.00**	scans timescales; only α
Single EWMA chart, a=0.5, b=0.1	0.85	0.98	the plain Roberts (1959) chart
CUSUM	0.67	0.90	one timescale, fixed drift
Mahalanobis distance (multivariate)	0.02	0.25	magnitude of the full vector
Page-Hinkley	0.02	0.08	baseline from the running stream
ADWIN-style window	0.01	0.08	baseline from the running stream

to about a hundred turns and are fixed across families rather than tuned per family. Per-cell benign moments μ_{ab}, σ_{ab} and the conformal quantile τ are estimated on the same benign calibration split. We use roughly $10/\alpha$ benign conversations for stable quantile resolution. Calibration traffic should reflect the conversation lengths expected at deployment, since the guarantee is marginal rather than conditional on length; substantial shifts require recalibration. In our latency benchmark the chart costs $\sim 5\,\mu$s per turn on CPU across all 43 cells, against ~ 33 ms for the added Mistral-7B forward pass on a single consumer GPU, so the chart is negligible relative to model inference. Because the conversation score is a running maximum, online and batch evaluation yield the same first crossing, and therefore the same detection turn and rescue lead.

Multiple comparisons. The detector takes the maximum over all grid cells and bank directions. We calibrate the conformal limit on this joint statistic rather than thresholding each component separately, so adding an axis requires recalibration but no separate multiple-comparisons correction. Theorem 1 guarantees false-alarm control for exchangeable traffic when the scoring rule is fixed before calibration. Our pipeline also uses calibration data to estimate normalization and some bank directions, so we evaluate its false-alarm rate on separate benign traffic. The observed rate is 0.042–0.069 across the five targets against a nominal 0.05 (Appendix A).

With the directional signal c_t held fixed, Table 8 compares standard sequential change-detection rules for turning its trajectory into an alarm (Basseville & Nikiforov, 1993; Tartakovsky et al., 2014). It also includes multivariate MEWMA on the full differential vector as the natural alternative to projection.

Why multi-scale EWMA. Table 8 compares rules on the first five turns of each conversation. The scalar rules use a family axis fitted on one half of the conversations and scored on the other; the multivariate rule instead uses the full normalized differential. This diagnostic scans a reduced 20-pair grid and counts alarms through the first harmful turn, so its rates are not directly comparable with the deployed-bank, strictly pre-harm rescue in Table 3. Gemma-2-9B on Echo Chamber has only four judged successes in the scored half and is omitted. A single EWMA is the strongest alternative, trailing the multi-scale chart by 0.05 on SHORT and 0.02 on LONG. The gain is small, but scanning avoids choosing one timescale in advance. CUSUM, which assumes one fixed drift scale, reaches 0.67 and 0.90. Page–Hinkley and ADWIN-style rules must estimate a baseline from at most four post-first-turn scores; neither exceeds 0.08. The multivariate Mahalanobis rule reaches 0.02 on SHORT and 0.25 on LONG.

Unknown onset. The chart does not assume when adversarial drift begins. Each cell compares a fast and a slow EWMA, so a new directional shift can separate them after an ordinary conversation has already taken place. Because model behavior can change over benign multi-turn interactions (Laban et al., 2026), the delayed-onset test uses real benign conversation turns rather than empty padding. Inserting zero, six or twelve benign turns leaves SafeMTData detection unchanged across four targets, while SafeDialBench varies across targets (Appendix L). Delayed-onset robustness is therefore family-dependent in these tests rather than universal.

Table 9: Magnitude against direction, read from the same differentials at the same layer. Scored per turn as AUROC against each family's structure-matched twin and, for contrast, against generic chat. Norms are conversation means.

			vs. matched twin		vs. WildChat	
Target	Family	$\|\delta\|_{\text{twin}}/\|\delta\|_{\text{atk}}$	Magn.	**Dir.**	Magn.	**Dir.**
Mistral-7B	Crescendo	1.03	0.464	**0.832**	0.345	**0.999**
	Echo Chamber	1.02	0.493	**0.813**	0.664	**1.000**
Llama-3.1-8B	Crescendo	1.06	0.425	**0.865**	0.228	**0.999**
	Echo Chamber	0.98	0.533	**0.819**	0.484	**0.999**
Qwen-2.5-7B	Crescendo	0.96	0.591	**0.876**	0.231	**1.000**
	Echo Chamber	0.94	0.641	**0.819**	0.381	**1.000**
Phi-3-mini	Crescendo	0.99	0.515	**0.862**	0.381	**1.000**
	Echo Chamber	0.98	0.552	**0.829**	0.515	**1.000**
Gemma-2-9B	Crescendo	1.01	0.490	**0.878**	0.324	**1.000**
	Echo Chamber	0.98	0.533	**0.842**	0.526	**1.000**

E SCORING AND READOUT ABLATIONS

WHAT THE DIFFERENTIAL CARRIES

Direction separates consistently; magnitude varies around chance. The differential δ_t records both how far context moves the representation and the direction in which it moves. We score these components separately, per turn, against each family's structure-matched twin. Magnitude is measured by $\|\delta_t\|$, while direction is measured by $\langle u_t, w \rangle$, with u_t the unit differential. Both use the same forward passes and layer, so the comparison changes only the readout applied to the differential (Table 9).

Magnitude reaches 0.641 at best and falls below chance in four of the ten cells. The twin's mean norm lies between 0.94 and 1.06 times the attack's, so attacks and twins move similar distances. Directional scoring reaches 0.813–0.878 on the identical conversations, at each cell's best layer under five-fold cross-validation. Against WildChat rather than the twin, direction saturates at 0.999–1.000 while magnitude does not, at 0.228–0.664. The twin is therefore the stricter control, removing differences in conversation form that make generic traffic easier to separate. Magnitude is computed from the same differential the detector scores, at the same layer, and cross-validated over conversations, yet provides weaker and less consistent separation. The structure-matched LAD comparison of §5.2 reaches the same conclusion using magnitude features learned from attack labels.

Cosine versus dot product. DCD scores the cosine $\langle u_t, w \rangle$. The dot product $\langle \delta_t, w \rangle$ multiplies it by the magnitude $\|\delta_t\|$, which Table 9 scores on its own, turn by turn. This comparison instead applies both scores to whole conversations through the deployed chart, with attacks compared against their twins in five-fold cross-validation at the layer where the cosine scores best. Their AUROCs differ by at most 0.034 across the ten target–family cells, and the dot product is higher in only three, so both rank conversations almost identically. At the chart's 5% conformal limit, however, the cosine has the higher true-positive rate in all ten cells, by 0.223–0.492 on Crescendo and 0.002–0.029 on Echo Chamber. Fitting the axis itself on raw differentials widens the largest gap to 0.510. Magnitude varies with how strongly a turn depends on prior context, which widens the spread of benign scores and raises the calibrated limit, so attacks keep their ranking but cross the limit less often. Scored against the WildChat conversations that calibrate the chart rather than held-out chat, every variant reaches 0.992–1.000 on both metrics, so the difference appears only under the twin control. We therefore use the cosine throughout.

Random-direction and magnitude controls. On whole conversations passed through the deployed chart, we compare the learned direction with random directions and with magnitude. Across the ten cells, random unit directions yield median AUROCs of 0.440–0.550, compared with 0.879–0.935 for the learned direction, each read at the cell's best layer under five-fold cross-validation. A random direction provides a null; the substantive control is the magnitude readout $\|\delta_t\|$, which

uses the same differential without its direction. At the common evaluation horizon, which restricts both families to the number of turns available in the shorter one, magnitude reaches 0.324–0.611 AUROC, and DCD exceeds it by 0.304–0.584 in every cell. Scoring Echo Chamber over its full ten turns lifts magnitude on all five targets and narrows DCD's lead on three, yet DCD still leads everywhere, by at least 0.167. Crescendo's natural length defines the common horizon, so its DCD–magnitude gap changes by at most 0.005. Thus, DCD's advantage over magnitude persists at the natural lengths of both twin-controlled families.

The mechanism contrast holds across cells. The mechanism analysis of §4.2 and §4.3 covers all ten cells formed by five targets and the two twin-controlled families, Crescendo and Echo Chamber. At each target's detection layer, we fit the drift direction on half the goals and evaluate it on the other half, keeping conversations that share a goal in the same split. The isolated refusal score ranges from 0.42 to 0.78 AUROC, whereas the differential along w reaches at least 0.997 in every cell; projecting that differential onto refusal does not separate the classes. These are turn-level diagnostics rather than detection results, since the deployed statistic is the conversation-level chart of §5. The primary comparison, also plotted in Figure 1, uses 288 benign conversations with the same escalating five-turn structure but a different generator. Repeating it against each family's same-generator twin yields differential AUROC of 0.79–0.86 at the detection layer.

Where the signal is read

Readout position and pooling. DCD reads the residual-stream state at the final input position immediately before reply generation, after the model has processed the system prompt, conversation history and current user turn. Pooling the last $K = 8$ positions does not improve this readout. Against the matched twins on three targets, pooling changes conversation-level AUROC by -0.060 to $+0.012$ and lowers true-positive rate in all six cells when both readouts use the same layer. Because that layer is selected for the single-position readout, we also test whether pooling benefits from another layer. Each readout is reported at the layer maximizing its own AUROC on the evaluated data, an oracle comparison separate from DCD's held-out layer selection. Even then, pooling improves AUROC by at most 0.012. It raises true-positive rate in two cells by at most 0.026 and lowers it in four, dropping Mistral Crescendo from 0.863 to 0.472.

Layer selection. Layer selection uses three disjoint partitions. The first fits a candidate direction at each layer. On the second, the maximum raw directional score across turns and bank axes gives one score per conversation, and the layer with the highest AUROC is selected. The untouched third provides the reported EWMA results (Appendix B). Single-turn refusal separation (Arditi et al., 2024) is a poor alternative selection criterion because it locates where the refusal direction is most legible. The drift directions retain only 0.03–0.09 of their norm under projection onto the refusal–harmfulness plane (§4.1), so the two criteria target different features and need not select the same depth. Across ten architecture-family cells, their rank correlation averages $+0.03$ and is negative in five, indicating little association. Selected depths also vary across families. For the three published families, selected layers fall at 24–36% relative depth in ten of fifteen cells; for Crescendo, they fall at 93–100%. Arditi et al. (2024) ablate across layers because later computation can reconstruct a feature removed at only one layer. Detection only reads the feature, so one legible layer suffices. Averaging all cached layers changes conversation-level AUROC by -0.03 to $+0.01$ across five targets and both twin-controlled families, within the variation across selection seeds. We use one layer to reduce cost, not because averaging consistently harms performance.

Why each axis has its own layer. Across all pairs of extracted depths, axes fitted from the same conversations have a median cosine of 0.218. By comparison, two estimates fitted at the same depth on disjoint halves agree at 0.720–0.907, against a chance floor of $1/\sqrt{d} \approx 0.017$. Changing depth therefore changes the fitted direction, not merely how clearly a fixed direction can be read. Sensitivity to depth also differs across families. Conversation-level AUROC varies across depth by 0.00–0.07 for the three published families, by 0.09–0.43 for Crescendo and by 0.05–0.27 for Echo Chamber. Crescendo is more depth-sensitive in four of five cells, while Echo Chamber overlaps the upper end of the published-family range. Crescendo reaches its best layer in the upper half of every target, whereas the published families are already at ceiling around one-third depth. Figure 2 shows the depth curves for the two twin-controlled families, Crescendo and Echo Chamber. A single layer

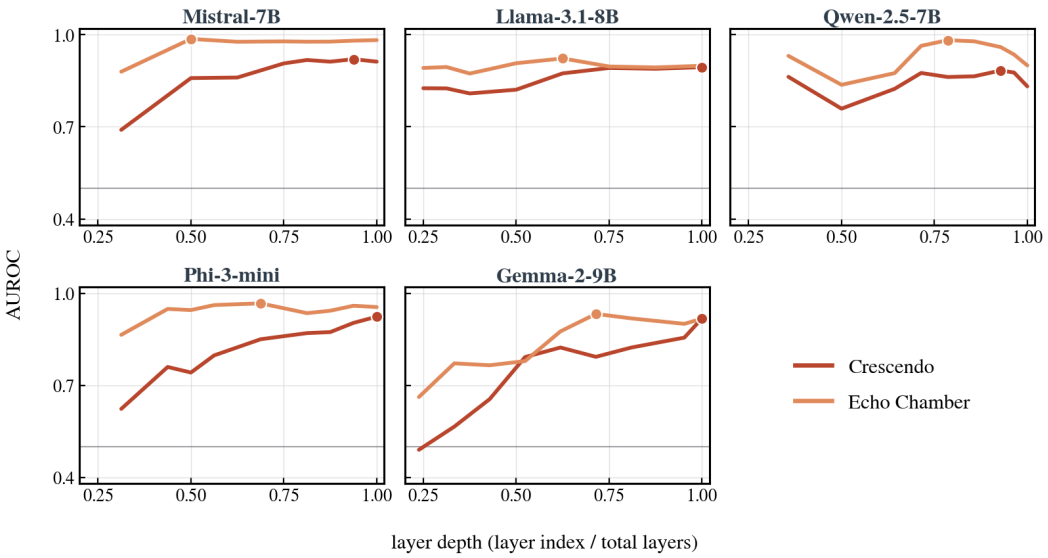


Figure 2: Conversation-level AUROC across depth for the two twin-controlled families. Dots mark the layers selected on held-out data.

per model would therefore compromise axes that are most legible at different depths, so the bank stores a selected layer with each axis.

Robustness of the calibration

Does a change of system prompt break calibration? Because the system prompt appears in both reads, changing it may shift the benign score distribution. We therefore test whether a limit calibrated under one prompt transfers to another. We replay the same 200 benign conversations under "You are a helpful assistant." and "You are a helpful, harmless, and honest assistant.", while regenerating the target's replies. This produces 903 paired turns on each target. The geometry summaries use the 703 turns after turn one; turn one is excluded because $\delta_1 = 0$ and therefore has no direction (Proposition 1). The differential changes measurably under the new prompt, with median relative differences of 34% on Mistral-7B and 26% on Phi-3-mini. Its direction remains largely preserved, with paired cosines of 0.941 and 0.967. The normalized projection of the prompt-induced change onto w is only 0.062 and 0.050, compared with isotropic baselines of 0.016 and 0.018, so most of the shift is off-axis. Consistent with this geometry, the realized false-positive rate is 0.040–0.060 across the four cross-prompt transfers and $9/200 = 0.045$ in all four within-prompt cells, against a nominal 0.05. All four cross-prompt Clopper–Pearson intervals include the nominal rate. This establishes transfer across one prompt substitution on two targets, not invariance to arbitrary system prompts. The direction w is fitted only from Crescendo and Echo Chamber against a matched benign twin, independently of the benign conversations used in this test.

A readout, not a lever

Causal controls. We test whether w causally controls refusal using standard activation-addition and ablation protocols (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023a). Removing the component along w does not change compliance. Small additions produce no consistent change in refusal across three targets. At larger scales, the effects of the drift direction, the reconstructed refusal direction of Arditi et al. (2024) and random controls overlap. On Llama-3.1-8B, refusal reaches 0.462 for w, 0.417 for the refusal direction and 0.438–1.000 across three random directions. On Mistral-7B, large additions reduce generation coherence across directions. The refusal-direction control therefore does not clearly separate from the random controls, so this experiment cannot distinguish an effect specific to w from a nonspecific consequence of large activation perturbations. We treat this activation-addition result as inconclusive about causality and use w only as a diagnostic readout.

Interchange intervention. We next test whether the coordinate DCD reads also controls harmful compliance, using a targeted edit rather than the scaled activation additions above. We select 41 conversations that produced judge-confirmed jailbreaks in the original evaluation and replay each under every intervention condition. During the forward pass that generates the response to the final payload turn, we replace the residual-stream coordinate along w at the readout layer with the mean value carried by benign conversations. This leaves all orthogonal components unchanged and determines the edit size from the data. A response-mode judge scores each resulting reply, with conditions paired within conversation. On replay without an intervention, compliance is 0.293 on Mistral-7B and 0.098 on Llama-3.1-8B.

Patching the drift coordinate changes compliance by $+0.049$ $[+0.000, +0.122]$ on Mistral-7B and $+0.000$ on Llama-3.1-8B. A random direction shifted by the same L_2 amount changes it by $+0.073$ and -0.024, respectively, so the drift edit produces no larger effect than the matched random control. The refusal coordinate serves as a positive control and succeeds on Llama but not Mistral. On Llama, setting it to its benign value raises compliance from 0.098 to 0.317, with nine paired conversations flipping toward compliance and none in the opposite direction ($p = 0.004$). This is the expected direction because benign contexts carry a lower refusal coordinate. On Mistral, the same edit changes compliance by $+0.049$, indistinguishable from the drift and random controls. The Llama result shows that the intervention can alter compliance through the refusal coordinate while finding no effect along the drift coordinate; the Mistral result remains inconclusive.

The patch sets the drift coordinate to its benign value, so a causal role predicts a reduction in compliance. The low no-intervention replay rates leave limited room for such a reduction, weakening the drift null. Although the successful Llama positive control shows that the design can detect an increase in compliance, it does not remove this floor for the predicted downward effect. Because conversations were selected by their original jailbreak success, the replay rates characterize this intervention set rather than corpus-level attack success. Finally, changing one scalar coordinate at one depth tests only that coordinate. It does not exclude causal effects requiring coordinated intervention across layers or a larger subspace.

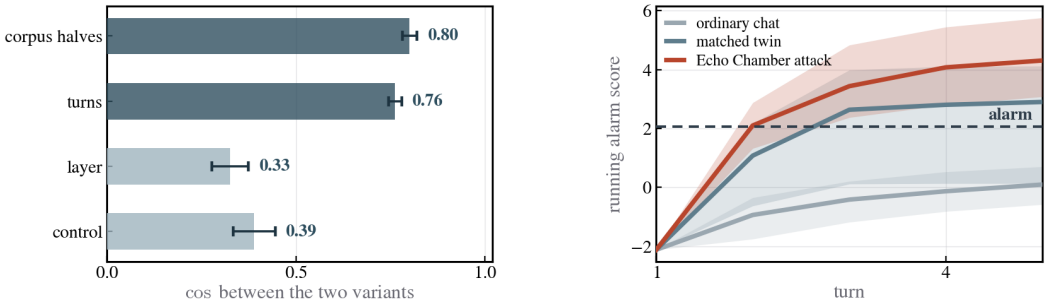
The value of the second read. The differential costs one additional forward pass. Table 10 compares it with the isolated and contextual readouts before any history exists and over turns 2–6, or fewer when a conversation is shorter. The second read removes history-free separation while preserving nearly all the later-turn AUROC of the single reads. For each target–family cell, directions are fitted on turns 2–6 of one quarter, layers are selected separately for each readout on a second quarter and results are reported on the untouched half, scoring each conversation by its turn-1 projection and by its maximum projection over turns 2–6. On turns 2–6, the differential reaches 0.862 AUROC, only 0.015 below the isolated read and 0.006 below the contextual read, with both paired intervals spanning zero. At turn one, however, the same fitted single-read directions already reach 0.777 and 0.724, whereas the differential is exactly 0.500 because $\delta_1 = \mathbf{0}$. Separation before any history exists is separation of the request text. The twins match procedure, scaffold, generator and turn count but use benign goals, so they do not hold that text fixed. In Section 4.3, replacing an attack's history with a benign conversation while holding the final request fixed reduces harmful compliance by 0.25–0.44 and DCD firing from 0.88–1.00 to 0.06–0.15, while the isolated read of the request does not change. The isolated read therefore cannot tell the two versions of the same request apart, although the model's behavior differs between them. For the contextual read, firing also falls to 0.09–0.27 under the same replacement (Appendix J). It also fires on 0.13–0.31 of these requests presented alone, so its firing does not require accumulated history. DCD scores the normalized difference between the contextual and isolated representations of the same request. That difference is zero when no history exists. The second read therefore provides a direct measure of the representation change induced by history.

F Direction estimation: is the axis stable?

Is w an artifact of the sample? Treating a concept as a direction follows the linear representation hypothesis (Park et al., 2024). The estimator is a difference of means, so a law of large numbers argument is available but not sufficient, since the question is whether *this* data yields a stable axis. For a direction pooled over the five attack families, its cosine with the full-data estimate rises from 0.84 at $n = 10$ to 0.94 at $n = 40$ and 0.985 at $n = 160$, while the across-seed standard deviation

Table 10: Isolated, contextual and differential readouts against the matched twins. Conversation-level AUROC from the turn-1 score, where no history exists yet, and from the maximum score over turns 2–6.

Readout	turn 1 (no history)	max over turns 2–6
isolated h_t^{iso}	0.777 [0.733, 0.819]	0.877 [0.855, 0.901]
contextual h_t^{ctx}	0.724 [0.675, 0.769]	0.868 [0.852, 0.884]
differential $h_t^{\mathrm{ctx}} - h_t^{\mathrm{iso}}$	0.500	0.862 [0.838, 0.888]
paired AUROC difference, single read minus differential		
isolated − differential	+0.277 [+0.233, +0.316]	+0.015 [−0.017, +0.053]
contextual − differential	+0.224 [+0.175, +0.269]	+0.006 [−0.021, +0.040]



(a) What the axis depends on.

(b) Echo Chamber on Mistral-7B, median and interquartile range.

Figure 3: The direction, and what accumulates along it.

falls from 0.03–0.06 to 0.004–0.005. With the benign reference fixed, attack samples drawn from disjoint goal pools agree at 0.901–0.915 with 40 conversations per fit and 0.956–0.974 with 160. A stronger check also splits the benign traffic, so the two estimates share no conversation, attack goal or benign reference. Agreement remains 0.859–0.873 with 40 conversations per fit and 0.914–0.930 with 160. Independently shuffled labels give 0.21–0.36 at $n = 40$ and 0.14–0.35 at $n = 160$, while directions constructed only from disjoint benign subsets agree at 0.08–0.12. Shared fitting data therefore do not explain the high agreement.

How much fitting data does detection need? The previous check measures the direction; this one measures the decision. We refit the deployed Crescendo and Echo Chamber axes from n attack conversations on three targets, holding the scored conversations out of the fit and recalibrating the conformal limit at each n. Using 40 of the 341–372 conversations that remain after holding out the scored successes brings rescue at any turn within 0.06 of the full-corpus rate in all six target–family cells, and rescue before the first harmful response within 0.07.

What the axis does and does not depend on. We vary one fitting choice at a time across the ten target–family cells with a valid structure-matched twin, Crescendo and Echo Chamber on five architectures. The split-half reference, averaged over the ten cells, is 0.828, the cosine between estimates of the same axis fitted on random disjoint halves at the same sample size; the chance scale is 0.017, averaged from $1/\sqrt{d}$ across architectures. The attack-half cosine quoted in §5.4 instead splits only the attack conversations against a shared benign mean, so it runs higher. The two attack halves taken in corpus order agree at 0.801, while axes fitted on early and late turns agree at 0.762. Both are within 0.08 of the reference (Figure 3a). By contrast, averaged over the same cells, replacing the matched twin with generic chat lowers agreement to 0.390 (67°), while moving to a second read depth lowers it to 0.325 (71°). The recovered direction is therefore specific to one attack mode, read at one depth and estimated against one control, which is why the detector maintains separately fitted axes.

What accumulates along the axis. Figure 3b shows the running alarm score for Echo Chamber on Mistral-7B against the conformal limit at $\alpha = 0.05$, with the median and interquartile range over conversations and the limit calibrated on ordinary chat of the same length. The ordinary-chat median stays below the limit. The matched-twin median crosses at turn 3 and stays above it, while its band overlaps the attack's for most of the conversation. Because the twin keeps the attack's procedure, scaffold, generator and turn count but uses a benign goal, it can accumulate similar drift and cross a limit calibrated on ordinary chat. Accordingly, the same target–family cell reaches 0.813 turn-level AUROC against the twin but 1.000 against ordinary chat (Table 9).

Convergence is not coverage. The preceding checks establish stability only within the sampled attack distribution; they do not address modes that steer in new directions. Cipher illustrates the gap. Its nearest remaining bank axis is 82° away on average, against the 90° orthogonality reference; Table 13 reports the per-family angles and the corresponding AUROCs. With the cipher direction withheld, the bank reaches 0.363–0.924 AUROC across the five targets; adding that direction raises the identical conversations to 1.000. A single axis pooled over the other four families performs worse. Its cosine with the cipher axis is negative on four targets, where AUROC falls to 0.136–0.265. Stable estimation of one mode's direction therefore does not guarantee coverage of a mode that steers elsewhere. The failure is geometric rather than an estimation failure, which is why the bank keeps one axis per known mode. Needing examples of a mode before covering it is an open-set boundary inherited by detectors learned from prior examples, long recognized in machine-learning-based intrusion detection and formalized in open-set recognition (Sommer & Paxson, 2010; Scheirer et al., 2013). DCD makes that boundary measurable. Once examples define a candidate mode, its angle to the existing bank predicts the open-set loss without running detection.

G Where the axis sits relative to single-turn safety

This appendix asks whether attack drift directions are combinations of the single-turn safety directions. If refusal and harmfulness spanned a drift axis w, projecting it into their plane would retain most of its norm. The projection is interpreted using two references beside every cell, a positive control that should land high and a shuffled-label null showing what finite-sample noise produces in the same geometry.

The safety plane is model-specific. Figure 4 applies the projection analysis of Table 1 separately to each attack family. Each panel shows one target's plane, spanned by that model's refusal and harmfulness directions. Because models have different residual spaces, the panels cannot be superimposed. Each panel therefore reports the fraction of an axis retained in its own model's plane, interpreted against that model's shuffled-label null rather than compared directly across panels.

Safety directions project more strongly. For the positive control, half of the contrastive prompts define the safety plane and the disjoint half estimates new refusal and harmfulness directions. Their projections retain 0.79–0.96 and 0.79–0.91 of their norms, respectively, showing that a genuine safety direction registers even when estimated from unseen prompts. We then project each attack axis into the plane estimated from all prompts, matching the construction used in Table 1. Across the twenty-five target–family cells, the retained norm is only 0.015–0.320.

Nearly all attack axes fall within the null. The gray cloud provides the relevant reference. We permute the attack and benign labels and refit the axis 300 times per target and family, producing directions with no label signal. Their median projection is 0.069, and their 95th percentile is 0.212. A uniformly random direction would project at only 0.022–0.026 in these dimensions. The shuffled-label null sits higher because it is built exactly as the attack axis is, in the same anisotropic geometry (Ethayarajh, 2019) and with the same estimation noise. The attack axis should therefore be compared with this null, not the random-vector floor. Twenty-four of the twenty-five cells fall within their own null's 95% region. Points mark the projected endpoints. Their distance from the origin, read against the null, is comparable across models; their angular position is not. The sixth panel magnifies the central region marked in each model panel and shows its overlap with the shuffled-label cloud. The one exception is cipher on Qwen-2.5-7B, at 0.320 against that family's 95th percentile of 0.260.

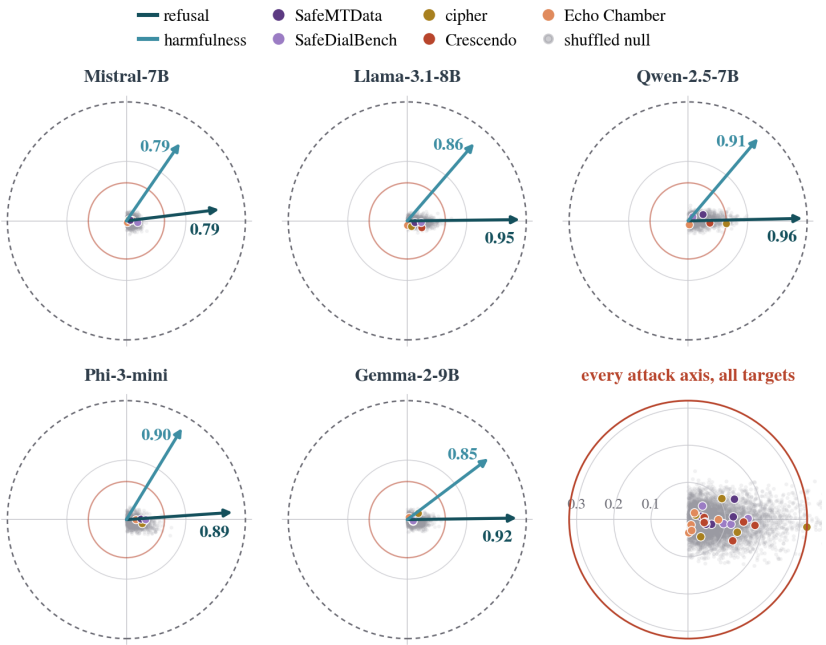


Figure 4: Every axis projected into its own target's single-turn safety plane. Arrows are held-out refusal and harmfulness; points are the five attack families; gray is 300 shuffled-label draws per family.

Safety directions do not span attack drift. Almost any direction lies mostly outside a two-dimensional plane at these dimensions, so a small projection alone does not establish usefulness. What the positive control and the null establish is that refusal and harmfulness do not span the measured drift axes. The null does not establish exact orthogonality. Across the five headline targets, the 95th percentile of the Table 1 null ranges from 0.08 to 0.25, and the per-family nulls above run up to 0.35, so small nonzero projections would remain unresolved. Detection along those axes is evaluated separately in §5.2.

The result holds in higher-dimensional safety subspaces. Refusal can occupy a multi-dimensional cone (Wollschläger et al., 2025), and safety behavior spans several orthogonal directions (Pan et al., 2025). We therefore extend the refusal–harmfulness plane with the top principal components of the per-prompt harmless-to-harmful shift at both readout positions, computed after removing the plane, giving nested subspaces S_k from 50 harmful and 50 harmless prompts; at $k = 2$, S_k is the plane of Table 1. Figure 5 shows how each direction's retained norm grows with k. At $k = 48$, which captures 68–76% of the remaining shift variance, w retains 0.19–0.35 of its norm, leaving at least 88% of its squared norm outside the subspace, while held-out refusal and harmfulness directions from disjoint prompts retain 0.87–0.98. On Qwen-2.5-7B, Llama-3.1-8B, Gemma-2-9B and Phi-3-mini, w stays within its shuffled-label null at every k ($p \geq 0.23$). On Mistral-7B, w shows a small excess over its null for $k \geq 8$ (unadjusted $p = 0.043$–0.053) while retaining at most 0.28 of its norm.

Post-training directions do not span the drift either. Following Pan et al. (2025), we also take the directions along which post-training moved the representation. For the four targets with a public base checkpoint (Table 11), we compare base and instruct activations on the same prompts at the same token positions, subtract the mean shift on harmless prompts from the shifts on harmful ones, and keep the leading singular directions. The refusal direction estimated from the same prompts retains 0.68–0.90 of its norm in the top 4–16 of them, yet w retains only 0.06–0.20 in their span with the plane and stays within its shuffled-label null at every k up to 16 ($p \geq 0.29$). The checkpoint difference reflects post-training as a whole rather than safety tuning alone.

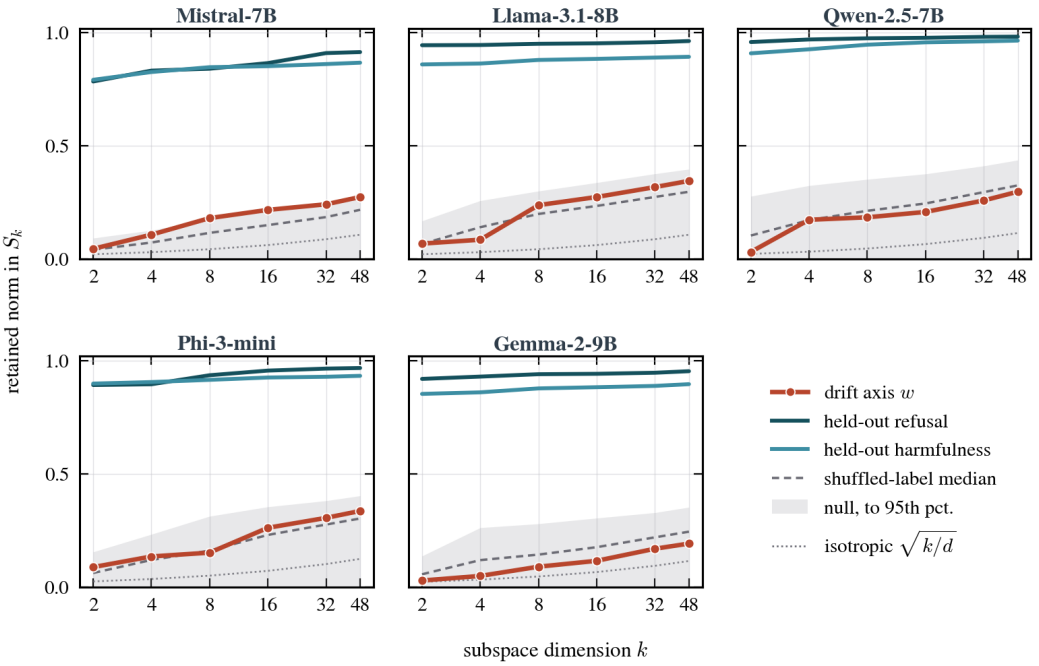


Figure 5: Retained norm inside the k-dimensional safety subspace S_k. At $k = 2$, S_k is the plane of Table 1; shading marks the shuffled-label null up to its 95th percentile.

Table 11: Alignment-shift subspace A_k: norm of the refusal direction retained in A_k, and of w retained in A_k with the plane, with the null median in parentheses.

	$k = 4$			$k = 8$			$k = 16$		
Target	ref. in A_k	w (null)	p	ref. in A_k	w (null)	p	ref. in A_k	w (null)	p
Mistral-7B	0.680	0.059 (0.091)	0.87	0.696	0.137 (0.120)	0.29	0.724	0.156 (0.139)	0.29
Qwen-2.5-7B	0.776	0.085 (0.140)	0.86	0.792	0.122 (0.168)	0.82	0.806	0.166 (0.204)	0.77
Llama-3.1-8B	0.865	0.081 (0.096)	0.70	0.870	0.183 (0.156)	0.32	0.881	0.202 (0.196)	0.45
Gemma-2-9B	0.884	0.072 (0.100)	0.81	0.893	0.093 (0.139)	0.87	0.902	0.106 (0.163)	0.95

The bank axes retain little safety-plane energy. We extend the main-text projection analysis from the joint Crescendo–Echo Chamber direction of Table 1 to all five mode-specific axes in the DCD bank. For unit-normalized axes, the squared projection gives the fraction of their energy inside the safety plane. Across all five bank axes, 24 of 25 target–family cells retain only 0.02%–3.20%, compared with 62%–92% for held-out safety directions. The remaining cipher–Qwen-2.5-7B cell retains 10.2% and is the null exception identified above. These energy bounds establish little overlap with the safety plane, not exactly zero overlap. We bootstrap by conversation rather than by turn. To limit finite-sample upward bias, we estimate overlap by split-half cross-projection; with more than 400 conversations, it agrees with the plug-in estimate to three decimals.

Both results hold on newer and larger models. To test a newer model family and a wider scale range, we evaluate six Qwen-3 models from 0.6B to 32.8B and add Qwen-2.5-72B to reach 72.7B, beyond the largest available dense Qwen-3 checkpoint (Table 12). Each cell uses a model-specific null from 200 conversation-level label shuffles. The projection falls within its null distribution on all seven, with a minimum permutation p of 0.10, while a held-out refusal direction retains 0.89–0.99 of its norm. At the same selected layers, the two-axis Crescendo–Echo Chamber bank reaches 0.843–0.895 matched-twin AUROC, so the off-plane geometry and the detection separation both persist across the tested range.

Table 12: Safety-plane projection norms and matched-twin AUROC across model scale.

Model family	Size	L	w proj.	null med.	perm. p	pos. ctrl. (ref./harm.)	AUROC
Qwen-3	0.6B	26	0.123	0.101	0.40	0.91 / 0.95	0.849
Qwen-3	1.7B	16	0.155	0.119	0.37	0.98 / 0.97	0.850
Qwen-3	4.0B	24	0.058	0.112	0.83	0.98 / 0.97	0.895
Qwen-3	8.2B	20	0.044	0.126	0.94	0.99 / 0.98	0.895
Qwen-3	14.8B	15	0.088	0.076	0.42	0.89 / 0.93	0.843
Qwen-3	32.8B	56	0.103	0.083	0.35	0.98 / 0.96	0.888
Qwen-2.5	72.7B	48	0.108	0.051	0.10	0.95 / 0.94	0.883

Table 13: Coverage against the angle to the rest of the bank. AUROC, averaged over the five targets.

Family	angle to bank	without own axis	with own axis
SafeDialBench	44°	0.953	0.980
SafeMTData	46°	0.977	0.996
Crescendo	63°	0.883	0.991
Echo Chamber	66°	0.873	0.991
cipher	82°	0.624	**1.000**

H THE COVERAGE BOUNDARY

Coverage follows distance to the bank. We test whether the bank can detect a family when its own axis is absent. For each target and family, we fit that axis on half the attack conversations and score the other half against held-out WildChat, using the same conversations with and without the axis. We select the layer on part of the fitting half using only the remaining bank. Crescendo and Echo Chamber also exclude each other from that bank because they share a generator. Scores take the maximum raw projection over five turns and the available axes, with every axis fitted against WildChat; Table 2 instead reports the deployed chart. Table 13 relates AUROC to the signed angle $\arccos\langle w_f, w_o\rangle$ to the nearest remaining axis, averaged over five targets. Without the family's own axis, AUROC generally falls as the angle grows, reaching 0.624 for cipher at 82°. Adding its axis raises AUROC to 0.980–1.000 on the same conversations. Holding the conversations fixed shows that the loss comes from directional coverage rather than family difficulty.

WHY POOLING NONALIGNED MODES LOSES SIGNAL

For nonaligned modes, the need for a bank follows directly from the estimator's geometry. Equation 4 fits an axis as the difference between attack and benign means. Pooling multiple modes replaces their distinct mean shifts with an average, so nonaligned components partially cancel. The calculation below quantifies this loss under equal mode weights and equal shift norms.

Write $v_i = \mu_i - \mu_b$ for the centered mean differential of mode i, with $\|v_i\| = m$, and let ρ be the average pairwise cosine among the k pooled modes. Let $w_{\text{pool}} = \sum_i \pi_i v_i / \|\sum_i \pi_i v_i\|$ with $\pi_i = 1/k$. The mean retention across the pooled modes and the retention of an absent mode s are

$$\bar{R}_{\text{in}} = \frac{1}{k}\sum_{i=1}^{k} \frac{\langle v_i, w_{\text{pool}}\rangle}{\|v_i\|} = \sqrt{\frac{1+(k-1)\rho}{k}}, \qquad R_s = \frac{\langle v_s, w_{\text{pool}}\rangle}{\|v_s\|} = \frac{\rho_s\sqrt{k}}{\sqrt{1+(k-1)\rho}}. \tag{7}$$

where ρ_s is that mode's average cosine with the pooled modes. The two limits provide a sanity check. When $\rho \to 1$, identical modes give retention 1, so pooling costs nothing. When $\rho = 0$, retention is $1/\sqrt{k}$, so the loss grows with the number of nonaligned modes.

Once k, ρ and ρ_s are measured, Equation 7 gives a prediction without fitted coefficients, which Table 14 tests directly. We test Equation 7 using the four non-cipher families as the pool ($k = 4$) and cipher as the absent mode. We compare its sign with the AUROC of the pooled axis.

Cipher's average cosine with the four pooled directions is negative on four targets and near zero on Phi-3-mini, so Equation 7 predicts negative or negligible retention. Correspondingly, the pooled axis reaches only 0.136–0.265 AUROC on the four negatively aligned targets and 0.502 on Phi-3-mini,

Table 14: The retention identity against measurement. The four non-cipher families form the pool; cipher is the mode held out of it. Equation 7 has no free parameters, so the predicted column is not fitted.

Target	ρ pool	ρ_s	predicted	AUROC pooled $\rightarrow$ own
Mistral-7B	0.369	-0.067	-0.092	$0.265 \rightarrow 1.000$
Llama-3.1-8B	0.526	-0.109	-0.135	$0.136 \rightarrow 1.000$
Qwen-2.5-7B	0.470	-0.058	-0.075	$0.245 \rightarrow 1.000$
Phi-3-mini	0.264	$+0.014$	$+0.021$	$0.502 \rightarrow 1.000$
Gemma-2-9B	0.295	-0.057	-0.083	$0.208 \rightarrow 1.000$

whereas cipher's own axis reaches 1.000 throughout. Pooling can therefore reverse a mode whose mean shift opposes the pooled direction, rather than merely weakening it.

The connection to Theorem 2 is through the scoring axis. Each mode induces its own attack mean and margin relative to benign traffic. Pooling rotates the axis and can shrink or reverse that margin, as Table 14 shows. The theorem still applies to a pooled axis whenever its conditional attack and benign means satisfy Assumption 1; the bank is not a precondition for the theorem but helps preserve its separation assumption across nonaligned modes.

Equation 7 assumes equal mode norms and weights. Its sign matches the pooled AUROC on all five targets, below chance where retention is predicted negative and at chance on Phi-3-mini. Because cipher's own axis reaches 1.000 AUROC everywhere, an effect-size ratio against it is undefined, so the identity is tested on sign rather than magnitude; it explains whether pooling retains or reverses directional alignment.

I THE MATCHED-TO-CHANCE CONTROL, IN BOTH DIRECTIONS

This control asks whether DCD and the guard baselines rely on the same evidence. Each attack conversation is paired with a held-out WildChat conversation receiving nearly the same score from one method, suppressing that method's separation on the matched sample, after which we measure what the other method still separates. We first match on the guard score and evaluate both DCD bank conditions. We then reverse the control for the own-axis-withheld bank, matching on its DCD score and evaluating the guard. These are conditional diagnostics rather than head-to-head comparisons; the direct comparison appears in §5.2.

Matching is performed without replacement and repeated over 30 random attack orderings so that no attack is systematically matched first. Pairs exceeding the caliper, the maximum permitted score difference, are discarded, and the tables report the surviving pair count and mean AUROC across orderings. Both members of each pair are scored over the same evaluation horizon, removing conversation length as a source of separation. Because finite-sample matching does not guarantee an AUROC of exactly 0.50, the matched method's realized AUROC is reported alongside the other method's result.

Table 15 matches on the guard score and measures DCD using both the deployed bank and a bank with the tested family's axis withheld. In the deployed condition, that family's axis is fitted on one half of its conversations and the other half is scored, which reduces the attack pool available for matching; withholding the axis gives the open-set condition of §5.3. Table 16 reverses the control for the own-axis-withheld DCD bank by matching on its score and measuring the guard. Figure 6 illustrates the guard-matched control on Mistral-7B. Each gray line connects an attack to a benign conversation with nearly the same guard score, making the pairs nearly vertical while any remaining DCD separation appears on the vertical axis.

WildGuard remains at 0.495–0.501 on its matched pairs, while deployed DCD reaches 0.898–1.000 across all six cells. Withholding the tested family's axis leaves DCD at 0.886–0.941 on SafeMTData and SafeDialBench but reduces it to 0.195–0.417 on cipher. Restoring the cipher axis raises both cells to 1.000. The contrast separates two findings. Deployed DCD uses evidence not captured by the guard score, while coverage by the remaining bank axes depends on the attack mode, as examined in §5.3.

Table 15: Guard-matched control. DCD is evaluated on pairs matched by their WildGuard scores.

				DCD AUROC	
Target	Family	pairs	Guard (matching check)	**deployed bank**	no own axis
Mistral-7B	SafeMTData	133	0.501	**0.965**	0.917
	SafeDialBench	78	0.498	**0.898**	0.886
	cipher	110	0.495	**1.000**	0.417
Llama-3.1-8B	SafeMTData	133	0.501	**0.965**	0.931
	SafeDialBench	78	0.498	**0.947**	0.941
	cipher	110	0.495	**1.000**	0.195

Table 16: Reverse matched control for the own-axis-withheld DCD bank. WildGuard is evaluated on pairs matched by their DCD scores.

Target	Family	pairs	DCD (matching check)	**Guard on pairs**	Guard, full sample
Mistral-7B	SafeMTData	58	0.504	**0.695**	0.711
	SafeDialBench	88	0.501	**0.900**	0.896
	cipher	134	0.500	**0.761**	0.770
Llama-3.1-8B	SafeMTData	47	0.502	**0.690**	0.711
	SafeDialBench	43	0.506	**0.880**	0.896
	cipher	75	0.580	**0.816**	0.770

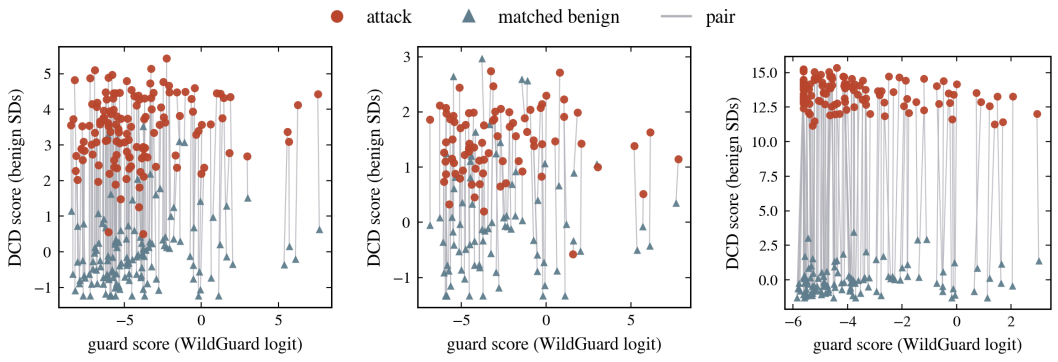
What the reverse control shows. On pairs matched by the own-axis-withheld DCD bank, WildGuard retains 0.690–0.900 AUROC, close to its 0.711–0.896 range on the full samples. The DCD matching check is 0.500–0.506 in five cells; the Llama cipher cell remains at 0.580, making that match less complete. For SafeMTData and SafeDialBench, the two controls show that neither readout contains all the evidence used by the other. DCD reaches 0.886–0.941 after matching on WildGuard, while WildGuard reaches 0.690–0.900 after matching on DCD. Cipher instead exposes the bank-coverage limitation because the reverse control withholds its only aligned axis. The 43–134 surviving pairs form the narrow DCD score-overlap region rather than a representative evaluation sample.

J STRESS TESTS OF THE DIRECTIONAL READOUT

This appendix asks whether DCD's advantage on ordinary chat survives stronger controls. We verify the guard implementation, intervene on conversation history, evaluate legitimate sensitive traffic and compare DCD with LAD against structure-matched twins.

Validating the guard baseline. Llama-Guard-3 emits a newline before its verdict, so we compare `safe` and `unsafe` at the verdict token rather than at the final prompt position, where neither token is in play. Read correctly, Llama-Guard-3 shows the largest paired gap of the three guards between a direct request and the same goal distributed over turns, $+9.606$ logits against WildGuard's $+6.914$. We also test whether scoring each user turn without its preceding conversation disadvantages the guard. With the same turn budget and each condition calibrated separately to 5% FPR, supplying the full conversation prefix changes mean AUROC by $+0.025$ and -0.016 and TPR by $+0.008$ and $+0.022$ on Mistral-7B and Llama-3.1-8B. Family-level AUROC changes range from -0.155 to $+0.073$, so including history brings no consistent benefit.

Removing the history. We test whether the DCD signal depends on accumulated history by holding each final request fixed. SafeMTData and SafeDialBench attacks selected as successful in an earlier screen are replayed with their original history, a turn-for-turn replacement from a real WildChat-1M conversation, their own history shuffled, or no history. The sets contain 52 SafeMTData and 78 SafeDialBench conversations on Llama-3.1-8B, and 16 and 32 on Mistral-7B. We regenerate responses in every condition and apply the same Granite response-mode judge. Thus even



(a) SafeMTData (b) SafeDialBench (c) cipher

Figure 6: Guard-matched pairs on Mistral-7B with the deployed bank. Attacks sit 3.4, 1.4 and 13.5 benign standard deviations above their matches from left to right; vertical scales differ.

Table 17: History interventions on originally successful attacks.

	Harmful compliance				DCD fires			
	Llama-3.1-8B		Mistral-7B		Llama-3.1-8B		Mistral-7B	
Condition	SafeMT	SafeDial	SafeMT	SafeDial	SafeMT	SafeDial	SafeMT	SafeDial
Unchanged history	0.50	0.53	0.63	0.81	1.00	0.95	1.00	0.88
History replaced	0.25	0.24	0.19	0.47	0.15	0.15	0.06	0.13
History shuffled	0.33	0.38	0.69	0.66	0.98	0.94	1.00	0.84
Request alone	0.27	0.14	0.38	0.41	0.00	0.00	0.00	0.00

the unchanged-history replay has compliance of 0.50–0.81, rather than one. We score the single-axis rescue monitor (§5.4) at the WildChat 5% limit, not the deployed bank.

Replacing the history lowers compliance in every cell and collapses DCD firing, whereas shuffling the same turns leaves firing within 0.04 of the unchanged condition despite moving compliance in both directions (Table 17). The signal therefore tracks what the history contains rather than the exact order of its turns. It is not simply reading compliance; within each condition the score separates complying from refusing conversations at only 0.22–0.53 AUROC. With the request alone DCD never fires, since $c_1 = 0$ by construction. The isolated read of that unchanged request is identical to twelve decimals across all four conditions on both targets, confirming that the intervention changes the history rather than the request. A contextual-read monitor fitted and calibrated by the same procedure also responds to the replacement, with firing falling from 0.91–1.00 to 0.09–0.27. It fires on 0.13–0.31 of the requests presented alone.

False alarms on legitimate sensitive traffic. A monitor that flags medical, legal or mental-health conversations is not deployable, and the drift axis has an obvious way to fail here, since a careful clinical conversation approaches difficult material gradually. We test it on the real sensitive-topic conversations and topic controls of Table 18, drawn from WildChat-1M as Appendix C describes. Assistant turns are regenerated by Mistral-7B and Llama-3.1-8B through the same pipeline as every other corpus, so the reads are the paper's reads.

At the paper's operating point, $\alpha = 0.05$ calibrated on ordinary chat, DCD flags 0.03–0.12 of real sensitive conversations on Mistral-7B, against 0.07–0.15 for the guards, and 0.17–0.20 on Llama-3.1-8B. The topic controls separate the two explanations. They are flagged at 0.065–0.095 on Mistral-7B and 0.135–0.160 on Llama-3.1-8B, against 0.050–0.070 by the guards, so on Mistral-7B the topic adds nothing the guards do not also pay and on Llama-3.1-8B it adds a few points on top of a draw that is already harder than ordinary chat. The detection rates in this block, for DCD and the guards alike, are those of Table 2, since these are the same detectors at the same operating point.

Table 18: Detection and false alarms on real sensitive traffic at two operating points. DCD ranges span the five family banks; bold marks the best detection rate in each block.

	TPR by family					FPR		
detector	SafeMT	SafeDial	cipher	Cresc.	Echo	sensitive	controls	ordinary chat
the paper's limit, 5% *on ordinary chat*								
DCD, Mistral-7B	**1.000**	0.526	**1.000**	**0.694**	**0.918**	0.030–0.120	0.065–0.095	0.040–0.065
DCD, Llama-3.1-8B	0.970	**0.837**	**1.000**	0.472	0.836	0.170–0.195	0.135–0.160	0.035–0.050
Llama-Guard-3	0.032	0.120	0.008	0.225	0.111	0.150	0.060	0.060
WildGuard	0.242	0.522	0.122	0.516	0.360	0.145	0.070	0.040
Granite	0.163	0.352	0.141	0.443	0.297	0.070	0.050	0.060
recalibrated on the mixed benign set								
DCD, Mistral-7B	**1.000**	**0.493**	**1.000**	**0.684**	**0.918**	0.050–0.085	0.050–0.110	0.055–0.085
DCD, Llama-3.1-8B	0.809	0.474	**1.000**	0.347	0.705	0.050–0.055	0.045–0.075	0.000–0.030
Llama-Guard-3	0.023	0.091	0.008	0.168	0.085	0.085	0.035	0.020
WildGuard	0.104	0.263	0.008	0.363	0.213	0.075	0.020	0.010
Granite	0.139	0.272	0.093	0.415	0.249	0.060	0.045	0.030

Half the sensitive conversations can instead be folded into the calibration set, the chart moments and the conformal limit taken there together, exactly as the paper calibrates on chat. Every detector then reads 0.050–0.085 on the 200 sensitive conversations held out from that fit, and within a few points of α on real and held-out chat. Detection is what pays for it, and the bill differs by target. Mistral-7B is almost unaffected, changing by at most 0.033 across the five families. Llama-3.1-8B loses real ground, cipher holding at 1.000 while the other four families give up 0.125–0.363. The lead over the best guard survives in nine of the ten cells, the exception being Llama-3.1-8B on Crescendo at 0.347 against Granite's 0.415, one of the four cells in which DCD already trails the best guard in Table 2, before any recalibration.

Another threshold cannot separate the two. Sensitive and attack score distributions overlap on the same scalar, so raising the limit reduces sensitive-traffic false alarms at some cost in attack detection. Separating legitimate sensitive development from adversarial drift needs information beyond the current score, such as an additional direction, rather than another threshold on the same axis. We leave that to future work.

Structure-matched twins are harder benign traffic still, resembling attacks in everything but goal. Across the five targets the rescue monitor of §5.4 fires on 0.04–0.28 of Crescendo twins and 0.38–0.76 of Echo Chamber twins, each twin scored by an axis fitted on the other twins. No twin enters the limit, which is calibrated on five- to six-turn chat and applied unchanged to every scored conversation, including the ten-turn Echo Chamber twins and cipher attacks of up to 13 turns.

Contrast with direction-agnostic trajectory detection. Those twins also separate DCD from the closest concurrent method. LAD also reads activation trajectories, but through magnitude features rather than a learned direction (Table 7); we reimplement it from the paper's description and score both on the same cached activations. The decisive test is against the *structure-matched* benign twins, which remove length and structure as confounds, since long eventful benign chats and long attacks both produce large activation movement.

Under the control LAD's magnitude features fall toward chance on all five targets (0.478–0.859 AUROC, below chance in two cells) while DCD holds at 0.832–0.925, and DCD leads in all ten cells (Table 19).

The common horizon does not create DCD's lead. Scoring Echo Chamber and its twin over their natural ten turns moves LAD from 0.767 to 0.718 and DCD from 0.896 to 0.879, widening the mean gap from 0.129 to 0.161. Crescendo already has its natural length of five turns. On *unmatched* benign traffic LAD does well, reaching 0.947 on Crescendo, because attack and benign conversations differ in length and activity there, which is exactly why the matched control is the meaningful test, and the same lesson as Appendix I one level up. Methods that measure how much a conversation moved look strong until the comparison is fair. We frame this as a mechanism contrast with concurrent work, not a claim of superiority over a tuned system.

Table 19: LAD against DCD on the structure-matched twins at the common five-turn horizon, AUROC. Bold marks the better of each pair.

	Crescendo		Echo Chamber	
Target	LAD	DCD	LAD	DCD
Mistral-7B	0.714	**0.851**	0.859	**0.910**
Llama-3.1-8B	0.557	**0.893**	0.618	**0.866**
Qwen-2.5-7B	0.478	**0.832**	0.806	**0.910**
Phi-3-mini	0.490	**0.898**	0.805	**0.922**
Gemma-2-9B	0.621	**0.925**	0.746	**0.870**
Mean	0.572	**0.880**	0.767	**0.896**

Representation-level defenses. Defenses that act on representations rather than text, such as circuit breakers (Zou et al., 2024), are trained on single-turn harmful completions; Bullwinkel et al. (2025) attribute their weakness against multi-turn attacks to exactly the generalization gap we measure geometrically in §4.

K ADAPTIVE ATTACKERS

We report additional details about the adaptive attacker, the experimental methodology and additional results.

Threat model. In every setting the attacker can query the target model, observe its replies and rewind the conversation. Its main capability is to reword user turns, while it cannot change the goal of the conversation, the number of turns or the weights of the target model. The monitor under attack is DCD with three axes, one each for SafeMTData, SafeDialBench and the cipher corpus, each read at the layer selected for it on the target model. Crescendo and Echo Chamber are deliberately excluded from the monitor's bank, allowing an attacker without access to the true axes to fit surrogate directions on these two families without recovering any axis used by the monitor.

We vary how much the attacker knows about the monitor. In the strongest (white-box) setting it also holds an exact copy of the monitor, i.e., the bank of axes $\{w^{(k)}\}$, the layer of each axis, the EWMA grid, the benign standardization constants and the conformal threshold τ, all read on the target model. In the intermediate settings, it knows two of the three components, i.e., the monitor can be wrong in the axes, in the layers or in the model it reads activations from, or it knows none of them and its copy is wrong in all three. With the wrong axes, the attacker uses axes fitted on Crescendo and Echo Chamber. With the wrong layer, it fits and reads the three axes at a layer the monitor does not use. With the wrong model, it fits the three axes on Phi-3-mini at the layer selected for that model and reads activations from it. Finally, with all three wrong, it uses Crescendo and Echo Chamber axes fitted on Phi-3-mini. Every copy is calibrated like the monitor on the same benign traffic.

The attacker keeps the first user turn unchanged, rewrites the following turns in sequence and lets the target generate a new reply to each. This setup is well defined only when the user turns follow a fixed script that does not depend on the target's previous replies, and when rewording can preserve the underlying request. We therefore run the adaptive attack on SafeMTData and SafeDialBench. Each conversation is scored under the condition that the bank holds an axis for the attacked family, fitted on a disjoint half of that family's goals. We do not use other datasets used in Section 5. Crescendo and Echo Chamber turns were each written in response to the generating partner's previous reply, so rewording one turn would leave later turns answering replies that no longer occur, while in the cipher corpus, rewording would break the substitution scheme that carries the request. We leave adaptive attacks tailored to these settings, rather than based on rewriting fixed user turns, to future work.

How the attack works. The attacker leaves the first user turn unchanged and rewrites every later turn in sequence. At turn t, it considers the original message together with B candidate rewrites. These rewrites are sampled from the target model at temperature 1 with nucleus sampling, using the

conversation exactly as it currently stands, including the attacker's previously accepted rewrites, the replies they produced, an instruction to reword the next user message without answering it and three worked examples. The examples are important because, in preliminary experiments, instruction-tuned models often continued the dialogue instead of rephrasing the user turn. Candidates that do not read as user messages, or that duplicate an already accepted rewrite, are discarded. Sampling proceeds in batches of 16 until B distinct rewrites are collected or a maximum of 640 proposals have been drawn.

Each candidate is scored by what the conversation score of the attacker's bank would be if it were sent, with the committed history included. On each intermediate turn, the attacker tries the candidates from the lowest score upward, reads the reply, and keeps the first one the target does not refuse. On the final turn, the attacker generates a reply for every candidate, keeps only those still judged harmful by the same judge (Padhi et al., 2025) used to determine jailbreak success and selects the harmful candidate with the lowest score. If none remains harmful, it falls back to the lowest-scoring candidate overall.

To keep all evaluation arms comparable, a conversation is removed from every arm, including the baseline, if its rewriter fails to produce enough candidates at any turn in any arm. This leaves a common evaluation set of 76 of the 80 Llama conversations and 88 of the 94 Mistral conversations that jailbreak their targets without rewriting.

Why the white-box attacker still gets caught. DCD continues to detect attacks operated by the white-box adaptive attacker for two main reasons.

First, the attacker must reduce the detection score while preserving jailbreak success and these objectives conflict. On the final turn, it generates a reply for each candidate and discards candidates whose replies the judge no longer considers harmful. At B=32, this rules out the lowest-scoring rewrite in 36 of 76 Llama conversations and 21 of 88 Mistral conversations. Second, the attacker must reduce the score enough to cross the detection threshold and evasion depends largely on the starting score. Conversations that evade detection after rewriting initially scored 1.14 and 0.93 times the threshold, compared with 1.42 and 1.46 for those still detected. Among conversations DCD detects before rewriting, the white-box attacker achieves joint success in 13 of 69 on Llama and 8 of 62 on Mistral. Its remaining joint successes come from conversations that already evaded detection before rewriting.

Why missing one component of the monitor makes the attacker weaker. The three incomplete attackers face the same limitation: each minimises a score different from the one the monitor uses, so its optimisation helps only partially. However, they differ in how much this limitation affects them.

For the attacker that does not know the correct axes, recall that the monitor takes the maximum over all axes in the bank. Reducing this maximum requires lowering the score along the highest-scoring axis, that here is the axis fitted on the attack family being evaluated. Then, an attacker using the wrong axes therefore optimises the wrong directions: in our experimental setting, it selects rewrites that reduce drift along the Crescendo and Echo Chamber directions, that differ from the axes used by the monitor (§5.3). Consequently, these rewrites may leave the score that determines the alarm unchanged. The other two attackers keep instead the relevant attack axes but measure drift at a different depth or in a different model. With the wrong depth, the attacker still seeks to reduce drift for the same conversation in the same model, but the fitted direction changes substantially with depth (Appendix F). Even so, rewrites that reduce drift at one depth tend to reduce it at neighbouring depths, consistent with the measurements we did, coming from the same model. With the wrong target model, the surrogate evaluates the same candidate turns using axes fitted in another representation space. The two models have their own widths and bases, so they share no direction vectors, yet rewrites that reduce conversational drift in one model tend to do so in the other.

To conclude, we note that Table 4 reflects this ordering. The attacker that knows the layer and model but uses the wrong axes is the weakest, while those using the relevant axes at the wrong depth or in the wrong model have an advantage. Only the white-box attacker directly minimises the deployed statistic, so any reduction it achieves is also a reduction in the monitor's score.

Table 20: Detection and joint success of every adaptive attacker at 8, 16, 24 and 32 rewrites per turn, scored by the target monitor.

	Llama-3.1-8B				Mistral-7B			
rewrites per turn B	8	16	24	32	8	16	24	32
conversations n	79	79	78	76	92	92	91	88
detection at the deployed limit								
original attacks	0.911	0.911	0.910	0.908	0.696	0.696	0.692	0.705
none of the three	0.924	0.899	0.885	0.868	0.793	0.793	0.780	0.795
layer and model	0.911	0.861	0.859	0.842	0.739	0.772	0.758	0.784
axis and model	0.899	0.823	0.782	0.763	0.707	0.707	0.703	0.716
axis and layer	0.911	0.835	0.833	0.789	0.717	0.685	0.725	0.705
all three (white box)	0.873	0.810	0.744	0.711	0.620	0.598	0.582	0.602
joint success								
original attacks	0.089	0.089	0.090	0.092	0.304	0.304	0.308	0.295
none of the three	0.063	0.089	0.103	0.118	0.207	0.196	0.209	0.193
layer and model	0.076	0.127	0.128	0.132	0.261	0.228	0.220	0.193
axis and model	0.076	0.152	0.179	0.197	0.293	0.293	0.286	0.273
axis and layer	0.089	0.139	0.141	0.184	0.250	0.293	0.253	0.284
all three (white box)	0.089	0.165	0.218	0.250	0.359	0.380	0.385	0.375

The query budget. Table 20 repeats every attacker, with different knowledge level, with B=8, 16, 24 and 32 rewrites per turn. For each budget, we evaluate only conversations for which every arm produced enough distinct rewrites and we rescore the original attacks on that same subset. On Llama, the white-box attacker lowers detection as the budget grows, from 0.873 at B=8 to 0.711 at B=32, against 0.911 and 0.908 for the original attacks, and its joint success rises accordingly, from 0.089 to 0.250. On Mistral both quantities are flat from B=16 onward, with detection at 0.58–0.60 and joint success at 0.375–0.385. Attackers without an exact copy of the monitor reduce detection less, reaching 0.763–0.868 on Llama at B=32, and never exceed the white-box attacker's joint success at the same budget. On Mistral, they leave detection close to or above the original attacks (0.685–0.795 versus 0.692–0.705), and none surpasses the original attacks in joint success at any budget.

A strong guard on the same conversations. We also evaluated the conversations generated by the white-box attacker at B=32 using Granite Guardian 3.0-8B together with the original conversations they were rewritten from. We score each user turn and take the maximum score over the conversation. This serves as a control on attack quality rather than as an additional defense, since only DCD was directly targeted by the attacker. On Llama, the guard flags 0.658 of the original conversations and 0.645 of the rewritten ones, compared with 0.908 and 0.711 for DCD. On Mistral, the corresponding rates are 0.659 and 0.614 for the guard, versus 0.705 and 0.602 for DCD. Before rewriting, the guard performs worse than DCD on both targets; the gap narrows after rewriting mainly because DCD is the detector being attacked. Importantly, rewriting does not increase the guard's detection rate. This suggests that the attacker did not evade DCD simply by making the conversations more overtly harmful. At the same time, because the guard itself was not targeted, its relative stability should not be interpreted as evidence of robustness, but rather as a consequence of being outside the attack objective.

L DELAYED ONSET

A monitor that fires on any long or eventful conversation would be of no practical use. We prepend $N = 0$, 6 or 12 benign turns to ten attack conversations from each of two families on four targets, scoring each attack with a direction fitted on the other nine $N = 0$ conversations of its family. SafeMTData holds at 1.00 in all twelve architecture–onset cells. SafeDialBench varies by target rather than declining with onset, spanning 0.50–1.00, with the largest transient dip on Llama-3.1-8B ($0.80 \to 0.50 \to 0.70$). From $N = 0$ to $N = 12$ detection falls by 0.10 on one target, is unchanged on another and rises by 0.10–0.20 on two. Across the tested prefixes, additional benign history does not systematically reduce detection, and held-out false alarms remain at or below the nominal rate.

The score reads the direction of the context shift, and ordinary conversation contributes little along that direction for prefixes up to twelve turns. SafeDialBench is already at 0.60–0.90 at $N = 0$ on three of four architectures, so its lower numbers are that family's separation being weaker in general rather than an onset effect.

Prefixes come from the WildChat calibration split and the false-alarm controls from the held-out split, so no conversation spliced in front of an attack appears in the set the limit was fitted on. Detection uses a limit fitted on all 290 long benign conversations per target. For an out-of-sample false-alarm check the limit is instead fitted on 145 and evaluated on the remaining 145, giving 0.007–0.048 for the SafeMTData axes and 0.021–0.041 for SafeDialBench against a nominal $\alpha = 0.05$. These rates count one to seven alarms, and the calibration half is itself only 145 conversations, below the roughly 200 needed to estimate a 5% tail, so they are a limited check rather than a precise one.