IIMB-WP NO. 742/2026

WORKING PAPER NO: 742

PsychoPass: Geometric Profiling of Multi-Turn Adversarial LLM Conversations

Muberra Ozmen
Coveo
mozmen@coveo.com

Subhabrata Majumdar
Assistant Professor, Decision Sciences
Indian Institute of Management Bangalore
smajumdar@iimb.ac.in

Year of Publication – June 2026

PSYCHOPASS: Geometric Profiling of Multi-Turn Adversarial LLM Conversations

Muberra Ozmen
Coveo
Montreal, QC, Canada
mozmen@coveo.com

Subhabrata Majumdar
Indian Institute of Management Bangalore
Bengaluru, India
smajumdar@iimb.ac.in

Abstract

Multi-turn jailbreak attacks on large language models (LLMs) reveal a mismatch in current guardrails: they operate on individual turns, while attacks unfold as trajectories across conversations. We propose a shift from content to dynamics, modeling conversations as paths in representation space and asking whether adversarial intent is encoded early in their geometry. We introduce PSYCHOPASS, a framework that extracts geometric features from conversation trajectories in embedding space to predict a potential attack *before* harmful content is produced. These features achieve near-perfect performance in naïve classifiers, which is largely explained by the inclusion of number of turns as a feature. After removing this confound, a smaller but consistent geometric signal remains, with classification performance that does not depend meaningfully on encoder choice. Crucially, this signal appears early in the conversation: attack outcomes remain above chance from short prefixes alone, more reliably than baseline guardrails. A supporting theoretical analysis explains these findings via a decomposition of length and shape, a detection bound based on prefix length, and encoder invariance. Together, these results show that adversarial conversations leave an early, representation-robust geometric fingerprint suitable for online monitoring [1].

1 Introduction

Large Language Models (LLMs) deployed in conversational interfaces are routinely targeted by multi-turn jailbreaks that escalate, backtrack, and reframe a forbidden request across several turns until the target system complies [14, 19]. Most runtime defenses aimed at preventing such attacks operate turn-by-turn: each user message is scored independently for policy violation, and each model response is filtered after generation. This design is fundamentally reactive, in that it can flag harmful *content* once it appears but cannot flag the *process* by which an attacker steers a conversation toward harm before that content is produced.

This paper asks whether the geometry of the trajectory of a conversation in a representational space can serve as an early warning signal for adversarial intent. Our premise is that if a multi-turn attack must systematically move a conversation toward a forbidden region, that directed movement should leave a geometric fingerprint, such as through curvature, drift, circularity, or temporal asymmetry of the conversation trajectory that distinguishes it from benign interaction before any single turn looks dangerous on its own (Figure 1).

To test our proposition, we embed multi-turn adversarial conversations via a sparse lexical encoder and a dense semantic encoder, extract trajectory-level shape statistics and temporal features, and run a sequence of three experiments to progressively isolate the geometric signal from confounds. The first experiment shows that naïve trajectory classification achieves near-perfect AUROC, but this is almost

[1] Link to repository: https://github.com/muberraozmen/psycho-pass

Preprint.

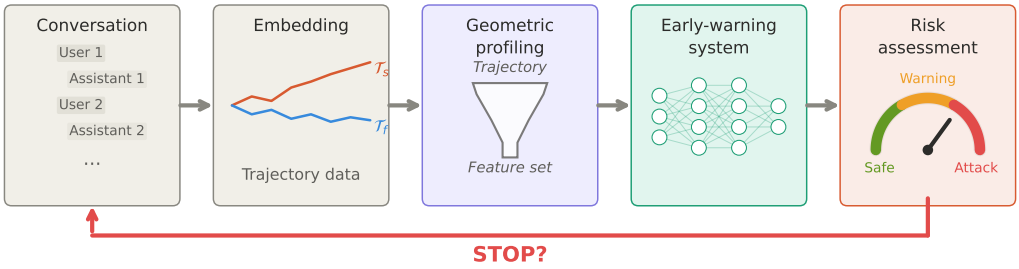

Figure 1: The PSYCHOPASS pipeline. We distill down the difference between trajectories of successful ($\mathcal{T}_s$) and failed ($\mathcal{T}_f$) attacks in the embedding space using geometric features, which are used to train an early-warning system to provide online risk assessment to detect attacks in progress.

entirely a length artifact: failed attacks exhaust their turn budget while successful ones terminate early, and any feature that scales with conversation length inherits this free signal. The second experiment removes the length confound by equalizing all trajectories to six turns. AUROC drops but remains reliably above chance, carried by stretch-decreasing dynamics, outlier timing, and low-frequency fluctuation; the choice of encoder has no statistically meaningful effect on classification. The third experiment asks whether this residual signal is available early enough to matter, and finds that it is: attack outcome stays above chance from prefixes as short as two turns, while a content-based safety classifier applied to the same prefixes collapses to chance. We complement the empirical study with a theoretical analysis. A decomposition of discriminability into length and shape components quantitatively predicts the performance drop between the first and second experiments, a finite-sample detection bound explains why short prefixes suffice in the third experiment, and a rank-preservation result explains the encoder invariance in the second experiment.

Related work How LLMs behave in the wild over long conversations spanning multiple turns is only beginning to be understood. Recently, Gooding and Grefenstette [3] showed that geometric features of conversation trajectories can serve as reward signals for alignment. Other studies have found that conversational history geometrically constrains future responses [20], model performance degrades as conversations grow longer [9], and non-reasoning LLMs show improved performance when an input prompt is repeated inside the same conversation [10]. These findings suggest that there are structural nuances in multi-turn interactions that single-turn analysis misses entirely.

Research on adversarial attacks on LLMs has followed a similar single-turn-to-multi-turn trajectory. Early jailbreak and prompt injection attacks relied on pre-crafted, static prefixes or suffixes [11, 23]. As guardrails and safety training became adept at stopping such attacks, researchers discovered that new strategies like Crescendo [19] and TAP [14] can spread and escalate an attack across turns to effectively circumvent perimeter defenses and make an LLM achieve a harmful objective. Several open-source libraries implement single- and multi-turn attacks. Two such well-known tools are garak [2] and PyRIT [17], which standardize multi-turn attack pipelines using harmful objectives from safety benchmarks like AdvBench and HarmBench [13].

Runtime defense methods to prevent attacks focus on filtering undesired inputs and outputs. Past research directions on such methods include encoder-only classifiers [15, 21], moderation LLMs [8, 16], embedding-based classifiers [1], and inference-time methods [6, 22]. While tools like Nova rules [18] can combine them to operationalize a guardrail that achieves acceptable latency-performance tradeoff in deployed AI systems, these methods often generalize poorly to unseen data. For example, Hackett et al. [5] showed that the best open-source prompt injection classifiers remain vulnerable to evasion techniques that perturb a malicious input prompt to obfuscate its content.

Existing runtime defenses work by filtering harmful content, not by discouraging malicious intent. This leaves the LLM underneath vulnerable to an adaptable attacker—often another LLM—to try alternative ways of achieving the same harmful goal if an initial attack is foiled. PSYCHOPASS aims to mitigate this risk, by using information inherent in the shape of an adversarial trajectory to enable real-time intervention of attacks unfolding over multiple turns.

2 Methodology

The PSYCHOPASS pipeline follows three stages: *Attack Generation*, which produces adversarial multi-turn conversations against target LLMs; *Trajectory Construction*, which embeds each conversation as a path followed in d-dimension space; and *Geometric Profiling*, which extracts statistical features from those paths.

Attack generation We use PyRIT [17] to orchestrate Crescendo [19] attacks, which utilizes a multi-turn escalation strategy with dynamic backtracking, against four target LLMs (see Appendix A for details).

Each attack is initialized with a seed objective drawn from one of the benchmarks bundled with PyRIT: AdvBench [23], HarmBench [13], and AIRT [17]. In total we draw seed from these three families, sampling two attacks per objective. Each attack is limited to a maximum of $T_{\max} = 8$ turns and 2 backtracks, where one turn is defined as one user message followed by one assistant message.

Three LLMs are involved in the attack generation process. The *adversarial bot* (`llama-3.1-8b-instruct`, temperature $t = 1.0$) acts as a proxy user that iteratively refines prompts to satisfy the seed objective. The *objective bot* is the system under evaluation; we experiment with four target LLMs (`llama-3.1-8b-instruct`, `gpt-oss-120b`, `mistral-small-3.2-24b-instruct`, `qwen-2.5-7b-instruct`), all at $t = 1.0$. The *scoring bot* (`gpt-oss-120b`, $t = 0.1$) is an automated judge that declares an attack a success when its score exceeds a threshold of 0.8. All bots are configured with a top-p value of 1.0 and are restricted to $1{,}024$ completion tokens per turn.

Trajectory construction A conversation $C = \{m_i\}_{i=1}^{N}$ of $N = 2 \times T_{\text{exec}}$ messages (steps), where T_{exec} denotes number of executed turns (with alternating user and assistant messages), is projected into a vector space via an encoder $\phi : \mathcal{M} \to \mathbb{R}^d$, yielding the *conversation trajectory* $\mathcal{T} = \{v_i\}_{i=1}^{N}$ with $v_i = \phi(m_i)$. We use two encoders:

- *Lexical* (ϕ_L): TF-IDF with up to $d = 4{,}096$ features, English stop-word removal, and a max-document-frequency of 0.95.
- *Semantic* (ϕ_S): the dense embedding model `qwen3-embedding-8b` with a context window of $32{,}000$ tokens which yields a dimensionality of $d = 4{,}096$.

To isolate role-specific signals and to support the predictive-horizon analysis, we additionally consider the *user trajectory* $\mathcal{T}_U$ (messages of adversarial bot), the *assistant trajectory* $\mathcal{T}_A$ (messages of objective bot), and the *trimmed trajectory* obtained either by keeping the first k turns $\mathcal{T}^{(:k)} = \mathcal{T}_{1:2k}$, or by removing the last k turns $\mathcal{T}^{(-k)} = \mathcal{T}_{1:(N-2k)}$.

Geometric profiling Seven L_2-norm based statistics (Table 2 top half) are used to summarize the geometry of trajectory $\mathcal{T}$. *Path length* (L) measures total conversational effort; *Displacement* (D) the net drift; *directness* (η) whether progression was goal-oriented or meandering; the *Step pace* pair ($(\bar{s}, \sigma_s)$) the average speed and its variability; *Velocity* (V) the directional progress per step; *Circularity* (ζ) whether the trajectory loops back through previously visited regions of the embedding space.

Catch22 [12] is a well-known method for extracting temporal features from time-series data. We adapt this method to capture higher-order temporal structure in our attack trajectories. We apply a targeted subset of the Catch22 feature set to the trajectory $\mathcal{T}$ (Table 2 bottom half). Since the embedding dimension d is large, computation is restricted to the top-K=100 dimensions with highest variance. Each feature is computed independently per dimension and then aggregated across dimensions by mean, maximum, minimum and standard deviation. *Time-reversal asymmetry* (TRA) is sensitive to the temporal arrow of a series; a nonzero value indicates statistical irreversibility. *Outlier timing* (OT^+, OT^-) measures the median normalized index of above- and below-mean outliers across an increasing sequence of thresholds; values near zero or one mean outliers concentrate early or late. *Stretch-high* ($\text{Str}^{\uparrow}$) is the longest run of consecutive above-mean values and captures the largest sustained excursion into a particular region of the embedding space. *Stretch-decreasing* ($\text{Str}^{\downarrow}$) is the longest run of non-positive first differences, i.e., the most sustained monotonic withdrawal in a latent dimension. *High fluctuation* (HF) is the trace of the covariance of a $3{\times}3$ transition matrix on the autocorrelation-down-sampled series, with larger values indicating more persistent low-frequency fluctuation in the bin sequence.

Table 1: Geometric features extracted from a trajectory $\mathcal{T} = \{v_i\}_{i=1}^{N} \subset \mathbb{R}^d$. $\Delta d_i = \|v_{i+1} - v_i\|_2$ denotes the i-th step length, $\mu = N^{-1}\sum_i v_i$ the centroid, $R_i = \|v_i - \mu\|_2$ the radius from μ, and $\epsilon > 0$ a small constant for numerical stability. $x = (x_1, \ldots, x_N)$ denotes the z-normalized values of a single coordinate of $\mathcal{T}$, $\bar{x}$ its mean, $\Delta x_t = x_{t+1} - x_t$ the first differences, and $\rho_x(k)$ its lag-k autocorrelation. $\Theta = \{0.01, 0.02, \ldots, \max_t |x_t|\}$ enumerates positive standardized thresholds; $\tau = \min\{k \geq 1 : \rho_x(k) \leq 0\}$ is the first zero-crossing of the autocorrelation; $x_{\downarrow\tau}$ is the resulting downsampled series; and $T \in [0,1]^{3\times 3}$ collects the empirical transition probabilities $T_{ij} = \Pr(s_{t+1} = j \mid s_t = i)$ over the equiprobable three-letter symbolization $s_t \in \{1,2,3\}$ of $x_{\downarrow\tau}$.

Type	**Feature**	**Definition**	**Interpretation**
L_2-norm shape statistics	L	$\sum_{i=1}^{N-1} \Delta d_i$	Total conversational effort
	D	$\|v_N - v_1\|_2$	Net semantic drift
	η	$D/(L+\epsilon)$	Goal-directedness $\in (0,1]$
	$\bar{s}$	$L/(N-1)$	Average speed
	σ_s	$\mathrm{std}(\Delta d)$	Average variability
	V	D/N	Directional progress per step
	ζ	$\mathrm{std}(R)/(\bar{R}+\epsilon)$	Tendency to revisit regions
Multivariate temporal (Catch22)	TRA	$\frac{1}{N-1}\sum_t (\Delta x_t)^3$	Statistical irreversibility
	OT^+, OT^-	Median normalized index of $\pm$ outliers	Outlier concentration early/late
	$\mathrm{Str}^{\uparrow}$	Longest run of consecutive $x_t > \bar{x}$	Sustained excursion above mean
	$\mathrm{Str}^{\downarrow}$	Longest run of $\Delta x_t \leq 0$	Longest monotonic withdrawal
	HF	$\mathrm{tr}\,\mathrm{Cov}(T(x_{\downarrow\tau}))$	Low-freq. fluctuation persistence

When temporal features are computed on very short trajectory prefixes (e.g. k=4, which yields N=4 messages per trajectory), certain Catch22 statistics can be undefined: the autocorrelation sequence of a four-point series may have no zero-crossing, leaving the lag τ undefined and making the downsampled series degenerate; likewise, the empirical transition matrix may be sparsely populated, causing missing feature values. Similarly, the L_2-norm circularity statistic ζ is explicitly missing for trajectories shorter than three steps. These missing values are handled differently by the two classifiers. For logistic regression, each feature is imputed with its per-feature median computed. For gradient boosting, we use Histogram-based Gradient Boosting Classification Tree, which natively accommodates missing entries by treating them as a separate branch direction at each split, requiring no imputation.

3 Experiments

Using the features above, we design and execute a sequence of experiments that progressively isolate the geometric signal in adversarial trajectories. All experiments share the same pool of generated conversations: 7,525 Crescendo attacks across the 4 target LLMs and the 3 seed benchmarks, with an overall attack-success rate of 69.6% (see Appendix B for statistics of generated attacks). Experiments differ only in trajectory configuration and whether the number of executed turns in the conversation is exposed to the classifier. To evaluate the features as impact factors on the attack outcome (success or failure), we train two simple classifiers: *logistic regression* (LR) with L1 regularization (strength 1.0) as a linear classifier, and *gradient boosting* (GB) (learning rate 0.1, depth 3 and L2 regularization 0.01) as a tree-based ensemble method. Throughout, the labeled set of conversations is split 80/20 into training and evaluation partitions. We report classification performance and *surviving factors*, defined as those whose permutation importance exceeds the standard deviation of their own noise (see Appendix C for formulations). For each conversation, geometric features are extracted separately from the full conversation trajectory $\mathcal{T}$, the user-only trajectory $\mathcal{T}_U$, and the assistant-only trajectory $\mathcal{T}_A$, and all three feature blocks are concatenated into a single feature vector presented to the classifier.

3.1 Experiment 1: Conversation length confound

We train both classifiers on the unfiltered set of 7,525 conversation using full trajectories under two conditions: with and without number of executed turns T_{exec} in the feature set. Regardless of encoder and classifier, both conditions yield near-identical and near-perfect classification, where AUROC remains at 0.991. The demonstration of important factors in Figure 3.1 makes the source of this

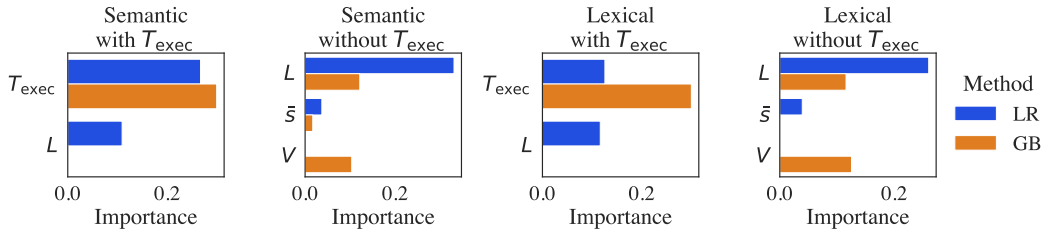

Figure 2: Permutation importance of top surviving factors in Experiment 1, per encoder and classifier. When included, T_{exec} is the dominating factor regardless of encoder and classifier, when removed, path-length proxy L inherits its role at comparable magnitude.

invariance explicit. When T_{exec} is in the feature set it is the dominant factor. When T_{exec} is removed, path length L takes over: it scales additively with the number of turns, so the dropped signal is reabsorbed almost exactly. The same pattern holds under both encoders, suggesting that the length confound is not a property of the embedding but of the geometry-to-length coupling.

Failed attacks exhaust the $T_{\max} = 8$ budget by construction, while successful attacks typically terminate in 3.3 turns on average (Appendix B). Any feature whose value scales with the number of steps therefore receives a free signal: this includes the path length L, velocity V and the average step pace $\bar{s}$. The implication is structural rather than methodological: comparing attacks across a wide range of turn counts is intrinsically unfair to genuinely shape-based features, because length-correlated proxies dominate classification regardless of the actual trajectory geometry. The remaining experiments restrict to subsets of the data in which length variation is either controlled or eliminated outright. This length sensitivity is not specific to our setup. Any experimental design that compares multi-turn conversations of variable length for any purpose risks conflating trajectory geometry with turn count. We recommend that future work in this space report results both with and without length equalization, treating the gap between the two as a diagnostic for the severity of the confound.

3.2 Experiment 2: Shape signals under fair length

Restricting to $\mathcal{T}^{(:6)}$ removes conversation length as a confound and leaves us with 2,926 length-equalized conversations across the four generators, of which 21.9% are successful escapes. We fit the logistic regression (LR) and gradient boosting (GB) classifiers, separately for the semantic (ϕ_S) and lexical (ϕ_L) encoders, repeating each run over 10 random seeds.

Table 3.2 reports the seed-averaged classification performance for the four resulting configurations. Once length information is removed, AUROC ranges from 0.653 to 0.700, and F1 from 0.400 to 0.437. The lexical encoder retains a small lead (+0.03 average AUROC and +0.03 average F1 over the semantic encoder), but the gap remains comparable to the seed-to-seed standard deviation observed within any single configuration ($\sigma_{\text{AUROC}} \in [0.017, 0.026]$, $\sigma_{\text{F1}} \in [0.022, 0.029]$). Recall stays high (0.66–0.69) while precision remains low (0.29–0.32), reflecting the class imbalance rather than any encoder-specific behavior. We therefore conclude that, with length controlled, the choice of encoder has no statistically meaningful effect on classification performance.

Table 2: Classification performance in Experiment 2, reported as mean ± standard deviation over 10 random seeds. All conversations are length-equalized ($\mathcal{T}^{(:6)}$, trimmed to the first six turns).

Encoder	Classifier	Precision	Recall	F1	AUROC
Semantic (ϕ_S)	LR	$0.295_{\pm 0.020}$	$0.675_{\pm 0.023}$	$0.411_{\pm 0.022}$	$0.672_{\pm 0.017}$
	GB	$0.287_{\pm 0.025}$	$0.659_{\pm 0.039}$	$0.400_{\pm 0.027}$	$0.653_{\pm 0.018}$
Lexical (ϕ_L)	LR	$0.323_{\pm 0.025}$	$0.682_{\pm 0.048}$	$0.437_{\pm 0.029}$	$0.700_{\pm 0.026}$
	GB	$0.315_{\pm 0.021}$	$0.691_{\pm 0.044}$	$0.432_{\pm 0.026}$	$0.680_{\pm 0.018}$

Table 3.2 summarizes which factor families survive across configurations. Cell marks reflect how reliably a family appears across seeds; a family is counted as present if it survives in any of the three trajectory types ($\mathcal{T}$, $\mathcal{T}_U$, $\mathcal{T}_A$). No family appears consistently across all four configurations; the surviving signal is clearly encoder-dependent. Under the lexical encoder, stretch-decreasing dynamics

($\mathrm{Str}^{\downarrow}$) and negative-side outlier timing (OT^{-}) are the dominant families, reaching ✓ in both LR and GB. Under the semantic encoder, the L_2-norm path descriptors (L) take the leading role, appearing in 8 and 10 of 10 seeds for LR and GB respectively. $\mathrm{Str}^{\downarrow}$ is the most broadly present family overall, reaching ✓ under the lexical encoder and (✓) under Semantic LR, but falling off under Semantic GB (4 seeds). The remaining families — HF, $\mathrm{Str}^{\uparrow}$, OT^{+}, and TRA — surface intermittently, each reaching at least (✓) in one or more configurations. Circularity (ζ) does not survive reliably in any configuration.

Table 3: Consistency of surviving factor families across encoders and classifiers in Experiment 2 (length-equalized, aggregated over 10 seeds). Cell-level marks: ✓ if family present in ≥ 7 of 10 seeds; (✓) in 5–6 seeds; ✗ otherwise.

	Lexical (ϕ_L)		**Semantic** (ϕ_S)	
Feature family	**LR**	**GB**	**LR**	**GB**
$\mathrm{Str}^{\downarrow}$	✓	✓	(✓)	✗
HF	✗	✗	(✓)	✗
L	✗	✗	✓	✓
OT^{-}	✓	(✓)	✗	(✓)
$\mathrm{Str}^{\uparrow}$	(✓)	(✓)	(✓)	✗
OT^{+}	(✓)	✗	✗	✓
TRA	✗	✓	(✓)	✗
ζ	✗	✗	✗	✗

Comparing against the experiment in Section 3.1, where the dominant signal was carried by the turn-count proxy itself, the picture under length-equalization is qualitatively different. The set of important features both *changes* and *expands*: length-correlated proxies give way to shape statistics, with stretch-decreasing dynamics ($\mathrm{Str}^{\downarrow}$) and outlier timing ($\mathrm{OT}^{-}$) leading under the lexical encoder, and path descriptors (L) persisting under the semantic encoder as a genuine shape signal once length is controlled. Encoder choice has a modest effect on performance (Lexical–Semantic AUROC gap of ~ 0.03, comparable to seed-to-seed noise) but a larger effect on which factor families emerge as survivors.

3.3 Experiment 3: Early warning potential

Experiment 3.2 established that, once length is held fixed at six turns, a stable geometric signal of around 0.65–0.70 AUROC remains. Whether this signal is operationally useful depends on *when* it becomes available: a fingerprint that requires the entire conversation to materialize is too late to act on, while one that surfaces in the opening turns supports online intervention. We test this directly by progressively trimming the tail of each conversation and re-fitting the same classifiers on the truncated prefix.

Concretely, we restart from the length-equalized pool of 2,926 conversations used in Experiment 3.2 ($\mathcal{T}^{(:6)}$), and remove the last k turns for $k \in \{0, \ldots, 4\}$, leaving prefixes of $6 - k$ turns ($\mathcal{T}^{(-k)}$). The k=0 row therefore retains the full six-turn conversation and serves as a single-seed in-figure replay of Experiment 3.2's Lexical+LR configuration. As an external reference, we additionally evaluate `llama-guard-4-12b` (Llama Guard) as a single-turn safety classifier applied to each truncated conversation; this is the dominant class of deployed runtime defense, which evaluates messages independently of trajectory shape.

Figure 3.3 summarizes classification performance as a function of the retained prefix length, viewed both as ROC (which is insensitive to the class imbalance) and as precision-recall (which is not). Two patterns stand out. First, the geometric classifier degrades *gracefully*: from the no-trim anchor (k=0, six retained turns) to last-turn trimmed (k=1); e.g., AUROC drops from 0.70 to 0.67 and PR-AUC from 0.38 to 0.35. Second, while even with only the first two turns visible (k=4) the geometric classifier remains well above chance with AUROC =0.61, `llama-guard-4-12b`'s performance drops to chance at $k = 2$ with AUROC =0.49.

The factor structure underlying these numbers is consistent with Experiment 3.2. Across all horizons, the surviving features are drawn from the same families that dominated the lexical-encoder results in Experiment 3.2: stretch-decreasing dynamics ($\mathrm{Str}^{\downarrow}$), negative-side outlier timing (OT^{-}), and

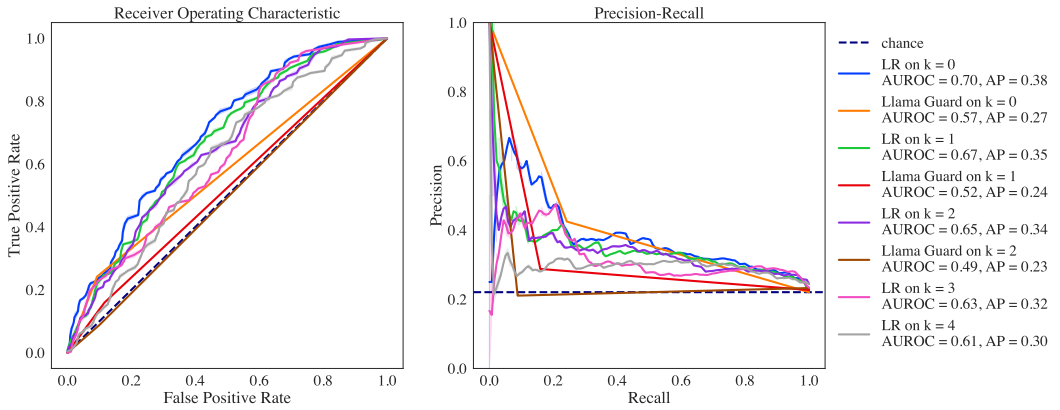

Figure 3: Receiver-operating-characteristic (left) and precision-recall (right) curves of the geometric LR classifier (lexical encoder, length-equalized prefixes) and the content-based `llama-guard-4-12b` (LG) baseline, across trim levels $k \in \{0, \ldots, 4\}$, where the retained prefix has length $6-k$. Dashed navy lines mark the chance reference: the diagonal in the ROC panel and the positive-class base rate (21.9%) in the PR panel. The full per-metric breakdown is given in Appendix D (Table D).

the L2-norm path descriptors (L, D, $\bar{s}$). At the longer prefixes ($k \in \{0, 1, 2\}$) the multi-variate temporal statistics, in particular $\mathrm{Str}^{\downarrow}$ and OT^{-}, lead the importance ranking, consistent with the lexical-encoder picture from Experiment 3.2; at the shortest prefixes ($k \in \{3, 4\}$), where temporal statistics become noisier on very short series, the L2-norm path length L rises to share the top of the ranking with OT^{-}, but the AUROC stays well above chance. The shape of an adversarial trajectory is therefore not an artifact of the entire conversation: by the second or third turn, enough of the geometry has accumulated to discriminate the eventual outcome at AUROC well within the seed-to-seed band of the full length-equalized analysis.

These results have a direct deployment consequence. A monitor that consults the lexical encoder, computes a small set of length-invariant geometric features over the user–assistant exchange so far, and queries a logistic regression classifier can flag in-flight jailbreaks at AUROC ≈ 0.65 from the first four turns of the conversation, while the dominant content-based defense slides to chance over the same prefix. The signal that distinguishes successful from failed multi-turn attacks is front-loaded, encoder-light, and available before the attack itself completes.

4 Theoretical analysis

The experiments in the previous section reveal three structural regularities: length dominates naïve classification (Experiment 3.1), encoder choice is irrelevant once length is held fixed (Experiment 3.2), and the residual shape signal that persists after length equalization is front-loaded in time (Experiment 3.3). In this section, we provide formal results that explain each regularity and yield testable quantitative predictions. Throughout, we model the binary label $Y \in \{0, 1\}$ (failed/successful attack) and a feature vector $\mathbf{f} \in \mathbb{R}^p$ extracted from the conversation trajectory. All proofs are in Appendix E.

Confounding effect of conversation length We first formalize the observation that length-correlated features dominate classification and predict the performance drop when length is controlled.

Definition 1. *Let $\ell(\mathbf{f})$ denote the* length component *of the feature vector, a subvector of features whose population means scale monotonically with the number of executed turns T_{exec}, and $\mathbf{g}(\mathbf{f})$ denote the* shape component*, the subvector of features whose population means are invariant to T_{exec} when computed on length-equalized trajectories. Write $\mathbf{f} = (\ell, \mathbf{g})$.*

Assumption 1. *The class-conditional distributions satisfy a Gaussian location model: for $y \in \{0, 1\}$, $\ell \mid Y{=}y \sim \mathcal{N}(\mu_\ell^{(y)}, \sigma_\ell^2)$, $\mathbf{g} \mid Y{=}y \sim \mathcal{N}(\boldsymbol{\mu}_g^{(y)}, \sigma_g^2 I_q)$, with $\ell \perp \mathbf{g} \mid Y$ (conditional independence of length and shape components given the label).*

The Gaussian location model is standard in signal detection theory [4] and is the implicit generative model behind Fisher's linear discriminant, which logistic regression approximates when class-

conditional distributions share a common covariance. Since both classifiers in our experiments produce decision boundaries that are monotone functions of a linear projection under this model, the assumption lets us derive closed-form expressions for discriminability that would otherwise require nonparametric estimation.

Proposition 1. *Under Assumption 1, the Bayes-optimal AUROC on the full feature vector* $\mathbf{f}$ *is*

$$\mathrm{AUROC}_{\mathrm{full}} \;=\; \Phi\left(\sqrt{\frac{\mathrm{SNR}_\ell^2 + \mathrm{SNR}_g^2}{2}}\right),$$

where $\mathrm{SNR}_\ell := \sigma_\ell^{-1}|\mu_\ell^{(1)} - \mu_\ell^{(0)}|, \mathrm{SNR}_g := \sigma_g^{-1}\|\boldsymbol{\mu}_g^{(1)} - \boldsymbol{\mu}_g^{(0)}\|$*, and* Φ *is the standard normal CDF. On length-equalized data (*$\mathrm{SNR}_\ell = 0$*), the AUROC reduces to* $\mathrm{AUROC}_{\mathrm{shape}} = \Phi\big(\mathrm{SNR}_g/\sqrt{2}\big)$.

From Experiment 3.1, $\mathrm{AUROC}_{\mathrm{full}} \approx 0.991$, so $\Phi^{-1}(0.991) \approx 2.37$ and $d'_{\mathrm{full}} = \sqrt{2} \cdot 2.37 \approx 3.35$. From Table 3.2, $\mathrm{AUROC}_{\mathrm{shape}} \approx 0.700$ (best length-equalized configuration), giving $d'_{\mathrm{shape}} = \sqrt{2}\,\Phi^{-1}(0.700) \approx 0.74$. The decomposition predicts $d'_{\mathrm{full}} = \sqrt{(d'_\ell)^2 + (d'_{\mathrm{shape}})^2}$, which yields $d'_\ell \approx \sqrt{3.35^2 - 0.74^2} \approx 3.27$. The length signal thus accounts for $(3.27/3.35)^2 \approx 95.1\%$ of the squared discriminability, consistent with the experimental observation that removing length information causes a dramatic AUROC drop while the residual shape signal, though modest in absolute terms, remains reliably above chance.

Minimum prefix length for reliable profiling Experiment 3.3 provides us with a number of geometric features that are informative for early detection of attacks. Given such a feature whose class-conditional distributions are separated, it is a natural question to ask: how many turns of observation are needed for reliable detection?

Assumption 2. *Let* h_k *denote the value of a scalar geometric feature computed from the first* k *turns of the trajectory. For each class* $y \in \{0, 1\}$*,* $h_k \mid Y{=}y$ *is sub-Gaussian with mean* $\mu_h^{(y)}(k)$ *and variance proxy* σ_h^2 *uniformly in* k*. The population separation satisfies* $\Delta(k) := |\mu_h^{(1)}(k) - \mu_h^{(0)}(k)| \geq \Delta_0 > 0$ *for all* $k \geq k_0$*, i.e. the feature becomes separated after a warm-up period* k_0.

Our geometric features are computed from trajectories in a bounded embedding space, so their range is mechanically bounded and sub-Gaussianity holds with a parameter that depends on this range. The separation condition formalizes the empirical observation from Experiment 3.2 that certain features (e.g. $\mathrm{Str}^{\downarrow}$, OT^{-}) maintain a nonzero gap between class means even after length equalization, and the warm-up period k_0 allows for features that require a minimum number of turns before they become well-defined.

Proposition 2. *Under Assumption 2, the Bayes error of a classifier on* h_k *satisfies*

$$P_{\mathrm{err}}(k) \;\leq\; \exp\left(-\frac{\Delta(k)^2}{8\,\sigma_h^2}\right) \quad \textit{for all } k \geq k_0.$$

Consequently, the minimum prefix length for error at most ε *is*

$$k^*(\varepsilon) \;=\; \min\Big\{k \geq k_0 : \Delta(k) \geq \sigma_h\sqrt{8\log(\varepsilon)}\Big\}.$$

If $\Delta(k) \geq \Delta_0$ *for all* $k \geq k_0$*, then* $k^*(\varepsilon) = k_0$ *whenever* $\Delta_0 \geq \sigma_h\sqrt{8\log(\varepsilon)}$.

The stability of AUROC across trim levels $k \in \{0, \ldots, 4\}$ in Experiment 3.3 (Table D) implies that the discriminative features attain their separation Δ_0 very early — effectively within one or two turns of observation. This is precisely the regime in which Proposition 2 predicts $k^*(\varepsilon) = k_0$: the feature separation is large enough relative to within-class variance that even a short observed prefix suffices. The proposition also clarifies why *different* features can have very different warm-up periods: features with larger Δ_0/σ_h ratios (such as L at short prefixes in Experiment 3.3) become reliable earlier, while features with smaller ratios require longer prefixes to overcome noise.

Representation invariance Experiment 3.2 shows that a sparse lexical encoder and a dense semantic encoder yield near-identical classification. We give a sufficient condition for this to happen.

Definition 2. *Two encoders $\phi_1, \phi_2 : \mathcal{M} \to \mathbb{R}^d$ are* rank-preserving *with respect to a feature functional $F : (\mathbb{R}^d)^n \to \mathbb{R}$ if, for all pairs of trajectories $\mathcal{T}, \mathcal{T}'$ of equal length,*

$$F(\phi_1(\mathcal{T})) \geq F(\phi_1(\mathcal{T}')) \iff F(\phi_2(\mathcal{T})) \geq F(\phi_2(\mathcal{T}')).$$

Proposition 3. *If two encoders are rank-preserving with respect to every feature functional in the set $\{F_1, \ldots, F_p\}$, then any classifier whose decisions depend only on univariate feature ranks has identical AUROC under both encoders. For linear classifiers, identical AUROC additionally requires that the within-class covariance structure of the feature vector is shared across encoders up to a global scalar.*

The near-perfect AUROC agreement between ϕ_L and ϕ_S in Table 3.2 suggests that the two encoders are approximately rank-preserving for the dominant features (path length, velocity, and the surviving Catch22 features). This is consistent with the Platonic Representation Hypothesis [7], which posits that diverse representation learners converge toward a shared statistical model of reality. If the underlying rank structure of trajectory geometries is a property of the data rather than the encoder, rank preservation follows naturally. For gradient boosting, Proposition 3 guarantees that this suffices for AUROC invariance. For logistic regression, the proposition requires the additional condition that the within-class covariance structure is shared across encoders; the near-identical performance of LR across encoders in Table 3.2 suggests this stronger condition holds approximately. The L_2-norm features (L, D, η, $\bar{s}$, V, ζ) depend on pairwise distances in the embedding space, and both encoders preserve the relative ordering of which trajectory pairs are more or less geometrically "extreme". The result also predicts that encoder invariance should *break* for features that depend on absolute scale rather than rank ordering. We observe this in Table 3.2, where the surviving Catch22 families differ substantially across encoders: $\mathrm{Str}^{\downarrow}$ and OT^{-} dominate under the lexical encoder while L leads under the semantic encoder.

5 Discussion and Conclusion

The most striking finding is how little of the naïve classification signal survives once conversation length is controlled: the AUROC drops from 0.99 to roughly 0.65–0.70, and our theoretical decomposition attributes 95.1% of the squared discriminability to length alone. This is a cautionary result for the broader conversation-analysis literature, where trajectory-level features are increasingly used as proxies for dialogue quality [3] or alignment [20]. Any study that compares conversations of unequal length risks mistaking a length artifact for a structural signal.

The residual geometric signal, while modest, has properties that make it practically interesting. Classification performance is similar across encoders, which means a deployment need not commit to an expensive embedding model. It is front-loaded, appearing within the first two to three turns, which is the window in which intervention can still prevent harm. And it is carried by interpretable features: monotonic withdrawal in latent dimensions (the target model gradually conceding), concentrated outlier timing (sharp shifts at specific turns), and persistent low-frequency fluctuation. These are not opaque classifier artifacts; they correspond to recognizable dynamics of how an attacker incrementally steers a conversation.

The baseline comparisons in Experiment 3.3 underscore a structural limitation of content-level defenses. The performance of `llama-guard-4-12b` collapses to chance when the final turns, the ones most likely to contain overt harmful content, are withheld. This is by design, because content filters are intended to detect harm, not intent. A trajectory-level monitor like ours would operate on a complementary signal and is therefore most valuable precisely where content filters are weakest, that is, early in the conversation before harmful content has been produced.

Limitations. Four limitations should be noted. *(i)* All experiments use a single attack strategy (Crescendo); whether the geometric fingerprint generalizes to other multi-turn methods (e.g. TAP, repeated-question attacks) remains open. *(ii)* The $T_{\max} = 8$ turn budget shapes the length confound; a different cap would shift both the success-rate distribution and the relative importance of length-sensitive features. *(iii)* Attack labels are assigned by a single LLM-as-judge with a fixed threshold; sensitivity to the judge or threshold is unexplored. *(iv)* Encoder invariance is established along a coarse sparse-vs.-dense axis; whether two dense encoders of different capacity agree as closely is not tested.

Future work. The immediate next step is cross-attack generalization: does a classifier trained on Crescendo trajectories transfer to TAP or repeated-question attacks, or does each strategy leave a distinct geometric signature? This would test whether the fingerprint reflects attacker behavior or model-specific response patterns. Beyond classification, replacing hard labels with calibrated trajectory-level probabilities would enable integration with existing guardrail pipelines, where a geometric risk score could modulate the sensitivity of content-level filters as a conversation progresses.

Acknowledgments and Disclosure of Funding

This research is supported by the Indian Institute of Management Bangalore Young Faculty Research Grant. SM acknowledges the creators of the anime Psycho-Pass for being the source of inspiration for this work.

References

[1] M. A. Ayub and S. Majumdar. Embedding-based classifiers can detect prompt injection attacks, 2024. URL https://arxiv.org/abs/2410.22284.

[2] L. Derczynski, E. Galinkin, J. Martin, S. Majumdar, and N. Inie. garak: A framework for security probing large language models, 2024. URL https://arxiv.org/abs/2406.11036.

[3] S. Gooding and E. Grefenstette. Interaction dynamics as a reward signal for llms, 2025. URL https://arxiv.org/abs/2511.08394.

[4] D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, New York, 1966.

[5] W. Hackett, L. Birch, S. Trawicki, N. Suri, and P. Garraghan. Bypassing llm guardrails: An empirical analysis of evasion attacks against prompt injection and jailbreak detection systems, 2025. URL https://arxiv.org/abs/2504.11168.

[6] K. Hines, G. Lopez, M. Hall, F. Zarfati, Y. Zunger, and E. Kiciman. Defending against indirect prompt injection attacks with spotlighting, 2024. URL https://arxiv.org/abs/2403.14720.

[7] M. Huh, B. Cheung, T. Wang, and P. Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR, 21–27 Jul 2024.

[8] IBM. Introducing the ibm granite 4.1 family of models, 2026. URL https://research.ibm.com/blog/granite-4-1-ai-foundation-models. Accessed: Apr 30, 2026.

[9] P. Laban, H. Hayashi, Y. Zhou, and J. Neville. Llms get lost in multi-turn conversation, 2025. URL https://arxiv.org/abs/2505.06120.

[10] Y. Leviathan, M. Kalman, and Y. Matias. Prompt repetition improves non-reasoning llms, 2025. URL https://arxiv.org/abs/2512.14982.

[11] X. Liu, N. Xu, M. Chen, and C. Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models, 2024. URL https://arxiv.org/abs/2310.04451.

[12] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones. catch22: Canonical time-series characteristics, 2019. URL https://arxiv.org/abs/1901.10200.

[13] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. Forsyth, and D. Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. URL https://arxiv.org/abs/2402.04249.

[14] A. Mehrotra, M. Zampetakis, P. Kassianik, B. Nelson, H. Anderson, Y. Singer, and A. Karbasi. Tree of attacks: Jailbreaking black-box llms with crafted prompts. *arXiv preprint arXiv:2312.02119*, 2024.

[15] Meta. Model card - prompt guard, 2024. URL https://huggingface.co/meta-llama/Prompt-Guard-86M. Accessed: Apr 30, 2026.

[16] Meta. Llama guard 4: Natively multimodal safeguard model, 2025. URL https://huggingface.co/meta-llama/Llama-Guard-4-12B. Accessed: Apr 30, 2026.

[17] G. D. L. Munoz, A. J. Minnich, R. Lutz, R. Lundeen, R. S. R. Dheekonda, N. Chikanov, B.-E. Jagdagdorj, M. Pouliot, S. Chawla, W. Maxwell, B. Bullwinkel, K. Pratt, J. de Gruyter, C. Siska, P. Bryan, T. Westerhoff, C. Kawaguchi, C. Seifert, R. S. S. Kumar, and Y. Zunger. Pyrit: A framework for security risk identification and red teaming in generative ai systems, 2024. URL https://arxiv.org/abs/2410.02828.

[18] T. Roccia. Nova: The prompt pattern matching, 2025. URL https://github.com/Nova-Hunting/nova-framework. Accessed: Apr 30, 2026.

[19] M. Russinovich, A. Salem, and R. Eldan. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack, 2025. URL https://arxiv.org/abs/2404.01833.

[20] A. Simhi, F. Barez, M. Tutek, Y. Belinkov, and S. B. Cohen. Old habits die hard: How conversational history geometrically traps llms, 2026. URL https://arxiv.org/abs/2603.03308.

[21] Vijil. Model card for vijil prompt injection, 2025. URL https://huggingface.co/vijil/mbert-prompt-injection. Accessed: Apr 30, 2026.

[22] J. Wang, F. Wu, W. Li, J. Pan, E. Suh, Z. M. Mao, M. Chen, and C. Xiao. Fath: Authentication-based test-time defense against indirect prompt injection attacks, 2024. URL https://arxiv.org/abs/2410.21492.

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

A Crescendo attack mechanism

Crescendo [19] is a multi-turn jailbreak strategy that escalates toward a forbidden objective across several turns rather than encoding it in a single prompt. Its premise is that single-turn safety filters score each message in isolation, so a request that would be refused outright can often be elicited indirectly: the attacker first establishes a conversational context that makes the harmful request look like a natural continuation, and only then closes the loop. We summarize the mechanism here; the full algorithm and prompt templates are described in the original paper.

Escalation. At each turn, the adversarial bot produces a prompt that takes the objective bot's previous response as a foothold and pushes one increment toward the seed objective—moving, for example, from background context, to specific details, to the explicit forbidden request. Each step stays within the conversational frame already established; this accumulating frame is what gives the attack its name. The seed objective itself is never revealed verbatim until the surrounding context has been primed enough that the objective bot is likely to comply.

Scoring and termination. After each assistant turn, the scoring bot (LLM-as-judge) rates how completely the response has satisfied the seed objective on a $[0, 1]$ scale. The attack is declared a *success* the first time this score crosses the threshold (0.8 in our setup); otherwise it is a *failure* once the turn budget $T_{\max} = 8$ is exhausted. Each turn comprises one user message followed by one assistant message.

Backtracking. When the objective bot refuses or produces a non-compliant response, the adversarial bot is allowed to *backtrack*: it discards the most recent user–assistant exchange and retries the same conversational position with a rephrased prompt. This lets the attacker route around individual refusals without resetting the entire dialogue. The number of backtracks per attack is capped at 2. The trajectory $\mathcal{T}$ that Sections 2–3 encode and analyze is the conversation that remains after the orchestrator commits to a single forward path—i.e., backtracked branches are pruned from the recorded turns.

Implementation. We use the Crescendo attack executor from PyRIT [17], which implements the loop above with the prompt templates released alongside the original paper. Bot identities and decoding parameters are listed in Section 2.

B Attack generation statistics

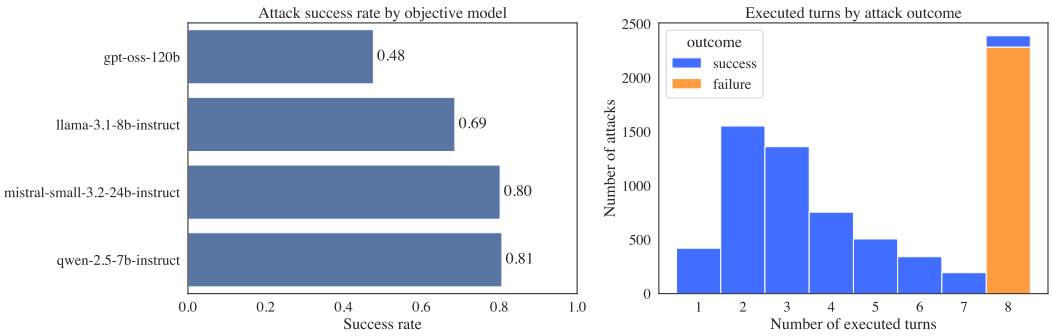

Figure 4: Aggregate statistics of the Crescendo attack pool used throughout the paper. *Left:* attack success rate per objective LLM (sorted ascending). *Right:* empirical distribution of the number of executed turns per attack, stacked by outcome; failures saturate the $T_{\max}$=8 budget by construction while successes are concentrated in the first three to four turns.

Figure B summarizes the conversational dataset that all three experiments share: 7,525 Crescendo attacks executed against four target LLMs, with each attack capped at $T_{\max} = 8$ turns and 2 backtracks. Attack difficulty varies substantially by target: `gpt-oss-120b` resists the largest fraction (48% success rate), `mistral-small-3.2-24b-instruct` and `qwen-2.5-7b-instruct` are the most permissive (80% and 81%), and `llama-3.1-8b-instruct`—which doubles as the adversarial bot, so this is the within-family setting where adversary and target share a base checkpoint—sits in between at 69%. The aggregate success rate over the full pool, 69.6%, is the value quoted in Section 3.

The right panel underwrites the length-confound diagnosis of Experiment 3.1. By construction of the attack strategy, every *failed* attack exhausts the turn budget exactly, so the failure histogram is a delta at $T_{\max} = 8$; *successful* attacks terminate strictly faster, with a mean of 3.3 turns and a long thin tail. Any feature whose value scales with conversation length therefore inherits a near-perfect signal from this near-deterministic coupling between outcome and turn count, which is precisely the artifact Experiment 3.1 sets out to isolate.

C Classification metrics

We report five standard binary-classification metrics throughout the paper. Let $y_i \in \{0, 1\}$ be the true label of conversation i (with 1 denoting a successful attack), $s_i \in [0, 1]$ a continuous score produced by the classifier, and $\hat{y}_i = \mathbf{1}[s_i \geq 0.5]$ the corresponding hard prediction at the default 0.5 threshold. The four entries of the confusion matrix are

$$\mathrm{TP} = \#\{i : y_i = 1,\ \hat{y}_i = 1\}, \qquad \mathrm{FP} = \#\{i : y_i = 0,\ \hat{y}_i = 1\},$$
$$\mathrm{TN} = \#\{i : y_i = 0,\ \hat{y}_i = 0\}, \qquad \mathrm{FN} = \#\{i : y_i = 1,\ \hat{y}_i = 0\}.$$

Threshold-dependent metrics. The first three metrics are evaluated at the default threshold (0.5) and characterize the classifier at a single operating point:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{F1} = \frac{2\,\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}.$$

Threshold-free metrics. The remaining two summarize the entire score-rank ordering and are independent of any threshold choice. The Area Under the ROC curve is most cleanly stated as a rank statistic,

$$\mathrm{AUROC} \;=\; \Pr\big(s_i > s_j \;\big|\; y_i = 1,\ y_j = 0\big),$$

i.e. the probability that a randomly drawn positive instance receives a higher score than a randomly drawn negative one; it is invariant to monotone score transformations, insensitive to class imbalance, and equal to $\frac{1}{2}$ for a random classifier. Average Precision (AP), the area under the precision-recall curve, is computed as a finite sum over the unique score thresholds:

$$\mathrm{AP} \;=\; \sum_{n} \big(\mathrm{Recall}_n - \mathrm{Recall}_{n-1}\big)\, \mathrm{Precision}_n,$$

where the index n runs over the sorted unique values of s_i. Unlike AUROC, AP is sensitive to the positive-class base rate π: a constant predictor achieves $\mathrm{AP} = \pi$, so AP values should be read against this floor rather than against the 0.5 AUROC baseline.

D Experiment 3: Classification performance

Table 4: Classification performance in Experiment 3 as a function of retained prefix length. Conversations are first length-equalized to six turns ($\mathcal{T}^{(:6)}$) and then truncated by removing the last k turns, leaving $6-k$ turns; k=0 retains the full six-turn conversation. `llama-guard-4-12b` (LG) is included as a content-based runtime safety baseline. All entries use the lexical encoder ϕ_L.

k	Prefix	**Method**	**Precision**	**Recall**	**F1**	**AUROC**
0	6	LR	0.318	0.721	0.442	0.697
		GB	0.317	0.705	0.438	0.682
		LG	0.425	0.242	0.308	0.575
1	5	LR	0.316	0.636	0.422	0.668
		GB	0.315	0.758	0.445	0.659
		LG	0.288	0.159	0.205	0.522
2	4	LR	0.315	0.603	0.414	0.648
		GB	0.286	0.603	0.388	0.622
		LG	0.211	0.088	0.124	0.494
3	3	LR	0.266	0.593	0.367	0.630
		GB	0.319	0.674	0.433	0.651
		LG	0.280	0.104	0.151	0.512
4	2	LR	0.307	0.593	0.405	0.609
		GB	0.275	0.550	0.367	0.586
		LG	0.167	0.043	0.068	0.488

Table D reports the per-metric classification performance behind the ROC and PR curves shown in Figure 3.3. For each trim level $k \in \{0, \ldots, 4\}$ (with retained prefix of $6 - k$ turns; k=0 is a single-seed replay of Experiment 3.2's lexical+LR setup) we list precision, recall, F1, and AUROC for the geometric logistic-regression and gradient-boosting classifiers (LR, GB) and for the `llama-guard-4-12b` content baseline (LG). Llama Guard's F1 column exposes the actual single-turn detection quality, which decays toward zero as the prefix shortens, while the geometric classifiers retain F1 around 0.4 and AUROC well above chance at every horizon.

E Proofs of Theoretical Results

Proof of Proposition 1. Under the Gaussian location model with equal covariances, the Bayes-optimal decision rule is Fisher's linear discriminant, whose projection yields the scalar

$$Z \;=\; \frac{\mu_\ell^{(1)} - \mu_\ell^{(0)}}{\sigma_\ell^2}\,\ell \;+\; \frac{(\boldsymbol{\mu}_g^{(1)} - \boldsymbol{\mu}_g^{(0)})^\top}{\sigma_g^2}\,\mathbf{g}.$$

By conditional independence, $\mathrm{Var}(Z \mid Y) = (\mu_\ell^{(1)} - \mu_\ell^{(0)})^2/\sigma_\ell^2 + \|\boldsymbol{\mu}_g^{(1)} - \boldsymbol{\mu}_g^{(0)}\|^2/\sigma_g^2 = \mathrm{SNR}_\ell^2 + \mathrm{SNR}_g^2$. Let $Z_1 \sim Z \mid Y{=}1$ and $Z_0 \sim Z \mid Y{=}0$ be independent draws from the two class-conditional distributions. Their difference $Z_1 - Z_0$ is Gaussian with mean $d' \cdot \sigma_Z$ and variance $2\sigma_Z^2$, where

$d' = \sqrt{\mathrm{SNR}_\ell^2 + \mathrm{SNR}_g^2}$ and $\sigma_Z^2 = \mathrm{Var}(Z \mid Y)$. The AUROC equals

$$\mathrm{AUROC} = P(Z_1 > Z_0) = \Phi\left(\frac{d'}{\sqrt{2}}\right),$$

which is the standard result for equal-variance Gaussian discriminants [see, e.g., 4]. Setting $\mathrm{SNR}_\ell = 0$ yields the length-equalized case: $\mathrm{AUROC}_{\mathrm{shape}} = \Phi\big(\mathrm{SNR}_g/\sqrt{2}\big)$. □

Proof of Proposition 2. For a classifier at the midpoint $t = (\mu_h^{(0)}(k) + \mu_h^{(1)}(k))/2$, an error under class y requires h_k to deviate from its class mean by at least $\Delta(k)/2$ in the direction of the opposing class. This is a one-sided event: under class y=0, misclassification requires $h_k \geq t$, i.e., $h_k - \mu_h^{(0)}(k) \geq \Delta(k)/2$; under class y=1, it requires $h_k \leq t$, i.e., $\mu_h^{(1)}(k) - h_k \geq \Delta(k)/2$. Under Assumption 2, $h_k \mid Y{=}y$ is sub-Gaussian with variance proxy σ_h^2, so the one-sided sub-Gaussian tail bound gives

$$P\big(h_k - \mu_h^{(y)}(k) \geq \Delta(k)/2 \mid Y{=}y\big) \;\leq\; \exp\left(-\frac{\Delta(k)^2}{8\,\sigma_h^2}\right).$$

The same bound applies to the other class by symmetry. Averaging over the two classes preserves the bound. The minimum prefix length follows by inverting the inequality $\exp(-\Delta(k)^2/(8\sigma_h^2)) \leq \varepsilon$. □

Proof of Proposition 3. AUROC is a rank statistic: it equals the probability that a randomly drawn positive instance receives a higher score than a randomly drawn negative instance, and therefore depends only on the rank ordering of classifier scores across samples.

For classifiers whose decisions depend only on univariate feature ranks, each decision step compares the value of a single feature against a threshold. Since rank preservation guarantees that the relative ordering of each feature's values is identical under both encoders, every such comparison yields the same outcome, and the final score ranking is preserved. AUROC invariance follows immediately.

For linear classifiers, the score is $s = \beta_0 + \sum_j \beta_j F_j(\phi(\mathcal{T}))$. Rank preservation of individual features does not in general imply rank preservation of a linear combination, since independent monotone transformations can alter relative feature scales. However, if the within-class covariance structure of the feature vector is shared across encoders up to a global scalar $c > 0$ (i.e., $\Sigma_{\phi_2} = c\,\Sigma_{\phi_1}$), then the optimal coefficient vectors under the two encoders are proportional, and the resulting scores are related by a positive affine transformation. Since AUROC is invariant to monotone score transformations, it is identical under both encoders. □