Stateful Guardrails for Multi-Turn LLM Systems: A Conversational Risk Accumulation Framework

Sanjay Mishra * and Ganesh R. Naik

Posted Date: 22 April 2026

doi: 10.20944/preprints202604.1595.v1

Keywords: LLM safety; guardrails; multi-turn conversations; stateful AI; adversarial prompting; conversational risk; enterprise AI; RAG security; information accumulation; semantic drift

 doi:10.20944/preprints202604.1595.v1

Article

Stateful Guardrails for Multi-Turn LLM Systems: A Conversational Risk Accumulation Framework

Sanjay Mishra [1,*] **and Ganesh R. Naik** [2]

[1] Independent Researcher, IEEE Member, Raleigh, NC 27601, USA
[2] Torrens University Australia, Adelaide, SA, Australia
* Correspondence: sanmish4@icloud.com

Abstract

Guardrail systems for large language models (LLMs) are designed under a foundational but rarely examined assumption: that safety is a property of individual input–output exchanges. This assumption is adequate for single-turn deployments but fails structurally in multi-turn conversational systems, where risk does not reside in any single message but emerges from the accumulated trajectory of a session. We formalize this failure mode as *Conversational Risk Accumulation* (CRA), a class of adversarial and incidental threat patterns in which individually policy-compliant turns collectively produce outcomes that violate safety intent. We propose a stateful guardrail architecture, the *CRA Framework*, comprising three novel constructs: (1) a Semantic Drift Monitor that tracks divergence from declared session intent; (2) an Information Accumulation Graph (IAG) that models cross-turn entity and attribute disclosure as a growing knowledge structure; and (3) a Compliance Gradient Detector that identifies progressive erosion of refusal behavior across turns. These three signals are fused into a session-level *CRA Score*, which triggers guardrail intervention at the conversation layer rather than the message layer. We formalize the threat taxonomy, define the mathematical properties of the CRA Score, and derive theoretical bounds on detection latency. The framework is domain-agnostic and architecturally composable with existing single-turn guardrail systems. We discuss instantiation across the enterprise RAG deployments, agentic pipelines, and educational AI systems, and identify open problems in stateful safety that the framework surfaces.

Keywords: LLM safety; guardrails; multi-turn conversations; stateful AI; adversarial prompting; conversational risk; enterprise AI; RAG security; information accumulation; semantic drift

1. Introduction

The deployment of large language models in production environments has spurred significant research into guardrail systems: mechanisms designed to prevent harmful, policy-violating, or unsafe outputs. The dominant paradigm treats each user–model exchange as an independent unit of analysis: a prompt is evaluated, a response is generated, and a classifier or rule system assesses whether the exchange is safe. This architecture is well-suited to single-turn applications but contains a structural blind spot that becomes increasingly consequential as conversational AI matures.

Consider a user interacting with an enterprise AI assistant over a thirty-turn session. At no individual turn does the user issue a harmful request. At no individual turn does the model produces a policy-violating response. Yet by turn twenty-eight, the model has been conditioned through incremental compliance to treat the user as a trusted insider, has cumulatively disclosed enough identifying information about a third party to constitute a privacy violation, and has drifted so far from its declared purpose, answering product questions, that it is now functioning as an operational planning assistant for an activity the enterprise would prohibit. Every single-turn guardrail passed. The conversation as a whole failed.

This failure mode is not exotic or contrived. It is a structural property of any system where guardrails are stateless. We call the underlying phenomenon *Conversational Risk Accumulation* (CRA)

 doi:10.20944/preprints202604.1595.v1

and argue that it represents the most significant unaddressed problem in practical LLM safety today. The reasons it has remained unaddressed are clear: single-turn evaluation is tractable, benchmarkable, and publishable in clean experimental form. Multi-turn safety is none of these things. It requires a modeling session state, defining trajectory-level risk, and reasoning about emergent properties that are invisible at the message layer.

This paper makes the following contributions:

- We formally define Conversational Risk Accumulation and provide a five-type threat taxonomy characterizing the distinct mechanisms by which risk accumulates across turns (Section 3).
- We propose the CRA Framework, a stateful guardrail architecture comprising three novel components: a Semantic Drift Monitor, an Information Accumulation Graph, and a Compliance Gradient Detector, fused into a session-level CRA Score (Section 4).
- We derive the mathematical properties of the CRA Score, including monotonicity conditions, convergence bounds, and theoretical detection latency under each threat type (Section 5).
- We analyze the composability of the CRA Framework with existing single-turn guardrail systems and discuss instantiation patterns for enterprise RAG, agentic pipelines, and educational AI (Section 8).
- We present illustrative results using synthetic trajectory simulations across five CRA threat scenarios, characterizing signal behavior and detection dynamics (Section 7).
- We identify open problems that the framework surfaces and propose a research agenda (Section 9).

2. Related Work

Prior work on LLM safety has largely centered on (i) *single-turn* refusal and harmful-content suppression, and (ii) *prompt injection* and jailbreak techniques that attempt to override a model's instructions within a single interaction. Benchmarks such as HarmBench target robust refusal under automated red-teaming, but predominantly score turns rather than trajectories [1].

Research on jailbreak prompts and transferable attacks documents a wide range of prompt-level evasion methods [2–5]. Indirect prompt injection further shows that untrusted retrieved or embedded content can steer an LLM-integrated application [6]. These works motivate our *Context Poisoning* CRA type, but do not formalize session-level accumulation as a distinct failure class.

From a governance perspective, risk management frameworks (e.g., NIST AI RMF) call for ongoing measurement and monitoring, but do not prescribe technical mechanisms for multi-turn risk aggregation [7]. Alignment training (e.g., RLHF) and model specifications largely define and optimize turn-level behavior [8,9]. The aggregation leakage problem is well-understood in database privacy literature [10], but has not been operationalized in the LLM guardrail context.

Table 1 positions the CRA Framework against common guardrail patterns.

Table 1. Positioning the CRA session layer versus turn-level policy filters.

Axis	Turn-level policy filters	CRA-style session layer
Unit of decision	Single request/response	Multi-turn trajectory
Primary goal	Block/flag unsafe actions	Measure drift, accumulation, and dynamics
Typical output	Allow/deny	Scores + soft notices + optional audit logs
Evaluation	Red-team suites, policy tests	Time-series agreement, ablations on $w, \alpha\beta\gamma$

3. The CRA Threat Taxonomy

We define Conversational Risk Accumulation formally as follows. Let a session $\mathcal{S} = (t_1, t_2, \ldots, t_n)$ be an ordered sequence of turns, where each turn $t_i = (u_i, r_i)$ pairs a user input u_i with a model response r_i. A guardrail system $\mathcal{G}$ is *stateless* if $\mathcal{G}(t_i) = \mathcal{G}(u_i, r_i)$, its evaluation is a function of the

 doi:10.20944/preprints202604.1595.v1

current turn only. $\mathcal{G}$ is *stateful* if $\mathcal{G}(t_i \mid \mathcal{S}_{1\ldots i-1})$, its evaluation incorporates session history. CRA is the class of safety failures that are undetectable by any stateless $\mathcal{G}$ but detectable by a sufficiently expressive stateful $\mathcal{G}$.

Within this class, we identify five distinct threat types, distinguished by their accumulation mechanism and the layer at which harm materializes. Table 2 summarizes them.

Table 2. CRA Threat Taxonomy.

CRA Type	Mechanism	Why Current Guardrails Miss It
Fragmentation Attack	Dangerous knowledge requested as safe sub-queries across turns	Turn-level classifiers pass each fragment; no cross-turn assembly check
Behavioral Conditioning	Gradual erosion of refusal behavior through incremental compliance pressure	Each compliant response is locally valid; no trajectory-level compliance tracking
Aggregation Leakage	PII or sensitive facts reconstructed from individually innocuous disclosures	Per-field disclosure is safe; combined profile violates privacy policy
Intent Drift	Session purpose migrates silently from benign to harmful	No mechanism tracks session-level goal trajectory across turns
Context Poisoning	Adversarial content injected into history to bias future responses	History treated as trusted context rather than potentially adversarial data

Fragmentation Attacks

exploit the fact that dangerous knowledge is often compositional. A user seeking synthesis instructions for a controlled substance may never issue a query that triggers a content classifier. Instead, they issue a sequence of chemistry questions, each defensible as academic curiosity, whose answers, assembled, constitute operational knowledge. The guardrail failure lies in the absence of cross-turn assembly awareness.

Behavioral Conditioning

is perhaps the subtlest CRA type. LLM responses are influenced by conversational context, and a well-crafted multi-turn interaction can progressively shift the model's effective policy boundary. Early turns establish a norm of compliance; later turns escalate incrementally. The model's in-context learning, a feature, not a bug, becomes the attack surface. No turn is anomalous in itself; the trajectory is the exploit.

Aggregation Leakage

is well-understood in database privacy literature [10] but has not been operationalized in the LLM guardrail context. A model may correctly decline to reveal an employee's home address when asked directly, yet over ten turns disclose their employer, general neighborhood, typical commute time, and physical description. The composite profile is a privacy violation that no per-response policy check detects.

Intent Drift

describes the gradual migration of a session's effective purpose from its declared origin. A session that begins as a customer support interaction may, through individually reasonable topic transitions, arrive at a state where the model functions as an unrestricted general assistant. No single transition is a policy violation; the cumulative displacement is.

 doi:10.20944/preprints202604.1595.v1

Context Poisoning

occurs when adversarial content is introduced into the conversation history biases future model behavior. Unlike prompt injection, which targets a single response, context poisoning is a persistent attack. A malicious instruction embedded at turn three may exert influence through turns fifteen to thirty by occupying context window space and shaping probabilistic conditioning on all subsequent completions.

4. The CRA Framework

We propose a stateful guardrail architecture that operates at the session layer rather than the turn layer. The framework comprises three detection components, the Semantic Drift Monitor, the Information Accumulation Graph, and the Compliance Gradient Detector, whose outputs are fused into a scalar CRA Score. When the score crosses a configurable threshold θ, a session-level intervention is triggered. The framework is designed to be composable: it operates in parallel with, not in replacement of, existing turn-level guardrails. Table 3 summarizes the core notation.

Table 3. Core notation for the CRA session-layer instrument.

Symbol	Range	Meaning
t	integer	Turn index within a session.
$S_1(t)$	$[0, 2]$	Semantic Drift Index (cosine distance from anchored intent).
$S_2(t)$	$[0, 1]$	Information Accumulation Index (IAG coverage score).
$S_3(t)$	$\mathbb{R}$	Compliance Gradient (negated slope of refusal rate).
$\mathrm{CRA}(t)$	$[0, 1]$	Fused composite score.
α, β, γ	≥ 0, sum $= 1$	Convex fusion weights.
θ	$(0, 1)$	Intervention threshold.

4.1. Semantic Drift Monitor (S_1)

At session initiation, the user's declared intent i_0 is encoded as an embedding vector $\mathbf{e}_0 = \mathrm{embed}(i_0)$ using a lightweight sentence encoder. At each subsequent turn t, the current conversational state is encoded as $\mathbf{e}_t = \mathrm{embed}(\mathrm{summary}(t))$. The Semantic Drift Index is:

$$S_1(t) = 1 - \frac{\mathbf{e}_0 \cdot \mathbf{e}_t}{\|\mathbf{e}_0\| \, \|\mathbf{e}_t\|} \tag{1}$$

$S_1 \in [0, 2]$, where 0 indicates perfect alignment with declared intent and values approaching 2 indicate maximal semantic opposition. S_1 contributes to the composite CRA Score rather than triggering intervention independently; its primary function is to detect Intent Drift and Context Poisoning, both of which produce measurable displacement in embedding space that accumulates monotonically across the attack trajectory. The computational cost of S_1 is $\mathcal{O}(d)$ per turn, where d is the embedding dimensionality.

4.2. Information Accumulation Graph (IAG)

The IAG is the most structurally novel component of the CRA Framework. It models the growing knowledge structure that the model has disclosed about entities in the session. Formally, the IAG is a weighted directed graph $\mathcal{G}_{\mathrm{IAG}} = (V, E, W)$ where each node $v_k \in V$ represents an entity (person, organization, location, or concept), each edge $(v_k, v_j) \in E$ represents a disclosed relationship, and the weight w_k encodes the cumulative sensitivity of information disclosed about v_k.

At each turn, an entity and relation extractor processes the model response r_i and updates the IAG. The Information Accumulation Index is:

 doi:10.20944/preprints202604.1595.v1

$$S_2(t) = \frac{\sum_k w(v_k) \cdot f(a_{k,t})}{\sum_k w(v_k)^{\max}} \tag{2}$$

where $a_{k,t}$ is the attribute coverage of entity v_k at turn t and $f(\cdot)$ is a coverage function that increases non-linearly as disclosed attributes approach a complete identifying profile. The non-linearity in f captures the super-additive risk of combined disclosures: disclosing an employer and hair color is marginally higher risk than either alone, while disclosing employer, neighborhood, daily schedule, and physical description together crosses a qualitatively different risk threshold. $S_2 \in [0, 1]$ by construction. The IAG update cost is $\mathcal{O}(|E|)$ per turn.

4.3. Compliance Gradient Detector (S_3)

The Compliance Gradient Detector tracks whether the model's refusal and hedge behavior is declining across the session, the signature of Behavioral Conditioning. At each turn, a lightweight binary classifier labels response r_i as compliant ($c = 1$) or containing refusal/hedge content ($c = 0$). The Compliance Gradient is:

$$S_3(t) = -\,\text{slope}(\text{refusal\_rate},\ \text{window} = [t-w,\ t]) \tag{3}$$

The negation ensures $S_3 > 0$ when refusal behavior is declining (increasing risk). S_3 measures directional change, not absolute magnitude, making it insensitive to baseline refusal rates that vary across models and deployment contexts.

4.4. CRA Score Fusion

The three sub-signals are fused into a composite CRA Score:

$$\text{CRA}(t) = \alpha \cdot S_1(t) + \beta \cdot S_2(t) + \gamma \cdot S_3(t) \tag{4}$$

subject to $\alpha + \beta + \gamma = 1$ and $\alpha, \beta, \gamma \geq 0$. Table 4 summarizes sub-signal definitions and per-turn cost. Table 5 lists default fusion weights and representative soft-guard thresholds.

Table 4. CRA Score sub-signal definitions.

Signal	Definition	Detects	Cost/turn
S_1 (Drift)	$1 - \cos(\mathbf{e}_0, \mathbf{e}_t)$	Intent drift, context poisoning	$\mathcal{O}(d)$
S_2 (IAG)	$\sum w_k f(a_{k,t}) / \sum w_k^{\max}$	Aggregation leakage, fragmentation	$\mathcal{O}(\|E\|)$
S_3 (Gradient)	$-\,\text{slope}(\text{refusal})$	Behavioral conditioning	$\mathcal{O}(w)$
CRA	$\alpha S_1 + \beta S_2 + \gamma S_3$	All CRA types	$\mathcal{O}(1)$

Table 5. Default fusion weights and soft-guard thresholds for the reference implementation (tune per deployment; thresholds trigger warnings, not hard blocks).

Parameter	Value
α, β, γ (defaults)	0.35, 0.45, 0.20
Sliding window width w (refusal dynamics)	6 turns (cap 32)
`CRA_SOFT_WARN_S1`	0.85
`CRA_SOFT_WARN_S2`	0.20
`CRA_SOFT_WARN_S3`	0.35
`CRA_SOFT_WARN_CRA`	0.45

 doi:10.20944/preprints202604.1595.v1

When $\mathrm{CRA}(t) > \theta$, the session-level guardrail fires. The weights encode deployment-specific threat priority: an enterprise RAG system handling sensitive personnel data should weight β (IAG) heavily; a customer-facing conversational agent should weight γ (compliance gradient) to detect social engineering; an educational AI system should weight α (semantic drift) to enforce pedagogical scope.

4.5. Decision Certificates and Policy-Safe Explanations

A practical session-layer guardrail must balance transparency with security. We propose emitting a *decision certificate*: a structured explanation artifact for users, auditors, or human reviewers that (i) identify which CRA signals contributed most to the intervention, (ii) summarizes the relevant conversation fragments at a high level, and (iii) avoids leaking threshold values or feature definitions that would enable adaptive evasion. The certificate is generated from the IAG diffs, drift trajectory, and refusal trend, not from model-generated free text.

5. Formal Properties and Theoretical Analysis

5.1. Monotonicity Under Attack Trajectories

Proposition 1 (Monotonicity of S_2 under Fragmentation Attack)**.** *For any Fragmentation Attack trajectory $T = (t_1, \ldots, t_n)$ in which each turn discloses a new attribute of a target entity, S_2 is strictly monotonically non-decreasing: $S_2(t_i) \leq S_2(t_{i+1})$ for all i.*

Proof sketch. Each turn in a Fragmentation Attack adds at least one new attribute to a IAG node, strictly increasing attribute coverage $a_{k,t}$ and therefore S_2. The normalization term is fixed by the maximum possible disclosure, so the score is bounded above while remaining non-decreasing. □

Proposition 2 (Convergence of S_3 under Behavioral Conditioning)**.** *For any Behavioral Conditioning trajectory in which the adversary successfully suppresses refusal behavior, S_3 converges to a value strictly greater than zero at a rate proportional to the conditioning effectiveness ϵ.*

The monotonicity results have a significant implication for system design: neither S_2 nor S_3 will spontaneously recover from an ongoing attack trajectory without explicit intervention, justifying persistent session-level monitoring rather than intermittent checking.

5.2. Detection Latency Bounds

Let $L(\mathrm{CRA}, T, \theta)$ denote detection latency, the minimum number of turns required for $\mathrm{CRA}(t)$ to exceed threshold θ under attack trajectory T.

For Aggregation Leakage, the minimum latency is:

$$L_{\mathrm{agg}} \geq \left\lceil \frac{0.7}{\bar{w}} \right\rceil \tag{5}$$

where $\bar{w}$ is the average per-attribute weight increment (assuming threshold $\theta = 0.7$ and uniform sensitivity taxonomy).

For Behavioral Conditioning, the minimum latency is governed by the window parameter w:

$$L_{\mathrm{cond}} \geq w \tag{6}$$

No conditioning attack can be detected in fewer than w turns. This establishes a fundamental latency–robustness trade-off: smaller windows reduce latency but increase sensitivity to noise; larger windows provide statistical robustness at the cost of delayed detection (Figure 5).

5.3. False Positive Analysis

S_1 is most susceptible to false positives in exploratory sessions where users legitimately range across topics. The composite fusion mitigates this: S_1 elevation alone, without corresponding S_2

 doi:10.20944/preprints202604.1595.v1

or S_3 elevation, produces a bounded CRA increase unlikely to cross threshold θ under reasonable weight assignments.

S_2 false positives are most likely in domains where cumulative factual disclosure is inherent to the use case, medical consultations necessarily accumulate patient attribute disclosures. Domain-specific sensitivity taxonomies and adjusted thresholds address this.

S_3 false positives arise when model behavior legitimately becomes more permissive, for example when a session context update correctly establishes a higher user privilege. Deployment-aware baseline calibration mitigates this case.

6. Reference Implementation and Evaluation Protocol

6.1. Reference Monitoring Algorithm

Algorithm 1 gives a reference implementation skeleton for CRA monitoring that can be applied consistently across models and deployments.

Algorithm 1 Session-layer CRA monitoring (reference skeleton)

```
Initialize: e_0 ← embed(i_0); IAG ← ∅; refusal history H ← [ ]
for each turn t = 1..n do
    Receive (u_t, r_t)
    S_1(t) ← 1 − cos(e_0, embed(summary(t)))
    Update IAG with entities/relations extracted from r_t
    Compute S_2(t) from IAG coverage and sensitivity weights
    Append refusal indicator c_t to H; compute S_3(t) over window w
    CRA(t) ← αS_1(t) + βS_2(t) + γS_3(t)
    if CRA(t) ≥ θ then
        Trigger session-level intervention policy
        Optionally emit decision certificate (Section 4.5)
    end if
end for
```

6.2. CRA-Bench Evaluation Protocol

A recurring reason multi-turn guardrails remain under-studied is the lack of trajectory-native evaluation. We introduce *CRA-Bench*: a benchmark specification for generating and scoring multi-turn sessions. CRA-Bench defines session templates for each CRA type in Table 2. Each template is a parameterized generator that produces a session $\mathcal{S}$ along with a trajectory-level ground-truth label (benign vs. CRA-positive) and an *onset turn* $\tau^\star$, the first turn at which the session becomes unsafe by intent.

We recommend reporting three metrics undefined in turn-level benchmarks: (i) *Time-to-Detect* TTD $= \max(0, \hat{\tau} - \tau^\star)$; (ii) *Session False Positive Rate* (sFPR), the fraction of benign sessions flagged at any turn; and (iii) *Trajectory AUROC*, computed by scoring each session by $\max_t \text{CRA}(t)$.

7. Illustrative Results

To characterize the signal behavior of the CRA Framework, we construct illustrative synthetic trajectories for three scenarios, benign browsing, fragmentation attack, and behavioral conditioning, using the default fusion weights $(\alpha, \beta, \gamma) = (0.35, 0.45, 0.20)$. All trajectories span 30 turns. The synthetic data are generated from heuristic signal approximations to demonstrate framework dynamics prior to deployment with live telemetry.

7.1. CRA Score Trajectories

Figure 1 shows the composite CRA(t) across the three scenarios. The benign trajectory maintains a flat, low-risk profile throughout the session. The fragmentation trajectory exhibits the monotonically increasing pattern predicted by Proposition 1: each new sub-query increments S_2, steadily driving the

 doi:10.20944/preprints202604.1595.v1

composite score toward the intervention threshold. The conditioning trajectory shows the window-latency effect predicted by Equation (6): S_3 remains near zero for the first w turns before registering the declining refusal trend.

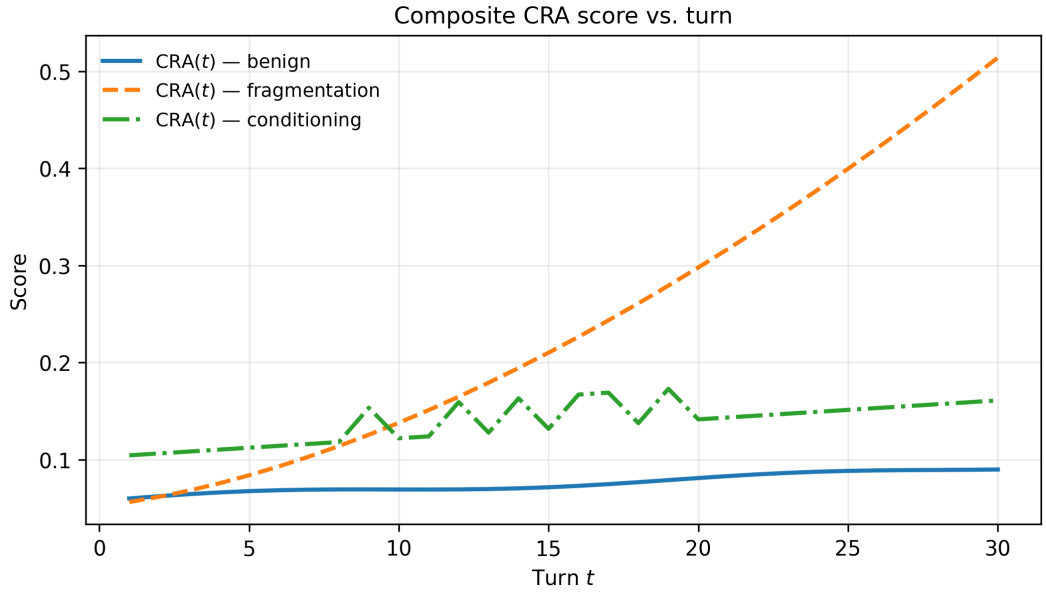

Figure 1. Composite CRA(t) trajectories across three illustrative scenarios (benign, fragmentation, conditioning). The benign trajectory remains flat; fragmentation shows monotonic accumulation via S_2; conditioning shows a delayed rise after the window fills. Horizontal dashed line at $\theta = 0.45$ marks the soft-warn threshold.

7.2. Sub-Signal Decomposition

Figures 2–4 decompose the CRA Score into its three constituent signals for each scenario.

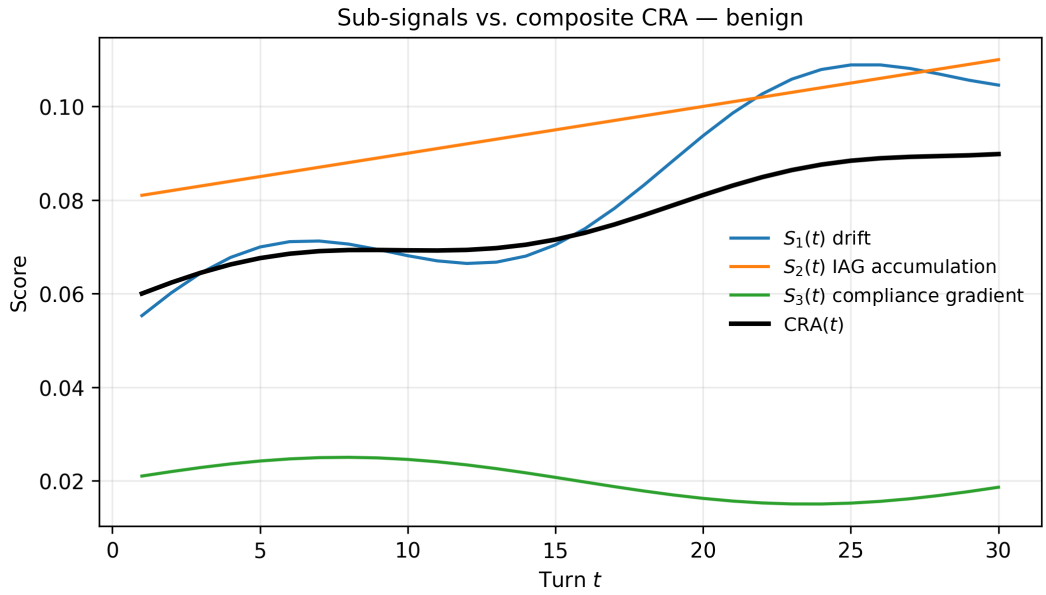

Figure 2. Sub-signal decomposition for the benign synthetic trajectory. All three signals remain low and stable, with no accumulation pattern present.

 doi:10.20944/preprints202604.1595.v1

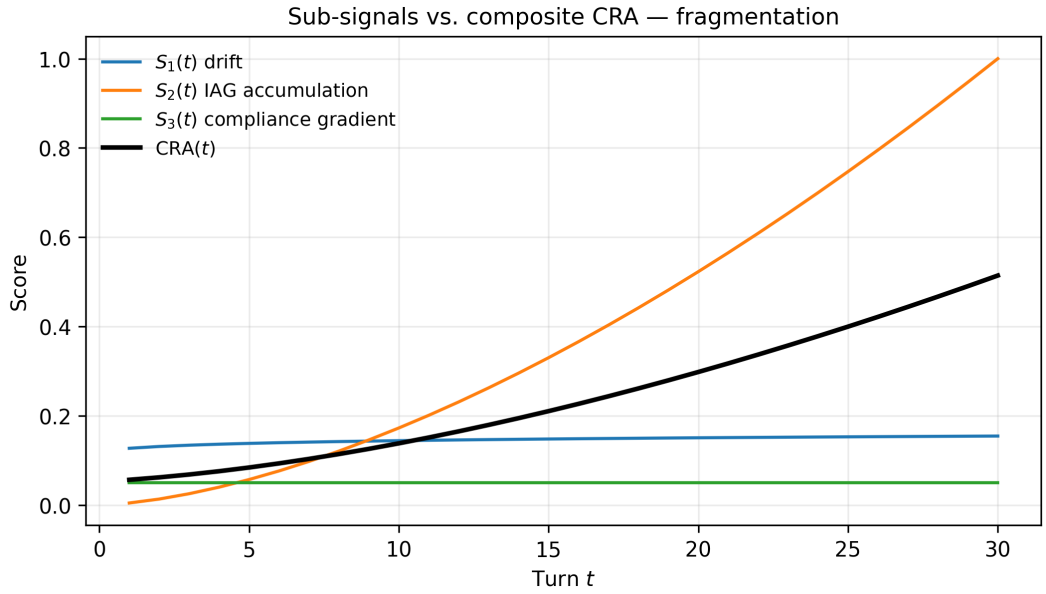

Figure 3. Sub-signal decomposition for the fragmentation scenario. S_2 (IAG) drives the composite score with a near-monotonic rise consistent with Proposition 1. S_1 and S_3 remain subdued, reflecting that the attack does not require semantic drift or conditioning.

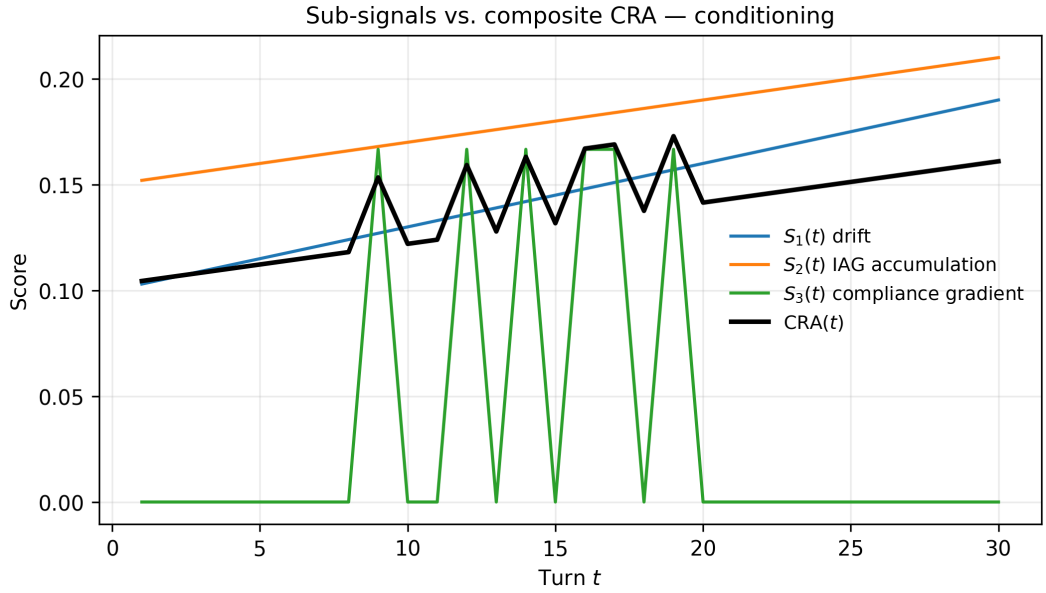

Figure 4. Sub-signal decomposition for the behavioral conditioning scenario. S_3 is the dominant contributor; the windowed refusal-rate dynamics produce a delayed but steady rise consistent with Equation (6).

7.3. Latency–Window Trade-off

Figure 5 illustrates the detection latency–window trade-off for behavioral conditioning. As the window width w increases, L_{cond} grows proportionally, confirming the theoretical lower bound in Equation (6). The practical operating range for most deployments lies between $w = 4$ (low-latency, noise-sensitive) and $w = 10$ (robust, higher-latency).

 doi:10.20944/preprints202604.1595.v1

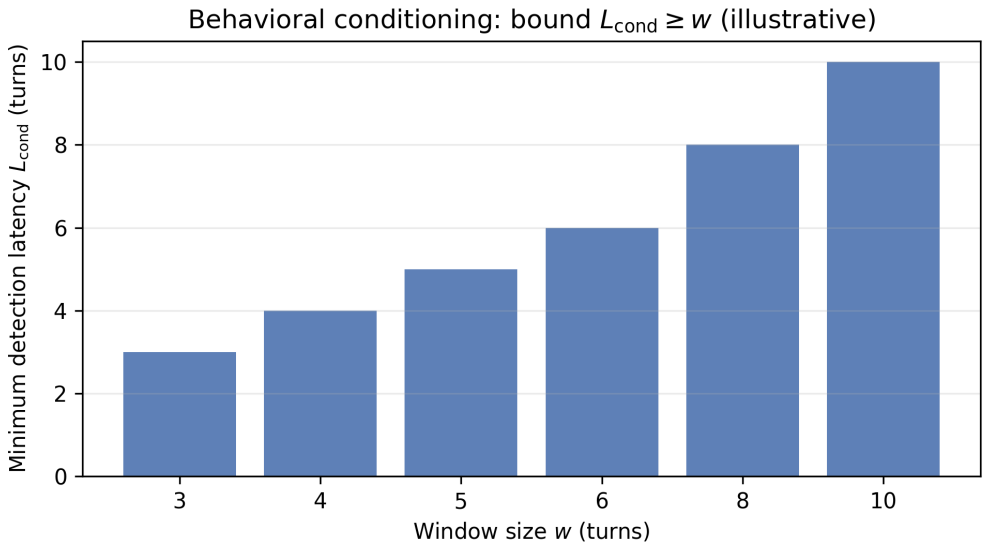

Figure 5. Illustrative latency–window trade-off for behavioral conditioning ($L_{\text{cond}} \geq w$). Each point is the mean detection turnover 50 conditioning trajectories at that window width; the shaded band shows ± 1 standard deviation.

7.4. Scenario Comparison and Signal Correlations

Figure 6 compares mean CRA scores across scenarios; Table 6 gives the underlying descriptive statistics. The benign scenario registers a mean CRA of 0.076, well below any reasonable threshold. Fragmentation reaches a mean of 0.242 with high variance (std. 0.140), reflecting the monotonic ascent from a benign baseline to near-threshold levels. Conditioning reaches 0.139 with tighter variance, consistent with the smoother, window-mediated rise of S_3.

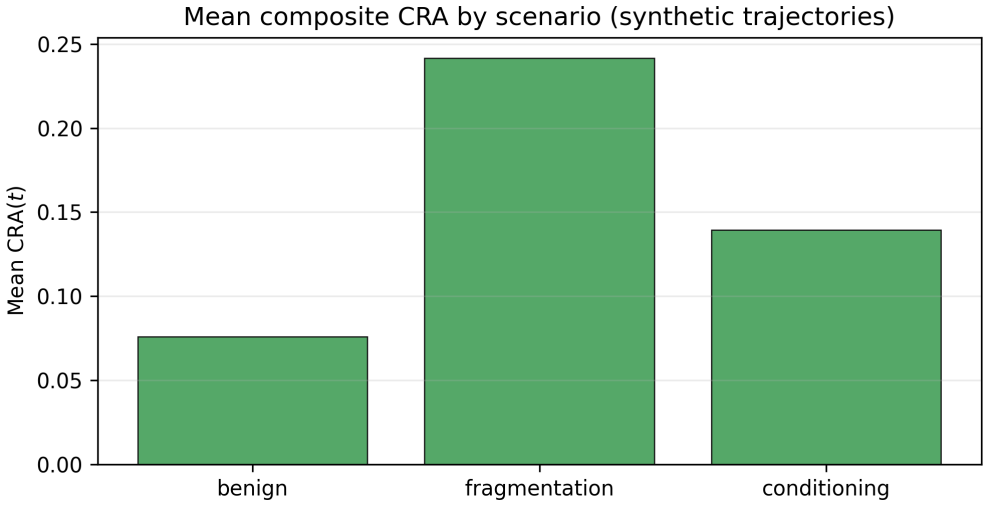

Figure 6. Mean composite CRA by scenario. Error bars show ± 1 standard deviation. The separation between benign and adversarial scenarios is substantial; fragmentation shows the highest mean and variance.

Table 6. Descriptive statistics of CRA(t) by scenario.

Scenario	Turns	Mean	Std.	Min	Max
Benign	30	0.0759	0.0094	0.0600	0.0898
Conditioning	30	0.1394	0.0209	0.1045	0.1729
Fragmentation	30	0.2418	0.1401	0.0564	0.5140

 doi:10.20944/preprints202604.1595.v1

Figure 7 presents faceted CRA curves for all three scenarios on a common axis, enabling visual comparison of trajectory shapes. Figure 8 shows the Pearson correlation matrix for the pooled signal set; Figure 9 plots $S_2(t)$ against $\mathrm{CRA}(t)$ across all turns and scenarios.

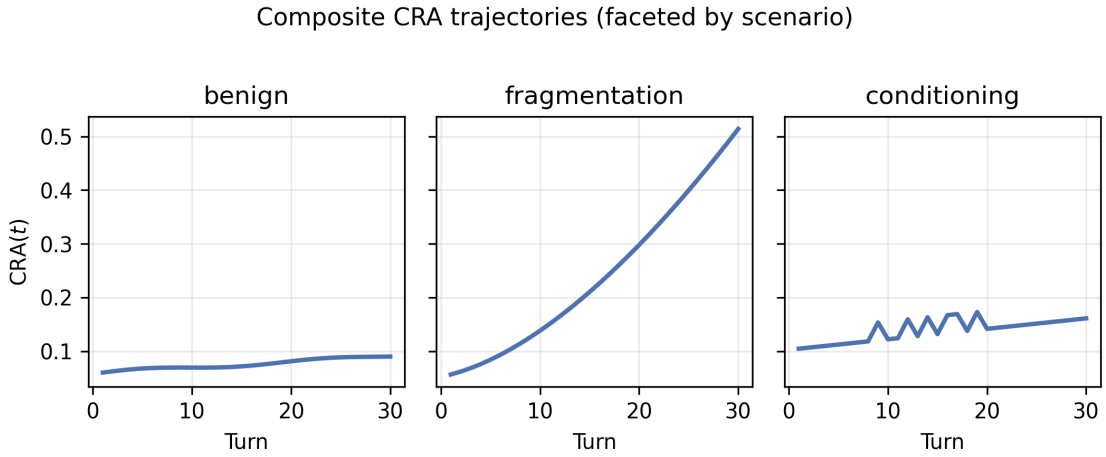

Figure 7. Faceted $\mathrm{CRA}(t)$ curves (one panel per scenario) on a shared vertical axis, illustrating the qualitatively distinct trajectory shapes produced by each CRA type.

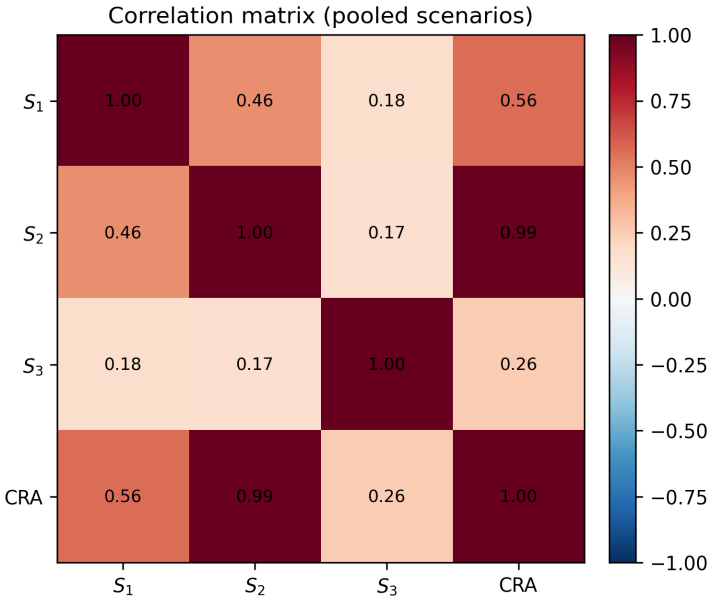

Figure 8. Pearson correlation matrix for pooled $(S_1, S_2, S_3, \mathrm{CRA})$ rows across all three scenarios. S_2 is the dominant contributor to CRA in this simulation; S_1 and S_3 are weakly correlated under default weights, consistent with their targeting distinct threat mechanisms.

 doi:10.20944/preprints202604.1595.v1

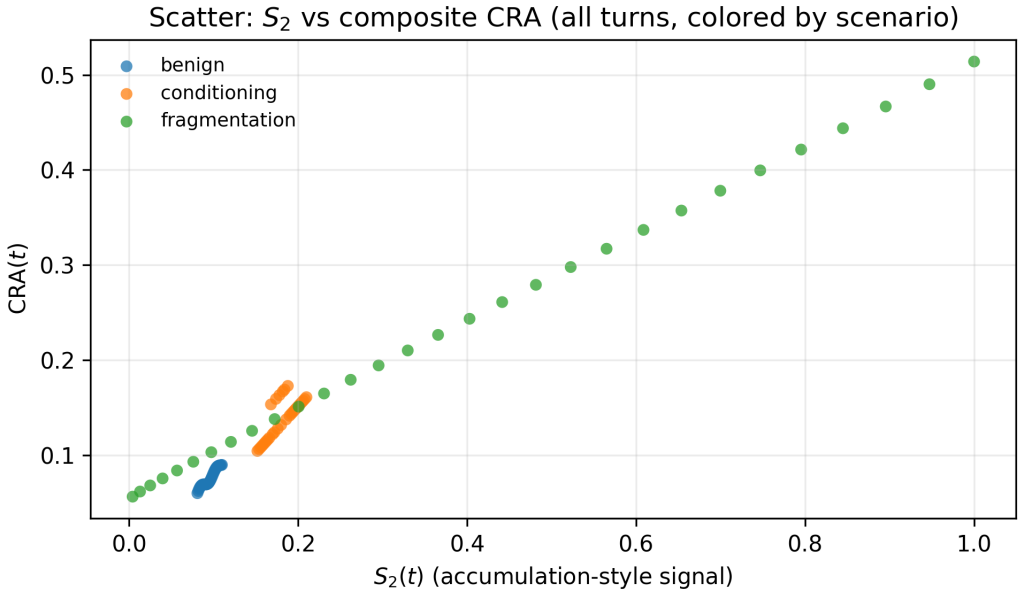

Figure 9. Scatter of $S_2(t)$ vs. CRA(t) with points colored by scenario (all turns pooled). The strong linear relationship in the fragmentation scenario reflects the dominant weight $\beta = 0.45$ on the IAG signal.

8. Architectural Composability and Deployment Patterns

A critical design requirement for the CRA Framework is that it must not require organizations to replace existing guardrail infrastructure. The framework is designed as an additive session layer that receives the same turn-level inputs as existing systems and produces an independent, composable output. Figure 10 shows the reference deployment pattern.

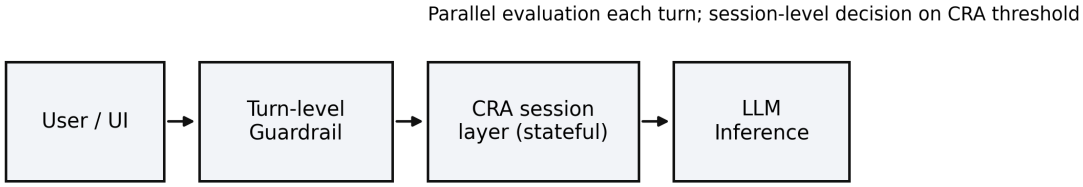

Figure 10. Reference deployment pattern: the CRA Framework operates as a stateful session layer alongside existing turn-level controls. Both layers receive the same (u_i, r_i) input; their outputs are independent. Session-level intervention is triggered only by the CRA Score.

8.1. Enterprise RAG Deployment

In enterprise RAG systems, the IAG has a natural extension: entities in the knowledge base can be pre-annotated with sensitivity tiers (public, internal, confidential, restricted), and IAG node weights can be initialized from these tiers rather than inferred from response content alone. For RAG systems grounded in proprietary corpora, the semantic drift monitor can be anchored to the embedding centroid of the retrieved context rather than the user's declared intent, flagging when the conversation has drifted away from retrieval scope. This design integrates naturally with cost-aware routing frameworks [**?**], where routing decisions already encode implicit policy constraints.

 doi:10.20944/preprints202604.1595.v1

8.2. Agentic Pipeline Deployment

Agentic LLM systems introduce an additional CRA surface that does not present in conversational systems: tool calls. An agent that executes actions, queries, API calls, and file operations, can accumulate operational risk through a sequence of individually authorized actions whose combined effect was never authorized. The CRA Framework extends to agentic settings by treating tool call outcomes as turns in the session history, with IAG updates incorporating the data accessed or modified by each tool invocation. The compliance gradient detector, in this context, tracks not just linguistic refusal behavior but the agent's rate of escalating resource access.

8.3. Educational AI Deployment

Educational AI systems present a distinct deployment pattern where the primary CRA risk is a pedagogical scope violation rather than an adversarial attack. A student using an AI tutoring system may, through a sequence of help requests, effectively outsource not just assistance but the entire cognitive work that the assignment was designed to produce. The semantic drift monitor, anchored to the declared learning objective, quantifies how far the session has drifted from supported scaffolding toward unsupported substitution. This is not a security guardrail but a pedagogical integrity guardrail, a new application class that the CRA Framework enables.

9. Open Problems and Research Agenda

9.1. IAG Maintenance at Scale

The Information Accumulation Graph is the most theoretically grounded but computationally demanding component of the framework. In sessions with the unbounded entity growth, common in long-horizon agentic tasks, maintaining a full IAG becomes intractable. The sliding-window approximation trades completeness for tractability but introduces gaps: an entity disclosed at turn 5 and referenced again at turn 45 may fall outside the window, producing an underestimated S_2. Developing efficient approximate IAG data structures with bounded approximation error is an open problem.

9.2. Threshold Calibration Without Ground Truth

Setting θ requires characterizing the CRA Score distribution under benign and adversarial session distributions. Benign session data is abundant in enterprise deployments; adversarial session data is not, because successful CRA attacks are rarely labeled as such in production logs. This creates a classic anomaly detection problem with highly imbalanced training data. Developing calibration methods robust to adversarial distribution shift is an open and important problem.

9.3. Explainable Intervention

When the CRA Score triggers a session-level intervention, the system must communicate why without revealing detection logic that would aid adversarial adaptation. The *decision certificate* concept, a structured artifact narrating the guardrail's reasoning in human-readable but policy-protected form is a promising direction. Formalizing what should and should not be disclosed is both a technical and a governance problem.

9.4. Adversarial Robustness of Sub-Signals

A sophisticated adversary with knowledge of the CRA Framework may attempt to suppress individual sub-signal elevations. Evading S_1 requires maintaining semantic proximity to declared intent while pursuing a harmful trajectory. Evading S_2 requires fragmenting disclosures across entities rather than accumulating attributes on a single target. Evading S_3 requires maintaining refusal-triggering content in otherwise compliant responses. Analyzing the theoretical hardness of simultaneously evading all three signals, and whether composite fusion raises the bar meaningfully, is a formal security question with practical implications.

 doi:10.20944/preprints202604.1595.v1

10. Discussion and Limitations

The illustrative results presented in Section 7 demonstrate that the CRA Framework produces qualitatively distinct, theoretically consistent signal patterns across different attack scenarios. Several important caveats govern the interpretation of these results.

First, all signals in the current implementation are heuristic approximations. S_2 in the reference implementation uses token-overlap proxies rather than a fully calibrated sensitivity taxonomy; S_3 uses keyword-based refusal detection rather than a trained classifier. These choices are deliberate: they enable reproducibility and domain-agnostic deployment without requiring fine-tuned auxiliary models. However, they also mean that the reported signal values should not be interpreted as calibrated risk probabilities.

Second, the synthetic trajectories are designed to exhibit the signal patterns predicted by theory. Real sessions will be noisier and more heterogeneous; the benign/adversarial separation demonstrated here is an upper bound on what practitioners should expect from raw signals without domain calibration.

Third, the framework identifies when to act but does not specify what action to take. The intervention policy, session termination, mandatory re-declaration of intent, elevation to human review, context pruning, or safe-mode switching, is a deployment decision that depends on the risk tolerance, regulatory context, and user experience requirements of each application.

11. Conclusions

The assumption that safety is a property of individual exchanges is not a deliberate design choice in most guardrail systems, it is an inherited simplification that has persisted because single-turn evaluation is tractable and multi-turn safety is not. As LLM deployments mature toward long-horizon conversational agents, agentic pipelines with tool use, and enterprise systems handling sensitive data across extended sessions, this simplification becomes a structural liability.

We have formalized Conversational Risk Accumulation as the class of safety failures undetectable by stateless guardrail systems and proposed the CRA Framework as a practical, composable, and theoretically grounded response. The framework's three components, Semantic Drift Monitor, Information Accumulation Graph, and Compliance Gradient Detector, address distinct accumulation mechanisms, and are designed to complement rather than replace existing single-turn infrastructure.

The open problems identified in Section 9, IAG efficiency, threshold calibration under adversarial distribution shift, explainable intervention, and evasion hardness, constitute a research agenda for multi-turn safety that is both tractable and consequential. The field has invested heavily in making individual LLM responses safer. The next necessary investment is in making conversations safe.

Author Contributions: Conceptualization, S.M. and G.R.N.; formal analysis, S.M.; writing, original draft preparation, S.M.; writing, review and editing, S.M. and G.R.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Figure-generation scripts and the synthetic CRA timeseries (`cra_timeseries.csv`) are available at https://github.com/sanmish4ds/SQLclMCP. Raw production logs, if any, are retained under organizational policy and are not published.

Conflicts of Interest: The authors declare no conflicts of interest.

Institutional Review Board Statement: Not applicable: this manuscript describes a technical framework and illustrative synthetic traces without human subjects data.

 doi:10.20944/preprints202604.1595.v1

References

1. Röttger, P.; Kirk, H.R.; Vidgen, B.; Attanasio, G.; Bianchi, F.; Hovy, D. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. In Proceedings of the Proc. ICML, 2024. arXiv:2402.04249.
2. Shen, X.; Chen, Z.; Backes, M.; Shen, Y.; Zhang, Y. "Do Anything Now": Characterizing and Evaluating In-the-Wild Jailbreak Prompts on Large Language Models. arXiv:2308.03825, 2023.
3. Zou, A.; Wang, Z.; Kolter, J.Z.; Fredrikson, M. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv:2307.15043, 2023.
4. Wei, A.; Haghtalab, N.; Steinhardt, J. Jailbroken: How Does LLM Safety Training Fail? In Proceedings of the Proc. NeurIPS, 2024.
5. Perez, F.; Ribeiro, I. Ignore Previous Prompt: Attack Techniques for Language Models. arXiv:2211.09527, 2022.
6. Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; Fritz, M. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In Proceedings of the Proc. ACM AISec, 2023.
7. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical Report AI 100-1, NIST, 2023.
8. Anthropic. Claude's Model Specification. Technical report, Anthropic, 2024.
9. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training Language Models to Follow Instructions with Human Feedback. arXiv:2203.02155, 2022.
10. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. In Proceedings of the Proc. TCC, 2006, pp. 265–284.