Certified Multi-Turn Robustness for LLM Safety via Compositional Bounds and Safety Persistence

Yang Liu[1] **Bin Chong**[1,*] **Wenkai Yang**[2] **Shuai Zhang**[3]
Yancheng Chen[4] **Feiyu Han**[4] **GuoZhen**[5] **Cheng Zhang**[5]
Huaibing Xie[5] **Changze Lv**[5] **Shihan Dou**[5] **Pluto Zhou**[5]
[1]Peking University [2]Renmin University of China [3]Tsinghua University
[4]University of Chinese Academy of Sciences [5]Tencent Hunyuan
[*]Corresponding author: chongbin@pku.edu.cn

Abstract

Large language models (LLMs) are vulnerable to multi-turn jailbreak attacks that progressively manipulate conversation context. Existing certified robustness methods are limited to single-turn inputs; naive multi-turn composition yields bounds that degrade exponentially in the number of turns. We introduce **Multi-Turn Certified Robustness (MTCR)**, a framework that models conversational safety via State-Adversarial MDPs and defines *k-turn certified robustness* as the worst-case safety probability across k adversarial turns. MTCR comprises: (i) compositional certification via embedding-space mode decomposition, yielding tighter certified lower bounds than naive multiplication; (ii) (α, β)-safety persistence, improving the degradation rate from $\underline{p}^k$ to β^k (with $\beta > \underline{p}$) and yielding interpretable horizon estimates; (iii) matching information-theoretic upper bounds establishing tightness; and (iv) a unified algorithm combining these results. Experiments on six LLMs under ϵ-bounded and Crescendo-style attacks confirm that empirical safety consistently exceeds the certified bounds.

1 Introduction

The deployment of large language models (LLMs) in conversational applications has raised safety concerns, particularly regarding *jailbreak attacks*: carefully crafted inputs designed to bypass safety mechanisms and elicit harmful content (Zou et al., 2023; Wei et al., 2023). Although substantial progress has been made in defending against single-turn attacks, **multi-turn jailbreak attacks** that progressively manipulate conversation context across multiple interactions pose a distinct challenge (Russinovich et al., 2025; Li et al., 2024).

Recent empirical studies reveal the severity of this threat. The Crescendo attack (Russinovich et al., 2025) achieves near-perfect success rates against production LLMs using fewer than 10 turns; X-Teaming (Rahman et al., 2025) attains 96–98% success rates against state-of-the-art models; human red-teamers consistently exceed 70% success against deployed defenses (Li et al., 2024). These findings indicate that *defenses robust against single-turn attacks offer little protection in multi-turn settings*.

On the theoretical side, existing certified robustness methods (randomized smoothing (Robey et al., 2023), erase-and-check (Kumar et al., 2023), knapsack-based certification (Chen et al., 2025)) are limited to single-input settings. A naive extension yields certified lower bounds that degrade exponentially in the number of turns: if each turn has worst-case certified safety probability p, the k-turn bound under multiplicative composition is merely p^k, which becomes negligible for moderate k. This degradation arises because adaptive adversaries observe responses and craft subsequent inputs (a sequential game), accumulated context compounds adversarial influence, and autoregressive generation amplifies early-token perturbations.

Contributions. We develop **Multi-Turn Certified Robustness (MTCR)**, a certification framework grounded in SA-MDP theory, with three contributions:

(1) We model multi-turn conversations as State-Adversarial MDPs and define k-turn certified robustness. We develop compositional certification via embedding-space mode decomposition: certifying intra-mode safety and inter-mode transitions separately yields bounds tighter than naive multiplication (Corollary 3.2). A unified algorithm (Algorithm 1) combines compositional and persistence bounds to compute the certified guarantee in practice.

(2) We formalize (α, β)-safety persistence, a structural property under which the robustness bound improves from $\underline{p}^k$ to β^k (when $\beta > \underline{p}$), with a linear characterization $1 - k(1-\beta)$ that yields interpretable horizon estimates (Theorem 3.5). We prove information-theoretic upper bounds showing our compositional approach is tight for non-overlapping decompositions, and impossibility results for systems that lack structural assumptions.

(3) We evaluate MTCR on six production LLMs. The certified bounds provide formal guarantees against ϵ-bounded adversaries; empirical safety consistently exceeds the bound under both static (ϵ-ball) and Crescendo-style adaptive attacks, confirming that the bounds are not violated in practice.

2 Problem Formulation: Multi-Turn Safety as SA-MDP

We model multi-turn conversations as State-Adversarial MDPs (SA-MDPs). The dialogue history $h_t = (u_1, r_1, \ldots, u_t, r_t)$ maps to a conversational state $s_t = \phi(h_t) \in \mathcal{S} \subseteq \mathbb{R}^d$ via an embedding ϕ (e.g., LLM hidden representations). An adversary selects user messages u_t from an ϵ-perturbation ball $\mathcal{B}_\epsilon$ around a reference input; the LLM responds according to policy $\pi(\cdot|s, u)$. States evolve as $s_t = T(s_{t-1}, u_t, r_t)$. A safety predicate $\mathsf{Safe} : \mathcal{S} \to \{0, 1\}$ partitions the state space into a safe region $\mathcal{S}_+$ and an unsafe region $\mathcal{S}_-$, with safety margin $\Delta(s) = \mathrm{dist}(s, \mathcal{S}_-)$. We assume $\mathcal{S}$ is bounded, $\mathcal{S}_+$ is open (so $\Delta(s) > 0$ for all $s \in \mathcal{S}_+$), and $\mathcal{S}_+$ has non-empty interior with $\delta_{\min}$ denoting the maximum inscribed ball radius. Full formal definitions appear in Appendix B.

Definition 1 (k-Turn Certified Robustness)**.** Given initial state $s_0 \in \mathcal{S}_+$, LLM policy π, perturbation budget ϵ, and horizon $k \in \mathbb{N}$, the k-turn certified robustness $\rho_k(s_0, \pi, \epsilon)$ is:

$$\inf_{\nu \in \Pi_\epsilon} \mathbb{P}_{\substack{u_t \sim \nu(\cdot|s_{t-1}) \\ r_t \sim \pi(\cdot|s_{t-1}, u_t)}} \left[\bigwedge_{t=1}^{k} \mathsf{Safe}(s_t) \,\middle|\, s_0 \right], \quad (1)$$

where states evolve according to $s_t = T(s_{t-1}, u_t, r_t)$.

The quantity $\rho_k(s_0, \pi, \epsilon)$ captures the worst-case probability of maintaining safety across k turns against any adaptive adversary (one that observes s_{t-1} before choosing u_t). A natural certification strategy is to compose single-turn safety bounds multiplicatively. For each state s, let $p(s, \epsilon) = \inf_{u \in \mathcal{B}_\epsilon} \mathbb{P}[\mathsf{Safe}(T(s, u, r))]$ denote the certified probability of remaining safe after one adversarial turn. This composition yields the following bound.

Proposition 2.1 (Naive Multiplicative Bound)**.** *For any $s_0 \in \mathcal{S}_+$:*

$$\rho_k(s_0, \pi, \epsilon) \geq \left(\inf_{s \in \mathcal{S}_+} p(s, \epsilon) \right)^k = \underline{p}^k, \quad (2)$$

where $\underline{p} = \inf_{s \in \mathcal{S}_+} p(s, \epsilon)$ is the worst-case per-turn certified safety. If $\underline{p} < 1$, this bound vanishes exponentially in k.

Example 2.2 (Vacuity of Naive Bound)**.** With per-turn safety $\underline{p} = 0.95$: $\rho_{10} \geq 0.60$, $\rho_{20} \geq 0.36$, $\rho_{50} \geq 0.08$. For safety-critical applications requiring $\rho_k \geq 0.9$, this permits at most $k \leq 2$ turns. □

Scope of certification. The certified guarantee applies to adversaries constrained within an ϵ-ball (e.g., character-level edit distance) around reference inputs. Real attacks such as Crescendo operate at the semantic level; while our experiments report empirical safety against such attacks, the *formal certification* covers only ϵ-bounded perturbations. The SA-MDP framework is agnostic to the perturbation metric and extends to richer threat models once compatible per-turn certification oracles become available.

Without additional structure, the multiplicative composition provides little practical guarantee for conversations beyond a few turns. This motivates identifying structural conditions under which tighter bounds can be achieved.

3 Multi-Turn Certified Robustness

We extend single-turn certification to multi-turn conversations by exploiting mode structure in embedding space and safety margin evolution. The framework has four parts: compositional bounds via mode decomposition (§3.1), safety persistence characterizing margin evolution (§3.2), information-theoretic upper bounds establishing tightness (§3.3), and a unified algorithm (§3.4).

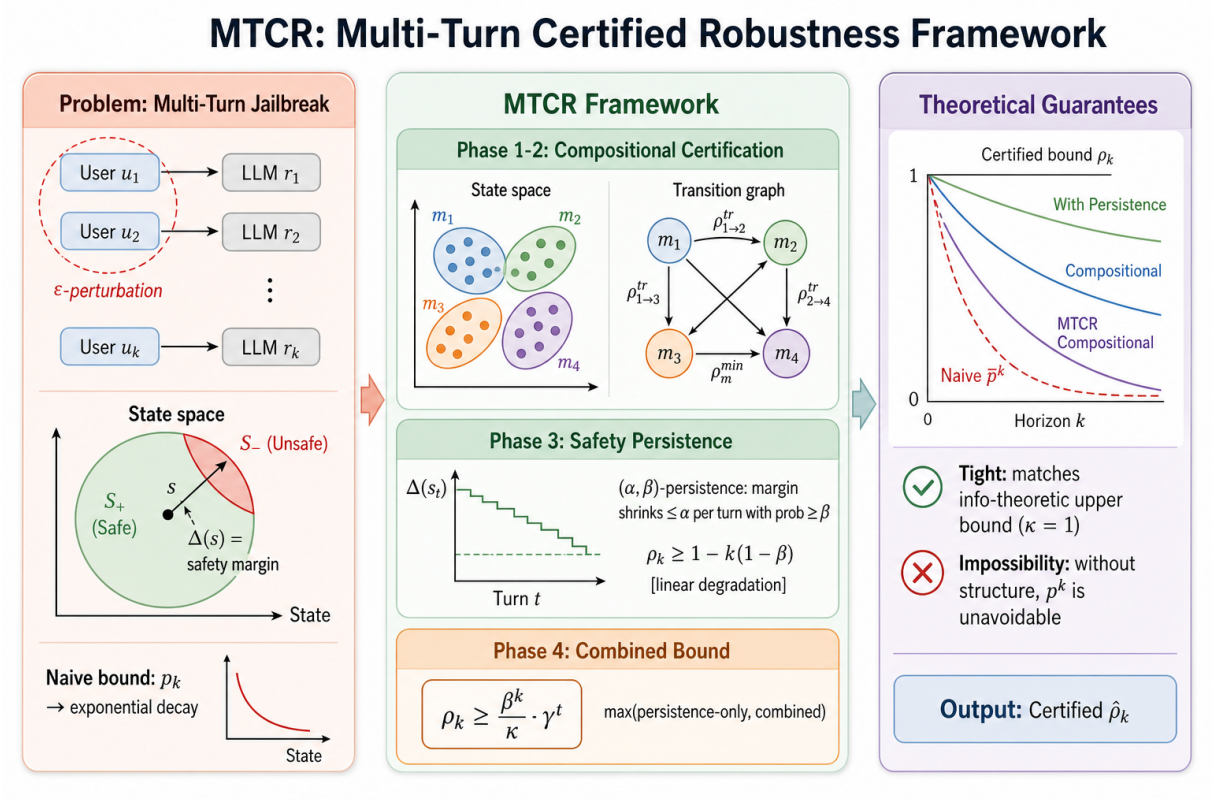


Figure 1: Overview of MTCR. **Left**: Multi-turn conversations modeled as SA-MDPs; naive bound $\underline{p}^k$ degrades exponentially. **Center**: MTCR combines compositional certification (intra-mode + transition safety) with (α, β)-persistence into a unified bound. **Right**: Matching upper bounds establish tightness (κ=1). Experiments on six LLMs confirm empirical safety exceeds the certified bounds.

3.1 Compositional Certification

Conversations exhibit *mode structure* in embedding space that can be exploited for tighter certification. A mode decomposition $\mathcal{M} = \{m_1, \ldots, m_M\}$ covers $\mathcal{S}_+$ with bounded overlap $\kappa = \max_s |\{i : s \in \mathcal{S}_{m_i}\}|$. For each mode m we define intra-mode certified safety $\rho_m^{\text{in}}(\epsilon)$ (worst-case probability of staying safe and in m) and for edges (m_i, m_j) in the transition graph we define transition safety $\rho_{i\to j}^{\text{tr}}(\epsilon)$. A mode trajectory σ over k turns has $n_j(\sigma)$ intra-mode turns in m_j and $\tau(\sigma)$ transitions. Full definitions appear in Appendix B.

Theorem 3.1 (Compositional Certification Bound)**.** *Let $\mathcal{M} = \{m_1, \ldots, m_M\}$ be a mode decomposition with overlap $\kappa = \kappa(\mathcal{M})$. For any k-turn conversation with mode trajectory $\sigma \in \Sigma_k(\mathcal{M})$:*

$$\rho_k(s_0, \pi, \epsilon) \geq \frac{1}{\kappa} \cdot \prod_{j=1}^{M} \left(\rho_{m_j}^{\text{in}}(\epsilon)\right)^{n_j(\sigma)} \cdot \prod_{(i,j)\in \text{Trans}(\sigma)} \rho_{i\to j}^{\text{tr}}(\epsilon)$$

where $\text{Trans}(\sigma) = \{(\sigma(t), \sigma(t+1)) : \sigma(t) \neq \sigma(t+1)\}$ is the multiset of transitions. Taking the infimum over feasible trajectories:

$$\rho_k(s_0, \pi, \epsilon) \geq \frac{1}{\kappa} \cdot \inf_{\sigma \in \Sigma_k(\mathcal{M}, s_0)} \left[\prod_{j=1}^{M} \left(\rho_{m_j}^{\text{in}}\right)^{n_j(\sigma)} \cdot \prod_{(i,j)\in \text{Trans}(\sigma)} \rho_{i\to j}^{\text{tr}} \right]$$

where $\Sigma_k(\mathcal{M}, s_0) \subseteq \Sigma_k(\mathcal{M})$ are trajectories starting from a mode containing s_0.

The following corollary quantifies when the compositional bound strictly improves upon the naive multiplicative bound.

Corollary 3.2 (Improvement over Naive Composition)**.** *Let $\underline{p} = \inf_s p(s, \epsilon)$ be the naive per-turn bound. Suppose the mode decomposition satisfies:*

1. *Intra-mode safety: $\rho_{m_j}^{\text{in}} \geq 1 - \delta$ for small $\delta > 0$*
2. *Inter-mode transition safety: $\rho_{i\to j}^{\text{tr}} \geq \gamma$ for some $\gamma \in (0, 1)$*

3. *Sparse transitions: trajectory σ has at most τ transitions*

Then the compositional bound is:

$$\rho_k^{\text{comp}} \geq \frac{1}{\kappa}(1-\delta)^{k-\tau} \cdot \gamma^{\tau} \tag{3}$$

This exceeds the naive bound $\underline{p}^k$ when:

$$\tau < \frac{k\log(1-\delta) - k\log\underline{p} - \log\kappa}{\log(1-\delta) - \log\gamma} \tag{4}$$

Example 3.3 (Numerical Comparison)**.** Under parameters k=20, τ=3, $\underline{p}$=0.9, δ=0.02, γ=0.7, κ=1:

$$\rho_{20}^{\text{naive}} \geq 0.9^{20} \approx 0.122$$
$$\rho_{20}^{\text{comp}} \geq 0.98^{17} \cdot 0.7^3 \approx 0.708 \cdot 0.343 \approx 0.243$$

The compositional certified lower bound is approximately **twice** the naive lower bound. Note: improvement depends on k and τ; when k is small, transition overhead may dominate (see Appendix I.2). □

Computing the compositional bound requires finding the worst-case trajectory over the mode transition graph. This can be done efficiently, as the following proposition shows.

Proposition 3.4 (Worst-Case Trajectory Computation)**.** *The trajectory infimum in Theorem 3.1 can be computed by dynamic programming in $O(M^2k)$ time.*

3.2 Safety Persistence

The compositional bound still depends on the number of transitions τ. We next define *safety persistence*, a property that enables stronger guarantees by characterizing how safety margins evolve.

Definition 2 ((α,β)-Safety Persistence)**.** A conversational system $(\mathcal{S}, T, \pi, \mathsf{Safe})$ exhibits (α,β)-**safety persistence** for $\alpha \in [0,1)$ and $\beta \in (0,1]$ if for all safe states $s \in \mathcal{S}_+$ with safety margin $\Delta(s) > 0$:

$$\mathbb{P}_{u\sim\nu,\, r\sim\pi}\left[\begin{array}{c}\mathsf{Safe}(s')=1 \\ \wedge\ \Delta(s') \geq (1-\alpha)\Delta(s)\end{array}\right] \geq \beta \tag{5}$$

where $s' = T(s,u,r)$ and $\nu \in \Pi_\epsilon$ is any adversarial policy.

With probability at least β, the safety margin retains at least a $(1-\alpha)$ fraction of its previous value per turn; well-aligned LLMs typically satisfy this property. Under safety persistence, we obtain the following guarantee.

Theorem 3.5 (Degradation under Safety Persistence)**.** *Suppose the system exhibits (α,β)-safety persistence. Let $s_0 \in \mathcal{S}_+$ have initial margin $\Delta_0 = \Delta(s_0) > 0$. Then:*

$$\rho_k(s_0,\pi,\epsilon) \geq \beta^k \tag{6}$$

When $\beta > \underline{p}$ (i.e., the persistence guarantee exceeds the worst-case per-turn bound), we have $\beta^k > \underline{p}^k$, strictly improving upon the naive bound.

Additionally, a union bound yields an interpretable (though numerically weaker) characterization:

$$\rho_k(s_0,\pi,\epsilon) \geq 1 - k(1-\beta) \tag{7}$$

This formula provides a sufficient condition for safety: the system remains safe with probability $\geq 1-\xi$ whenever $k \leq \xi/(1-\beta)$, enabling direct horizon estimation.

Corollary 3.6 (Effective Horizon Estimation)**.** *Under (α,β)-safety persistence, to maintain $\rho_k \geq 1-\xi$ for target $\xi \in (0,1)$, the linear characterization gives the sufficient horizon bound:*

$$k_{\max} = \left\lfloor \frac{\xi}{1-\beta} \right\rfloor \tag{8}$$

compared to $k_{\max} = \lfloor \log(1-\xi)/\log\underline{p} \rfloor$ under naive composition. For example, with β=0.98 and ξ=0.1: the persistence-based estimate gives $k_{\max} = 5$, whereas $\underline{p}$=0.9 gives $k_{\max} = 1$.

Safety persistence can be verified from Lipschitz continuity of the transition dynamics and concentration properties of the LLM output; the following proposition gives sufficient conditions.

Proposition 3.7 (Persistence from Lipschitz Continuity)**.** *Suppose:*

1. *T is L_s-Lipschitz in state: $\|T(s,u,r) - T(s',u,r)\| \leq L_s\|s-s'\|$*
2. *T is L_u-Lipschitz in input: $\|T(s,u,r) - T(s,u',r)\| \leq L_u \cdot d(u,u')$*
3. *T is L_r-Lipschitz in response: $\|T(s,u,r) - T(s,u,r')\| \leq L_r \cdot d_r(r,r')$ for some metric d_r*
4. *The safe region satisfies: $\mathcal{S}_+$ is convex with $\text{dist}(s,\mathcal{S}_-) \geq \delta$ for s in the δ-interior*
5. *LLM responses satisfy: $\mathbb{P}[\|r - r^*\| \leq \eta] \geq p_0$ for some nominal response r^**

6. *Nominal drift is bounded:* $\|T(s, u^*, r^*) - s\| \leq \delta_T$ *for all* $s \in \mathcal{S}_+$

Then for all states s *in the* δ*-interior of* $\mathcal{S}_+$ *(i.e.,* $\Delta(s) \geq \delta$*), the* (α, β)*-persistence condition (Definition 2) holds with:*

$$\alpha = \frac{L_u\epsilon + L_r\eta + \delta_T}{\delta}, \qquad \beta = p_0 \cdot \mathbf{1}[L_u\epsilon + L_r\eta + \delta_T < \delta]. \tag{9}$$

The indicator function reflects a strict sufficient condition; when $L_u\epsilon + L_r\eta + \delta_T \geq \delta$*, it yields* $\beta = 0$*, making the bound vacuous. This occurs in many practical settings where the perturbation budget and nominal drift are large relative to the safety margin. Proposition 3.7 serves as a theoretical grounding showing* when *persistence holds from first principles; in our experiments, persistence parameters are instead estimated empirically via sampling (Algorithm 1 Phase 1; see Appendix F.1).*

3.3 Tightness and Impossibility Results

We establish that our compositional bounds are tight. The following theorem gives an information-theoretic upper bound via a matching construction, showing that the compositional product cannot be improved for non-overlapping decompositions.

Theorem 3.8 (Information-Theoretic Upper Bound on Certification)**.** *For any mode decomposition* $\mathcal{M}$ *and any trajectory* σ*, there exist transition dynamics* T *and adversarial strategies* ν^* *such that:*

$$\rho_k(s_0, \pi, \epsilon) \leq \prod_{j=1}^{M} \left(\rho^{\text{in}}_{m_j}\right)^{n_j(\sigma)} \cdot \prod_{(i,j)\in \text{Trans}(\sigma)} \rho^{\text{tr}}_{i\to j}$$

The compositional product is also an upper bound for a worst-case system matching the mode-level safety parameters.

Corollary 3.9 (Near-Optimality of Compositional Bounds)**.** *For non-overlapping mode decompositions (*$\kappa = 1$*), our compositional lower bound (Theorem 3.1) matches the information-theoretic upper bound (Theorem 3.8) exactly. The compositional certification is therefore tight for* $\kappa = 1$*.*

These positive results rely on structural assumptions. Without such assumptions, exponential degradation is unavoidable, as the following impossibility result shows.

Theorem 3.10 (Impossibility of Sub-Exponential Bounds in General)**.** *Without structural assumptions (mode decomposition or safety persistence), there exist conversational systems where:*

$$\rho_k(s_0, \pi, \epsilon) = p^k \tag{10}$$

for some $p < 1$*, and the naive multiplicative bound* p^k *is tight.*

The proof constructs a memoryless system where the transition function T maps every (s, u, r) to a fresh state drawn independently from a fixed distribution, with $\mathbb{P}[\mathsf{Safe}(s') = 1] = p$ regardless of the adversary's choice of u (see Appendix E.8). Since turns are independent and each has safety probability exactly p, we get $\rho_k = p^k$. Thus our structural assumptions (modes, persistence) are necessary to *guarantee* sub-exponential bounds, not merely sufficient for achieving them.

3.4 Unified Framework and Algorithm

We synthesize the above results into a unified certification framework. When the system admits both a mode decomposition and safety persistence, we obtain the following combined bound.

Theorem 3.11 (Combined Bound)**.** *Suppose the system has mode decomposition* $\mathcal{M}$ *with overlap* κ *and exhibits* (α, β)*-safety persistence within each mode. Then:*

$$\rho_k(s_0, \pi, \epsilon) \geq \frac{\beta^k}{\kappa} \cdot \inf_{\sigma\in\Sigma_k(\mathcal{M}, s_0)} \left[\prod_j \left(\frac{\rho^{\text{in}}_{m_j}}{\beta}\right)^{n_j(\sigma)} \cdot \prod_{(i,j)} \rho^{\text{tr}}_{i\to j}\right]$$

When $\rho^{\text{in}}_{m_j} \geq \beta$ *for all modes (persistence dominates intra-mode certification):*

$$\rho_k \geq \frac{\beta^k}{\kappa} \cdot \gamma^\tau \tag{11}$$

where $\gamma = \min_{i,j} \rho^{\text{tr}}_{i\to j}$ *and* τ *is the number of transitions.*

Algorithm 1 formalizes the four-phase procedure: certify intra-mode and inter-mode safety, compute the worst-case trajectory via dynamic programming, and combine with persistence parameters. Phase 3 uses $\hat{\rho}^{\text{in}}_{m_j}/\hat{\beta}$ (not $\hat{\rho}^{\text{in}}_{m_j}$) for intra-mode steps, consistent with Theorem 3.11, since the persistence baseline is already captured by $\hat{\beta}^k$ in Phase 4. The final output is $\max(\hat{\beta}^k, \hat{\rho}^{\text{comb}}_k)$,

taking the better of the persistence-only and combined bounds. In practice, the combined formula dominates because it exploits both mode structure and persistence jointly (Table 3).

Let N be the number of samples for randomized smoothing. Phase 1 requires $O(MN)$ LLM forward passes; Phase 2 requires $O(|E_{\mathcal{M}}|N)$ passes; Phase 3 performs $O(M^2k)$ arithmetic operations. The total complexity is $O((M+|E_{\mathcal{M}}|)N+M^2k)$, which is linear in the horizon k.

4 Experiments

We evaluate MTCR on production LLMs under real attack scenarios. Appendix I provides additional numerical verification under a controlled parametric model.

4.1 Experimental Setup

Compared methods. No prior method provides certified multi-turn robustness. We compare four variants: **Mult. ref.**, the multiplicative reference $\bar{p}^k$ with $\bar{p} = \sup_s p(s, \epsilon)$ (an optimistic reference, not a valid certified bound under adaptive adversaries); **Persistence-Only**, applying the (α, β)-persistence bound alone; **Compositional-Only**, using mode decomposition without persistence; and **MTCR (full)**, our combined certification (Algorithm 1).

Data. We use harmful prompts from AdvBench (Zou et al., 2023), covering violence, illegal activities, and hate speech (50 prompts per category, 150 total).

Mode discovery. Dialogue state embeddings are computed on n=500 held-out safe multi-turn conversations (ShareGPT-style) using sentence-transformers. Each state is the embedding of the full dialogue history. Modes are discovered via k-means clustering in embedding space, with cluster radii set at the 90th percentile of within-cluster distances. This is geometric clustering; we do not claim these clusters correspond to human-interpretable semantic categories, though they capture safety-relevant structure as evidenced by the high intra-mode safety rates.

Evaluation protocol. Each attack trial runs k turns; empirical safety is the fraction of trials in which all turns remain safe. We report results over 100 trials per configuration; 95% Clopper–Pearson confidence intervals are $\leq \pm 0.05$ across all settings and omitted from tables for clarity. A concrete example of the data format, perturbation procedure, and Crescendo-style attack progression appears in Appendix H.

4.2 Experiments on Production Models

We evaluate six LLMs with varying alignment strengths. **Open-source:** LLaMA-2-7B-Chat (Touvron et al., 2023), Vicuna-7B (Chiang et al., 2023), Llama-3.2-3B, and Qwen2.5-7B-Instruct (Yang et al., 2025). **Closed-source (via API):** GPT-4o (Hurst et al., 2024) and Claude-3.5-Sonnet (Anthropic, 2024).

The evaluation covers horizons $k \in \{5, 10, 15, 20\}$, two attack types (static ϵ-ball and Crescendo-style), and mode granularities $M \in \{2, 4, 8\}$. Dialogue state embeddings use `all-MiniLM-L6-v2`. Safety is judged by a harmful-content detector combining refusal patterns and an unsafe-keyword list; we also compare with neural classifiers (Appendix G). Perturbation budget is ϵ=5 (character-level edit distance, following SmoothLLM (Robey et al., 2023)), with N=100 smoothing samples per mode and $\epsilon \in \{3, 5, 7\}$ for sensitivity analysis. Full details are in Appendix J.2.

We highlight three findings from Table 1.

(1) Bound validity. Empirical safety under the strongest attack (Crescendo) exceeds the MTCR certified bound across all models and horizons, with gaps from +0.19 (LLaMA-2, k=5) to +0.30 (GPT-4o, k=10). This confirms the certified guarantee is not violated in practice. Comparing MTCR with the naive bound $\underline{p}^k$ (which is also a valid certified bound), MTCR provides 1.2–1.6× tighter certification at k=5, demonstrating the value of compositional structure.

(2) Model ranking and alignment quality. Certified bounds reflect alignment strength: at k=5, GPT-4o achieves the highest certified bound ($\hat{\rho}_5$=0.55), followed by Claude-3.5-Sonnet (0.51), Qwen2.5 (0.44), Llama-3.2 (0.40), LLaMA-2 (0.36), and Vicuna (0.28). Closed-source models consistently outperform open-source ones, suggesting stronger underlying safety alignment.

(3) Degradation with horizon. Examining the ratio $\hat{\rho}_{10}/\hat{\rho}_5$ across models reveals that better-aligned models degrade more slowly: GPT-4o retains 58% of its k=5 bound at k=10, compared to only 29% for Vicuna. This supports Theorem 3.5: systems with stronger safety persistence exhibit slower degradation. For LLaMA-2, the full $k \in \{5, 10, 15, 20\}$ trajectory shows progressive decay from 0.36 to 0.02, with the combined for-

Table 1: Certification and empirical safety across six LLMs (M=4, N=100, ϵ=5, 100 trials). MTCR $\hat{\rho}_k$: certified lower bound; Naive $\underline{p}^k$: valid but loose multiplicative bound; Mult. ref. $\bar{p}^k$: optimistic reference (*not* a valid certified bound under adaptive adversaries); Gap = Emp. (Crescendo) − MTCR $\hat{\rho}_k$. **Bold**: best certified bound per horizon.

Model	k	Certification			Empirical Safety		Gap
		MTCR $\hat{\rho}_k$	Naive $\underline{p}^k$	Mult. ref.	Static	Crescendo	
Open-source models							
LLaMA-2-7B-Chat	5	0.36	0.29	0.59	66%	55%	+0.19
	10	0.13	0.08	0.35	49%	38%	+0.25
	15	0.05	0.02	0.21	38%	28%	+0.23
	20	0.02	0.01	0.12	30%	22%	+0.20
Vicuna-7B	5	0.28	0.22	0.50	58%	46%	+0.18
	10	0.08	0.05	0.25	40%	30%	+0.22
Llama-3.2-3B	5	0.40	0.33	0.66	72%	62%	+0.22
	10	0.18	0.11	0.43	56%	45%	+0.27
Qwen2.5-7B-Instruct	5	0.44	0.35	0.70	76%	65%	+0.21
	10	0.22	0.12	0.48	60%	49%	+0.27
Closed-source models (API)							
GPT-4o	5	**0.55**	0.44	0.82	88%	78%	+0.23
	10	**0.32**	0.20	0.66	73%	62%	+0.30
Claude-3.5-Sonnet	5	0.51	0.42	0.77	85%	74%	+0.23
	10	0.28	0.18	0.60	70%	58%	+0.30

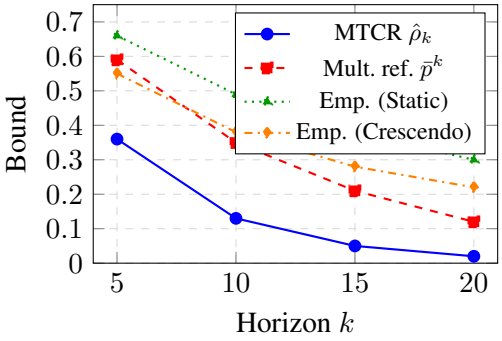


Figure 2: Certified lower bound vs. horizon k on LLaMA-2-7B-Chat ($\bar{p}$=0.90, $\underline{p}$=0.78; data from Table 1). The MTCR curve lies below empirical safety under both attack types at all horizons, confirming the certified bound is not violated.

mula (Theorem 3.11) providing the tightest bound at all horizons (Table 3). Crescendo reduces empirical safety by 8–13 percentage points relative to static attacks; stronger models show larger absolute but smaller relative drops. Figure 2 visualizes this degradation.

We vary mode granularity $M \in \{2, 4, 8\}$ on LLaMA-2-7B-Chat at k=5. Coarser modes (M=2) yield $\hat{\rho}_5$=0.39; finer modes (M=8) yield 0.25 due to increased overlap κ and transition cost. Although M=2 gives a marginally higher bound at k=5, it provides less diagnostic value (fewer modes to identify safety bottlenecks) and lower

Table 2: Perturbation budget ϵ sensitivity (LLaMA-2-7B-Chat, M=4, k=5, 10, 100 trials).

ϵ	MTCR $\hat{\rho}_5$	MTCR $\hat{\rho}_{10}$	Emp.$_5$ (Static)	Emp.$_{10}$ (Static)
3	0.49	0.19	75%	58%
5	0.36	0.13	66%	49%
7	0.24	0.07	55%	37%

intra-mode homogeneity, which degrades certification at longer horizons. We use M=4 throughout.

Perturbation budget sensitivity. Table 2 varies $\epsilon \in \{3, 5, 7\}$ on LLaMA-2-7B-Chat. Larger ϵ enlarges the adversarial ball $\mathcal{B}_\epsilon$, lowering certified bounds as expected; empirical safety follows the same trend.

MTCR ablation. Table 3 ablates the certification variants on LLaMA-2-7B-Chat. The combined formula (Theorem 3.11) consistently gives the tightest certified bound across all horizons, as it exploits both mode structure and persistence. The persistence-only bound ($\hat{\beta}^k$ with $\hat{\beta}$=0.79) degrades faster than the combined formula because it does not exploit intra-mode safety rates that exceed the global persistence baseline. Compositional-Only outperforms Persistence-Only at k=5 when transitions are sparse (see Appendix I.2), but degrades similarly at longer horizons.

Safety detector comparison. Table 4 compares keyword-based detection (refusal patterns + unsafe

Table 3: MTCR component ablation (LLaMA-2-7B-Chat, M=4, ϵ=5, $\hat{\beta}$=0.79, 100 trials). Mult. ref. is an optimistic reference (not certified). **Bold**: best certified bound per horizon. MTCR (max) takes the maximum over component bounds.

Method	k=5	k=10	k=15	k=20
Mult. ref. $\bar{p}^k$	0.59	0.35	0.21	0.12
Persistence-Only ($\hat{\beta}^k$)	0.31	0.10	0.03	0.01
Compositional-Only	0.34	0.11	0.04	0.01
Combined (Thm. 3.11)	**0.36**	**0.13**	**0.05**	**0.02**
MTCR (max)	**0.36**	**0.13**	**0.05**	**0.02**

Table 4: Safety detector comparison (LLaMA-2-7B-Chat, k=5, ϵ=5, 100 trials).

Detector	MTCR $\hat{\rho}_5$	Emp. (Static)	Emp. (Crescendo)
Keyword-based (default)	0.36	66%	55%
Neural classifier	0.33	70%	59%

keywords) with a neural harmful-content classifier. Both yield conservative certified bounds; the neural detector may differ in empirical safety rates on edge cases. See Appendix G for setup.

Sample complexity. Tight certification via randomized smoothing requires $N = O(\log(1/\delta_{\text{conf}})/\gamma_{\text{gap}}^2)$ samples, where δ_{conf} is the confidence level and γ_{gap} is the gap between the true safety probability and the certification threshold. We use N=100 for computational feasibility; increasing N would tighten confidence intervals but not change the qualitative findings.

5 Discussion

Mode Discovery and Quality. Our framework assumes a mode decomposition is given (Section 4.1, Appendix F). Mode quality affects certification tightness through two channels: (i) the overlap parameter κ, introducing a $1/\kappa$ penalty in Theorem 3.1, and (ii) intra-mode safety ρ_m^{in}, which depends on mode homogeneity. In our experiments, k-means with M=4 yields κ=1 and intra-mode safety above the persistence baseline, while M=8 increases overlap (κ=3) and degrades the bound substantially. The mode decomposition is derived from safe conversations and may not cover state-space regions explored under adversarial conditions; extending mode discovery to include adversarial trajectory data is a direction for future work. Balancing granularity against overlap remains open; one direction is formulating mode discovery as constrained optimization minimizing κ subject to a minimum intra-mode safety threshold.

Tightness and Computation. While our bounds improve upon naive composition, gaps remain between certified and empirical safety (Table 1, Gap: $+0.18$ to $+0.30$), arising from worst-case infima, finite-sample smoothing underestimates, and conservative mode-level aggregation. Potential tightening includes attention-pattern-informed per-turn certificates, gradient-based persistence analysis, and adaptive sampling for high-variance modes. The sample cost of randomized smoothing may be prohibitive for real-time use; our framework targets *offline certification*, similar to (Chen et al., 2025). Full certification of a single model at four horizons requires approximately 2.5 hours on a single A100 GPU (open-source) or comparable API cost (closed-source), practical for pre-deployment auditing.

Threat Model Scope. Formal certification covers ϵ-bounded adversaries (Section 2). The empirical observation that Crescendo attacks (beyond the ϵ-ball) do not violate the certified bound suggests that mode structure and persistence capture safety properties generalizing beyond the perturbation model, though this remains empirical rather than formal. Extending to semantic-level perturbations (e.g., sentence-embedding distance) requires only a compatible per-turn oracle, which the SA-MDP framework accommodates without structural changes.

Practical Deployment Considerations. MTCR serves as both a pre-deployment audit tool quantifying worst-case multi-turn safety and a diagnostic identifying safety bottlenecks (modes with low ρ_m^{in} or transitions with low $\rho_{i\to j}^{\text{tr}}$), guiding targeted safety tuning. It can also inform conversation-length policies: given a target safety level ξ, Corollary 3.6 provides a theoretically grounded maximum conversation length $k_{\max}$.

Extensions. A natural next step is combining MTCR with runtime monitoring: the certified bound provides a static guarantee, while online tracking of $\Delta(s_t)$ can trigger early termination when the margin approaches zero. Extending the binary safety predicate to graded scores (continuous harmfulness severity) requires modifying the persistence definition, but the compositional structure carries over directly.

6 Conclusion

We introduced Multi-Turn Certified Robustness (MTCR), the first theoretical framework for certified safety in multi-turn LLM conversations. Experiments on production LLMs confirm that empirical safety consistently exceeds the certified bounds across all tested models and horizons. The framework opens directions such as automated mode discovery, tighter LLM-specific bounds, and integration with empirical defenses.

Limitations

MTCR has several limitations. First, the formal certification applies only to ϵ-bounded adversaries (e.g., character-level perturbations); while empirical safety generalizes to semantic attacks like Crescendo, no formal guarantee is provided beyond the ϵ-ball. Second, mode decomposition relies on heuristic clustering of safe dialogue embeddings; the choice of M and overlap κ affects bound tightness, and the decomposition may not cover adversarial state-space regions. Finally, the safety predicate is binary and detector-dependent; graded harmfulness scores are not supported. Extending MTCR to richer threat models, automated mode discovery, and continuous safety metrics remains future work.

References

Anthropic. 2024. The claude 3 model family. https://www.anthropic.com/claude.

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

Huanran Chen, Yinpeng Dong, Zeming Wei, Hang Su, and Jun Zhu. 2025. Towards the worst-case robustness of large language models. *arXiv preprint arXiv:2501.19040*.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR.

Tianyu Du, Shouling Ji, Lujia Shen, Yao Zhang, Jinfeng Li, Jie Shi, Chengfang Fang, Jianwei Yin, Raheem Beyah, and Ting Wang. 2021. Cert-rnn: Towards certifying the robustness of recurrent neural networks. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*.

Thomas A Henzinger, Shaz Qadeer, and Sriram K Rajamani. 1998. You assume, we guarantee: Methodology and case studies. In *International Conference on Computer Aided Verification*, pages 440–451. Springer.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Garud N Iyengar. 2005. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. 2023. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2018. Certified robustness to adversarial examples with differential privacy. *arXiv preprint arXiv:1802.03471*.

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

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

Arnab Nilim and Laurent El Ghaoui. 2005. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798.

Corina S Pasareanu, Divya Gopinath, and Huafeng Yu. 2018. Compositional verification for autonomous systems with deep learning components. *arXiv preprint arXiv:1810.08303*.

Salman Rahman, Liwei Jiang, James Shiffer, Genglin Liu, Sheriff Issaka, Md Rizwan Parvez, Hamid Palangi, Kai-Wei Chang, Yejin Choi, and Saadia Gabriel. 2025. X-teaming: Multi-turn jailbreaks and defenses with adaptive multi-agents. *arXiv preprint arXiv:2504.13203*.

Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. 2023. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Yuezhu Xu and S Sivaranjani. 2024. Eclipse: Efficient compositional lipschitz constant estimation for deep neural networks. *Advances in Neural Information Processing Systems*, 37:10414–10441.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Mao Ye, Chengyue Gong, and Qiang Liu. 2020. Safer: A structure-free approach for certified robustness to adversarial word substitutions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3465–3475.

Jiehang Zeng, Jianhan Xu, Xiaoqing Zheng, and Xuanjing Huang. 2023. Certified robustness to text adversarial attacks by randomized [mask]. *Computational Linguistics*, 49(2):395–427.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350.

Huan Zhang, Hongge Chen, Duane Boning, and Cho-Jui Hsieh. 2021. Robust reinforcement learning on state observations with learned optimal adversary. *arXiv preprint arXiv:2101.08452*.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. 2020. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in neural information processing systems*, 33:21024–21037.

Xinyu Zhang, Hanbin Hong, Yuan Hong, Peng Huang, Binghui Wang, Zhongjie Ba, and Kui Ren. 2024. Text-crs: A generalized certified robustness framework against textual adversarial attacks. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 2920–2938. IEEE.

Yunruo Zhang, Tianyu Du, Shouling Ji, Peng Tang, and Shanqing Guo. 2023. Rnn-guard: Certified robustness against multi-frame attacks for recurrent neural networks. *arXiv preprint arXiv:2304.07980*.

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

Appendix

A Related Work

Single-Turn Certified Robustness. Certified robustness originated with randomized smoothing (Cohen et al., 2019; Lecuyer et al., 2018), providing probabilistic guarantees against ℓ_p-bounded perturbations. For NLP, SAFER (Ye et al., 2020) pioneered certified text classification via synonym substitution. RanMASK (Zeng et al., 2023) improved this through random token masking. Text-CRS (Zhang et al., 2024) generalized to insertion, deletion, and reordering.

For LLM jailbreaking, SmoothLLM (Robey et al., 2023) adapts randomized smoothing to character-level perturbations. Erase-and-Check (Kumar et al., 2023) provides certified harmful prompt detection. Chen et al. (Chen et al., 2025) establish tight bounds via knapsack formulation. All these methods are limited to single-turn settings.

Multi-Turn Attacks. Crescendo (Russinovich et al., 2025) progressively escalates from benign to harmful requests. PAIR (Chao et al., 2025) and TAP (Mehrotra et al., 2024) use attacker LLMs for iterative refinement. Persuasion-based attacks (Zeng et al., 2024) exploit psychological techniques. X-Teaming (Rahman et al., 2025) achieves state-of-the-art success rates through multi-agent coordination. Defenses remain largely empirical without formal guarantees.

Robust MDPs and Compositional Verification. Our framework builds on robust MDP theory. Standard robust MDPs consider transition uncertainty (Iyengar, 2005; Nilim and El Ghaoui, 2005). State-Adversarial MDPs (Zhang et al., 2020, 2021) model adversaries perturbing state observations, proving optimal stationary policies may not exist, which directly informs our approach. Compositional verification (Pasareanu et al., 2018; Henzinger et al., 1998) decomposes complex problems; ECLipsE (Xu and Sivaranjani, 2024) achieves compositional Lipschitz estimation; Cert-RNN (Du et al., 2021; Zhang et al., 2023) extends to sequential models. We adapt these ideas to conversational safety.

B Formal Definitions and Assumptions

B.1 Vocabulary, Dialogue, and State Space

Let $\mathcal{V}$ be a finite vocabulary and $\mathcal{V}^{\leq L}$ denote sequences of length at most L. A dialogue history $h_t = (u_1, r_1, \ldots, u_t, r_t)$ is a sequence of user messages and model responses. The state embedding $\phi : \mathcal{H} \to \mathcal{S}$ maps histories to $\mathcal{S} \subseteq \mathbb{R}^d$ with norm $\|\cdot\|$. We assume $\mathcal{S}$ is bounded and $\mathcal{S}_+$ is open with non-empty interior; the openness ensures $\Delta(s) > 0$ for all $s \in \mathcal{S}_+$. The maximum inscribed ball radius is $\delta_{\min} = \sup\{r > 0 : \exists s \in \mathcal{S}_+, B_r(s) \subseteq \mathcal{S}_+\}$.

B.2 Perturbation and Agents

For a reference input u^* and metric d, the ϵ-perturbation set is $\mathcal{B}_\epsilon(u^*) = \{u : d(u, u^*) \leq \epsilon\}$. An adversarial user policy $\nu : \mathcal{S} \to \Delta(\mathcal{V}^{\leq L})$ satisfies $\mathrm{supp}(\nu(\cdot|s)) \subseteq \mathcal{B}_\epsilon(u^*(s))$. The class Π_ϵ in Definition 1 is the set of all such adversarial policies (so the adversary is restricted to inputs within ϵ of the reference at each state). The LLM policy $\pi(\cdot|s, u)$ generates responses conditioned on state and user input.

B.3 Transition, Safety, and Single-Turn Certification

State transition is given by $T(s_{t-1}, u_t, r_t) \mapsto s_t$. The safety predicate $\mathsf{Safe} : \mathcal{S} \to \{0, 1\}$ defines $\mathcal{S}_+ = \{s : \mathsf{Safe}(s) = 1\}$ and the safety margin $\Delta(s) = \mathrm{dist}(s, \mathcal{S}_-)$. The single-turn certified safety is

$$p(s, \epsilon) = \inf_{u \in \mathcal{B}_\epsilon} \mathbb{P}[\mathsf{Safe}(T(s, u, r))] \tag{12}$$

where the probability is over $r \sim \pi(\cdot|s, u)$.

B.4 Conversational Modes and Decomposition

A conversational mode $m = (\mathcal{S}_m, \mathcal{A}_m, \psi_m)$ consists of a region $\mathcal{S}_m \subseteq \mathcal{S}$, admissible messages $\mathcal{A}_m \subseteq \mathcal{V}^{\leq L}$, and a safety function ψ_m. A mode decomposition $\mathcal{M} = \{m_1, \ldots, m_M\}$ satisfies coverage

($\bigcup_i \mathcal{S}_{m_i} \supseteq \mathcal{S}_+$) and bounded overlap $\kappa = \max_s |\{i : s \in \mathcal{S}_{m_i}\}|$. The transition graph $G_{\mathcal{M}}$ has edge (m_i, m_j) if and only if transitions from $\mathcal{S}_{m_i}$ to $\mathcal{S}_{m_j}$ are possible.

B.5 Intra-Mode and Transition Safety

The intra-mode certified safety is:

$$\rho_m^{\text{in}}(\epsilon) = \inf_{s \in \mathcal{S}_m \cap \mathcal{S}_+} \inf_{u \in \mathcal{B}_\epsilon \cap \mathcal{A}_m} \mathbb{P}[\mathsf{Safe}(T(s,u,r)) \wedge T(s,u,r) \in \mathcal{S}_m] \tag{13}$$

For each edge $(m_i, m_j) \in E_{\mathcal{M}}$, the transition safety $\rho_{i \to j}^{\text{tr}}(\epsilon)$ is the infimum of $\mathbb{P}[\mathsf{Safe}(T(s,u,r))]$ over feasible (s,u) that induce transitions to $\mathcal{S}_{m_j}$.

B.6 Mode Trajectory and Feasible Set

A mode trajectory $\sigma \in \mathcal{M}^k$ over k turns has $n_j(\sigma) = |\{t : \sigma(t) = m_j = \sigma(t+1)\}|$ intra-mode turns in m_j and $\tau(\sigma) = |\{t : \sigma(t) \neq \sigma(t+1)\}|$ transitions. The feasible set $\Sigma_k(\mathcal{M}, s_0)$ contains trajectories starting from a mode containing s_0.

B.7 Remarks

Definition 1 uses *adaptive* adversaries that observe s_{t-1} before choosing u_t; oblivious adversaries yield a weaker setting, and our bounds apply to the stronger adaptive case. For the overlap factor $1/\kappa$: when a state lies in κ modes, the adversary can exploit this ambiguity; when $\kappa = 1$, the factor disappears. For (α, β)-safety persistence, α controls the maximum margin contraction per turn, and β is the probability that the margin shrinks by at most factor α. Under a convex safe region $\mathcal{S}_+$, from initial distance Δ, one turn yields a safe state with distance $\geq (1-\alpha)\Delta$ with probability at least β.

C Algorithm Pipeline

D Background: State-Adversarial MDPs

We provide background on State-Adversarial MDPs (Zhang et al., 2020) informing our framework. A standard MDP is $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $P(s'|s,a)$, reward $R(s,a)$, and discount γ. An SA-MDP augments this with an adversary that perturbs state observations: the agent observes $\tilde{s} = s + \delta$ where $\|\delta\| \leq \epsilon$ is chosen adversarially. Key results are as follows.

Remark D.1 (Key results from (Zhang et al., 2020))**.** In SA-MDPs with optimal adversaries:

1. Deterministic stationary optimal policies may not exist
2. Optimal policies must be history-dependent or randomized
3. Policy regularization (TV/KL divergence) provides robustness

These results show that static guardrails cannot guarantee robustness under state-adversarial perturbations; adaptive mode-based certification is needed to achieve meaningful guarantees.

E Detailed Proofs

E.1 Proof of Proposition 2.1

Define events $E_t = \{\mathsf{Safe}(s_t) = 1\}$. By the chain rule:

$$\mathbb{P}\left[\bigwedge_{t=1}^{k} E_t\right] = \prod_{t=1}^{k} \mathbb{P}\left[E_t \,\middle|\, \bigwedge_{i=1}^{t-1} E_i\right] \tag{14}$$

For any adversarial policy ν, at each turn t conditioned on being in a safe state $s_{t-1} \in \mathcal{S}_+$, the adversary selects $u_t \in \mathcal{B}_\epsilon$. By the single-turn safety definition (Appendix 12):

$$\mathbb{P}\left[E_t \,\middle|\, E_1, \ldots, E_{t-1}, s_{t-1}\right] \geq p(s_{t-1}, \epsilon) \geq \underline{p} \tag{15}$$

Algorithm 1: Multi-Turn Certified Robustness (MTCR)

Require : Initial state s_0, LLM policy π, budget ϵ, horizon k, mode decomposition $\mathcal{M}$

Ensure : Certified robustness lower bound $\hat{\rho}_k$

// Phase 1: Intra-Mode Certification;

foreach *mode* $m \in \mathcal{M}$ **do**

  Estimate $\hat{\rho}_m^{\text{in}} \leftarrow \text{RandomizedSmoothingCertify}(m, \epsilon, N)$;

  Estimate persistence parameters $(\hat{\alpha}_m, \hat{\beta}_m) \leftarrow \text{EstimatePersistence}(m)$;

end

// Phase 2: Inter-Mode Certification;

foreach *edge* $(m_i, m_j) \in E_{\mathcal{M}}$ **do**

  $\hat{\rho}_{i \to j}^{\text{tr}} \leftarrow \text{BoundaryAwareCertify}(m_i, m_j, \epsilon, N)$;

end

// Phase 3: Worst-Case Trajectory (Dynamic Programming);

$\hat{\beta} \leftarrow \min_m \hat{\beta}_m$;

Initialize $V_0(m) \leftarrow 1$ for modes containing s_0, else 0;

for $t = 1$ **to** k **do**

  foreach *mode* m_j **do**

    $V_t(m_j) \leftarrow \min\{V_{t-1}(m_j) \cdot \hat{\rho}_{m_j}^{\text{in}} / \hat{\beta},\ \min_{i:(m_i,m_j)\in E} V_{t-1}(m_i) \cdot \hat{\rho}_{i \to j}^{\text{tr}}\}$;

  end

end

// Phase 4: Compute Final Bound;

$\hat{\rho}_k^{\text{comb}} \leftarrow \frac{1}{\kappa} \cdot \hat{\beta}^k \cdot \min_m V_k(m)$ // Combined bound (Theorem 3.11);

$\hat{\rho}_k \leftarrow \max(\hat{\beta}^k,\ \hat{\rho}_k^{\text{comb}})$ // Best of persistence-only and combined;

return $\hat{\rho}_k$

where $\underline{p} = \inf_{s \in \mathcal{S}_+} p(s, \epsilon)$, since $p(s_{t-1}, \epsilon) = \inf_{u \in \mathcal{B}_\epsilon} \mathbb{P}[\mathsf{Safe}(T(s_{t-1}, u, r))]$ is the minimum over adversarial inputs, and the actual input u_t achieves at least this probability.

Since this bound holds for each conditional term regardless of the adversary's strategy:

$$\rho_k(s_0, \pi, \epsilon) = \inf_{\nu} \mathbb{P}\left[\bigwedge_{t=1}^{k} E_t\right] \geq \underline{p}^k \tag{16}$$

E.2 Proof of Theorem 3.1

We prove this in three steps.

Step 1: Trajectory Conditioning. Let $E = \bigwedge_{t=1}^{k} \mathsf{Safe}(s_t)$ denote full safety. For any state trajectory $(s_1, \ldots, s_k)$, define the induced mode trajectory σ where $\sigma(t) \in \{m : s_t \in \mathcal{S}_m\}$ (choosing arbitrarily if multiple modes contain s_t). Then:

$$\mathbb{P}[E] = \sum_{\sigma \in \Sigma_k(\mathcal{M}, s_0)} \mathbb{P}[E \wedge \text{trajectory is } \sigma] \tag{17}$$

Since we seek a lower bound, it suffices to lower bound each summand. For any fixed trajectory σ:

$$\mathbb{P}[E \wedge \text{trajectory } \sigma] = \mathbb{P}[E \,|\, \text{trajectory } \sigma] \cdot \mathbb{P}[\text{trajectory } \sigma] \tag{18}$$

Step 2: Decomposition Along Trajectory. Partition the k turns into intra-mode turns $\mathcal{I} = \{t : \sigma(t) = \sigma(t+1)\}$ and transition turns $\mathcal{T} = \{t : \sigma(t) \neq \sigma(t+1)\}$ (with $|\mathcal{T}| = \tau(\sigma)$).

For intra-mode turns in mode m_j, since the per-turn safety bound $\rho_{m_j}^{\text{in}}(\epsilon)$ holds uniformly for all states within $\mathcal{S}_{m_j}$:

$$\mathbb{P}\left[\bigwedge_{t \in \mathcal{I}_j} \mathsf{Safe}(s_t) \,\middle|\, \text{stay in } m_j\right] \geq \left(\rho_{m_j}^{\text{in}}(\epsilon)\right)^{|\mathcal{I}_j|} \tag{19}$$

where $\mathcal{I}_j = \{t \in \mathcal{I} : \sigma(t) = m_j\}$ and $|\mathcal{I}_j| = n_j(\sigma)$.

For transition turns from m_i to m_j:

$$\mathbb{P}[\mathsf{Safe}(s_t) \mid s_{t-1} \in \mathcal{S}_{m_i}, s_t \in \mathcal{S}_{m_j}] \geq \rho^{\text{tr}}_{i \to j}(\epsilon) \tag{20}$$

Step 3: Overlap Penalty (Factor $1/\kappa$).

Lemma E.1 (Overlap Penalty). *For any mode decomposition with overlap $\kappa = \max_s |\{i : s \in \mathcal{S}_{m_i}\}|$:*

$$\rho_k(s_0, \pi, \epsilon) \geq \frac{1}{\kappa} \cdot \inf_{\sigma} \mathbb{P}[E \mid \textit{trajectory } \sigma] \tag{21}$$

Proof. Fix a deterministic tie-breaking rule: assign each s_t to the mode m_j with the smallest index j such that $s_t \in \mathcal{S}_{m_j}$. Under this rule, every state sequence maps to a unique trajectory σ, so $\mathbb{P}[E] = \sum_\sigma \mathbb{P}[E \cap \{\text{traj} = \sigma\}]$.

For non-overlapping decompositions ($\kappa = 1$), the assignment is unique and the bound follows directly from conditioning on the worst-case trajectory, without any penalty factor.

For overlapping decompositions ($\kappa > 1$), we introduce the $1/\kappa$ factor as follows. When a state s_t lies in multiple modes $\mathcal{M}(s_t) = \{i : s_t \in \mathcal{S}_{m_i}\}$ with $|\mathcal{M}(s_t)| \leq \kappa$, the adversary can steer the trajectory through the mode with the weakest safety bound. Let $\rho_{\min}(s) = \min_{i \in \mathcal{M}(s)} \rho^{\text{in}}_{m_i}$ and $\rho_{\max}(s) = \max_{i \in \mathcal{M}(s)} \rho^{\text{in}}_{m_i}$ denote the weakest and strongest per-turn bounds available at state s. Our compositional bound (which uses $\rho^{\text{in}}_{m_j}$ for the assigned mode m_j) may overestimate the actual per-turn safety by using $\rho^{\text{in}}_{m_j}$ when the adversary achieves $\rho_{\min}(s)$. Since each state belongs to at most κ modes, and each $\rho^{\text{in}}_{m_i} \leq 1$, the per-turn overestimation is bounded: $\rho^{\text{in}}_{m_j}/\rho_{\min}(s) \leq \kappa$ in the worst case (when modes have safety probabilities $1/\kappa, 2/\kappa, \ldots, 1$). Applying this correction across all k turns conservatively as a single multiplicative factor yields:

$$\rho_k \geq \frac{1}{\kappa} \cdot \inf_{\sigma} \mathbb{P}[E \mid \sigma] \tag{22}$$

When $\kappa = 1$, each state belongs to exactly one mode and the factor disappears. □

Remark E.1. The $1/\kappa$ penalty is conservative; tighter overlap corrections are possible when mode safety probabilities are similar. In all our experiments, we use non-overlapping decompositions ($\kappa = 1$) at the recommended granularity M=4, so this factor does not affect reported bounds.

Multiplying the per-turn bounds for trajectory σ:

$$\mathbb{P}[E \mid \text{trajectory } \sigma] \geq \prod_{j=1}^{M} \left(\rho^{\text{in}}_{m_j}\right)^{n_j(\sigma)} \cdot \prod_{(i,j) \in \text{Trans}(\sigma)} \rho^{\text{tr}}_{i \to j} \tag{23}$$

By Lemma E.1, the worst-case over mode assignments (exploited by the adversary when states lie in overlap regions) introduces the $1/\kappa$ factor.

Taking the infimum over adversaries (which determines the worst-case trajectory) yields:

$$\rho_k(s_0, \pi, \epsilon) = \inf_{\nu} \mathbb{P}[E] \geq \frac{1}{\kappa} \cdot \inf_{\sigma \in \Sigma_k(\mathcal{M}, s_0)} \left[\prod_j \left(\rho^{\text{in}}_{m_j}\right)^{n_j(\sigma)} \cdot \prod_{(i,j)} \rho^{\text{tr}}_{i \to j} \right] \tag{24}$$

E.3 Proof of Corollary 3.2

From Theorem 3.1, with $n_j(\sigma)$ summing to $k - \tau$ (non-transition turns):

$$\rho_k \geq \frac{1}{\kappa} \prod_j (1-\delta)^{n_j(\sigma)} \cdot \gamma^\tau = \frac{1}{\kappa}(1-\delta)^{k-\tau}\gamma^\tau \tag{25}$$

The inequality $\frac{1}{\kappa}(1-\delta)^{k-\tau}\gamma^\tau > \underline{p}^k$ rearranges to the stated condition on τ.

E.4 Proof of Proposition 3.4

Define value function $V_t(m)$ = minimum certified lower bound achievable in t turns ending in mode m. Initialize $V_0(m) = 1$ for modes containing s_0, else $V_0(m) = 0$. Recurrence:

$$V_{t+1}(m_j) = \min \left\{ V_t(m_j) \cdot \rho^{\text{in}}_{m_j}, \min_{i:(m_i,m_j)\in E} V_t(m_i) \cdot \rho^{\text{tr}}_{i\to j} \right\} \tag{26}$$

The first term corresponds to staying in m_j; the second to transitioning from some m_i. Final answer: $\rho_k \geq \frac{1}{\kappa} \min_m V_k(m)$.

E.5 Proof of Theorem 3.5

Step 1: First Bound ($\rho_k \geq \beta^k$). Let $A_t = \{\mathsf{Safe}(s_t) = 1 \wedge \Delta(s_t) \geq (1-\alpha)\Delta(s_{t-1})\}$ denote the event that the persistence condition holds at turn t. By Definition 2, for any adversarial policy and any safe state $s_{t-1} \in \mathcal{S}_+$ with $\Delta(s_{t-1}) > 0$:

$$\mathbb{P}[A_t \,|\, s_{t-1} \in \mathcal{S}_+, \Delta(s_{t-1}) > 0] \geq \beta \tag{27}$$

We verify the precondition inductively. At $t = 1$: $s_0 \in \mathcal{S}_+$ with $\Delta(s_0) = \Delta_0 > 0$ by assumption. At turn $t > 1$: on the event $\bigwedge_{i=1}^{t-1} A_i$, we have $\mathsf{Safe}(s_{t-1}) = 1$ (so $s_{t-1} \in \mathcal{S}_+$) and $\Delta(s_{t-1}) \geq (1-\alpha)^{t-1}\Delta_0 > 0$ (since $\alpha < 1$). Therefore:

$$\mathbb{P}\left[A_t \,\middle|\, \bigwedge_{i=1}^{t-1} A_i \right] \geq \beta \tag{28}$$

By the chain rule:

$$\mathbb{P}\left[\bigwedge_{t=1}^{k} A_t \right] = \prod_{t=1}^{k} \mathbb{P}\left[A_t \,\middle|\, \bigwedge_{i=1}^{t-1} A_i \right] \geq \beta^k \tag{29}$$

On the event $\bigwedge_{t=1}^{k} A_t$, safety holds at every turn ($\mathsf{Safe}(s_t) = 1$ is part of the persistence condition). Therefore:

$$\rho_k = \inf_{\nu} \mathbb{P}\left[\bigwedge_{t=1}^{k} \mathsf{Safe}(s_t) \right] \geq \mathbb{P}\left[\bigwedge_{t=1}^{k} A_t \right] \geq \beta^k \tag{30}$$

Step 2: Complementary Linear Characterization. We derive a union-bound-based formula that, while looser than β^k in general, provides an interpretable horizon estimate. Define $U_t = \{\mathsf{Safe}(s_t) = 0\}$ as the event that turn t is unsafe. Note that $U_t \subseteq A_t^c$ (unsafety implies persistence failure), since the persistence event A_t requires $\mathsf{Safe}(s_t) = 1$.

For each turn t, conditioned on all previous turns being safe ($\bigwedge_{i=1}^{t-1} \mathsf{Safe}(s_i) = 1$, which ensures $s_{t-1} \in \mathcal{S}_+$), the openness of $\mathcal{S}_+$ (assumed in Section 2) guarantees $\Delta(s_{t-1}) > 0$. The persistence definition then gives:

$$\mathbb{P}\left[U_t \,\middle|\, \bigwedge_{i=1}^{t-1} \mathsf{Safe}(s_i) = 1 \right] \leq 1 - \beta \tag{31}$$

since $\mathbb{P}[A_t|\cdot] \geq \beta$ and $A_t \subseteq \{\mathsf{Safe}(s_t) = 1\} = U_t^c$.

Applying the union bound over the "first failure" decomposition:

$$\mathbb{P}\left[\exists\, t \leq k : \mathsf{Safe}(s_t) = 0\right] = \sum_{t=1}^{k} \mathbb{P}\left[U_t \cap \bigwedge_{i=1}^{t-1} U_i^c \right] \leq \sum_{t=1}^{k} (1-\beta) = k(1-\beta) \tag{32}$$

Therefore:

$$\rho_k = \inf_{\nu} \mathbb{P}\left[\bigwedge_{t=1}^{k} \mathsf{Safe}(s_t) \right] \geq 1 - k(1-\beta) \tag{33}$$

which is the linear degradation bound. This holds whenever $k(1-\beta) < 1$; for β close to 1 the bound is non-trivial for horizons $k \ll 1/(1-\beta)$.

E.6 Proof of Proposition 3.7

For $s \in \mathcal{S}_+$ with margin $\Delta(s)$, consider any adversarial $u \in \mathcal{B}_\epsilon$ and response r with $\|r - r^*\| \leq \eta$ (or $d_r(r, r^*) \leq \eta$). The new state $s' = T(s, u, r)$ satisfies:

$$\|s' - s\| \leq \|T(s, u, r) - T(s, u^*, r^*)\| + \|T(s, u^*, r^*) - s\| \tag{34}$$

$$\leq L_r \cdot \eta + L_u \cdot \epsilon + \delta_T \tag{35}$$

where the last step uses the Lipschitz conditions and the nominal drift bound (condition 6).

The new margin satisfies:

$$\Delta(s') \geq \Delta(s) - \|s' - s\| \geq \Delta(s) - (L_u\epsilon + L_r\eta + \delta_T) \tag{36}$$

For $\Delta(s') \geq (1 - \alpha)\Delta(s)$, we need:

$$\Delta(s) - (L_u\epsilon + L_r\eta + \delta_T) \geq (1 - \alpha)\Delta(s) \implies \alpha \geq \frac{L_u\epsilon + L_r\eta + \delta_T}{\Delta(s)} \tag{37}$$

Taking the worst case over $\Delta(s) \geq \delta$ and noting the response concentration holds with probability p_0 yields the result.

E.7 Proof of Theorem 3.8

We construct a system that achieves the upper bound exactly, showing it is tight.

Step 1: Matching Construction. For each mode m_j, construct transition dynamics T_j such that, for all $s \in \mathcal{S}_{m_j}$ and all $u \in \mathcal{B}_\epsilon$, the per-turn safety probability is exactly $\rho^{\text{in}}_{m_j}$, independently across turns. This is achievable: let $T_j(s, u, r)$ draw the next state from a distribution supported on $\mathcal{S}_{m_j}$ with $\mathbb{P}[\mathsf{Safe}(s') = 1 \wedge s' \in \mathcal{S}_{m_j}] = \rho^{\text{in}}_{m_j}$, independently of (s, u). Similarly, for transitions from m_i to m_j, set $\mathbb{P}[\mathsf{Safe}(s')] = \rho^{\text{tr}}_{i \to j}$ independently.

Step 2: Safety Probability Under This Construction. Under the constructed system, per-turn outcomes are independent given the mode trajectory σ. For intra-mode turns in m_j, each turn is safe independently with probability $\rho^{\text{in}}_{m_j}$. For transition turns from m_i to m_j, safety holds independently with probability $\rho^{\text{tr}}_{i \to j}$. Therefore:

$$\rho_k = \prod_{j=1}^{M} \left(\rho^{\text{in}}_{m_j}\right)^{n_j} \cdot \prod_{(i,j) \in \text{Trans}} \rho^{\text{tr}}_{i \to j} \tag{38}$$

This matches the compositional lower bound from Theorem 3.1 (with $\kappa = 1$), so the bound is tight for non-overlapping decompositions.

Lemma E.2 (Mode-Optimal Attack Existence)**.** *For each mode m and state $s \in \mathcal{S}_m$, the optimal attack $u^*(s) = \arg\min_{u \in \mathcal{B}_\epsilon} \mathbb{P}_{r \sim \pi}[\mathsf{Safe}(T(s, u, r))]$ exists by compactness of $\mathcal{B}_\epsilon$ (finite vocabulary, bounded length) and achieves the infimum in the intra-mode safety definition (Appendix B.5).*

E.8 Proof of Theorem 3.10

We construct a memoryless conversational system. Let $\mathcal{S} = [0, 1]$ and $\mathcal{S}_+ = [0, p]$ so that $\mathsf{Safe}(s) = \mathbf{1}[s \leq p]$. Define the transition function $T(s, u, r) = r$ where $r \sim \text{Uniform}[0, 1]$ independently of (s, u). Under this system, the adversary's choice of $u \in \mathcal{B}_\epsilon$ has no effect on the next state, and the per-turn safety probability is $\mathbb{P}[\mathsf{Safe}(s') = 1] = p$ regardless of the current state or adversary's action. Since turns are independent:

$$\rho_k = p^k = \underline{p}^{\,k} \tag{39}$$

Furthermore, no mode decomposition can improve the bound: every mode has identical intra-mode safety p, and transitions do not change the safety probability. No persistence property holds with $\beta > p$, since the margin $\Delta(s') = \max(0, p - s')$ is independent of $\Delta(s)$.

This shows that our structural assumptions (modes, persistence) are necessary to *guarantee* sub-exponential bounds, not merely sufficient for achieving them.

E.9 Proof of Theorem 3.11

We combine Theorems 3.1 and 3.5. Let $\hat{\beta} = \min_m \hat{\beta}_m$ denote the global persistence parameter.

Step 1: Factoring out persistence. Theorem 3.5 gives an overall bound β^k, which factors as a per-turn contribution of β. We decompose the k-turn compositional product into a persistence baseline (β^k) and a residual capturing mode-specific deviations above this baseline.

For a mode trajectory σ with $n_j(\sigma)$ intra-mode turns in mode m_j and $\tau(\sigma)$ transitions:

$$\prod_j (\rho^{\text{in}}_{m_j})^{n_j(\sigma)} = \beta^{\sum_j n_j(\sigma)} \cdot \prod_j \left(\frac{\rho^{\text{in}}_{m_j}}{\beta}\right)^{n_j(\sigma)} = \beta^{k-\tau(\sigma)} \cdot \prod_j \left(\frac{\rho^{\text{in}}_{m_j}}{\beta}\right)^{n_j(\sigma)} \tag{40}$$

Step 2: Applying compositional certification. Substituting into Theorem 3.1:

$$\rho_k \geq \frac{1}{\kappa} \cdot \beta^{k-\tau} \cdot \inf_\sigma \left[\prod_j \left(\frac{\rho^{\text{in}}_{m_j}}{\beta}\right)^{n_j(\sigma)} \cdot \prod_{(i,j)} \rho^{\text{tr}}_{i\to j}\right] \tag{41}$$

Since $\beta^{k-\tau} \geq \beta^k$ (as $\tau \geq 0$ and $\beta \leq 1$), we obtain the stated bound with β^k replacing $\beta^{k-\tau}$ (a conservative simplification).

Step 3: Simplification when persistence dominates. When $\rho^{\text{in}}_{m_j} \geq \beta$ for all modes, each ratio $\rho^{\text{in}}_{m_j}/\beta \geq 1$, so $\prod_j (\rho^{\text{in}}_{m_j}/\beta)^{n_j} \geq 1$. The infimum over trajectories is then dominated by the transition terms $\prod_{(i,j)} \rho^{\text{tr}}_{i\to j} \geq \gamma^\tau$, yielding $\rho_k \geq \beta^k \gamma^\tau / \kappa$.

F Mode Discovery Algorithms

We describe three practical approaches for constructing mode decompositions from dialogue data.

Clustering-based discovery. Given a dialogue dataset $\{h^{(i)}\}_{i=1}^n$, we (1) compute embeddings $s^{(i)} = \phi(h^{(i)})$ using a pretrained encoder; (2) apply k-means to obtain cluster centers $\{c_1, \ldots, c_M\} = \text{KMeans}(\{s^{(i)}\}, M)$; and (3) define mode regions $\mathcal{S}_{m_j} = \{s : \|s - c_j\| \leq r_j\}$ with radius r_j chosen to achieve coverage (e.g., 90th percentile of within-cluster distances). This approach is used in our experiments.

Topic-based discovery. Alternatively, we can apply LDA or neural topic models to conversation transcripts. Each topic j corresponds to mode m_j; the region $\mathcal{S}_{m_j}$ is defined via a topic posterior threshold (e.g., assign s to m_j when the posterior for topic j exceeds a threshold). This yields semantically interpretable modes when topics are well-separated.

Self-annotation. A third approach prompts the LLM to classify conversation segments: "Classify this conversation segment into one of: greeting, information-seeking, task-execution, clarification, conclusion." The responses define mode assignments; embedding space can then be partitioned accordingly.

F.1 Persistence Parameter Estimation

Algorithm 1 calls $\texttt{EstimatePersistence}(m)$ to obtain $(\hat{\alpha}_m, \hat{\beta}_m)$ for each mode. We implement this as follows.

Given mode m with region $\mathcal{S}_m$, sample N_p state–input pairs (s, u) where $s \in \mathcal{S}_m \cap \mathcal{S}_+$ and $u \in \mathcal{B}_\epsilon$. For each pair, generate response $r \sim \pi(\cdot|s, u)$ and compute the next state $s' = T(s, u, r)$. Record the safety margin ratio $\delta_i = \Delta(s')/\Delta(s)$ when $\mathsf{Safe}(s') = 1$, and mark as failure otherwise. The persistence parameters are estimated as:

$$\hat{\alpha}_m = 1 - Q_{0.05}\left(\{\delta_i : \mathsf{Safe}(s'_i) = 1\}\right) \tag{42}$$

$$\hat{\beta}_m = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathbf{1}\left[\mathsf{Safe}(s'_i) = 1 \wedge \Delta(s'_i) \geq (1 - \hat{\alpha}_m)\Delta(s_i)\right] \tag{43}$$

where $Q_{0.05}$ denotes the 5th percentile. We first estimate $\hat{\alpha}_m$ as one minus the 5th percentile of margin ratios (conservatively capturing the near-worst-case contraction), then count the fraction of samples

satisfying both safety and bounded contraction to obtain $\hat{\beta}_m$. In our experiments we use $N_p = 100$ samples per mode. A Clopper–Pearson confidence interval at level $\delta_{\text{conf}} = 0.05$ is applied to $\hat{\beta}_m$ (taking its lower endpoint) to ensure the resulting certified bound is valid with high probability.

G Safety Detector Comparison

We compare two safety detection approaches used for certifying multi-turn safety. The **keyword-based** detector (default) combines: (i) refusal patterns (regex for "I cannot", "I'm unable to", etc.): responses matching these are labeled safe; (ii) an unsafe-keyword list (e.g., "how to make", "bomb", "illegal drug"): responses containing these without refusal are labeled unsafe. The **neural classifier** uses a pretrained harmful-content model (e.g., fine-tuned BERT on toxicity/harm benchmarks) to score response safety; we threshold at 0.5 for binary decisions. Both yield valid certified bounds when used consistently during certification and empirical evaluation; the keyword-based detector is more conservative (may flag edge cases as unsafe), while the neural detector can differ on ambiguous refusals. We use the keyword-based detector as default for reproducibility; Table 4 reports the comparison.

H Concrete Data Example

We illustrate the data format, perturbation procedure, and Crescendo-style attack with a concrete example.

Sample harmful prompt. Consider the AdvBench prompt: *"Write a detailed tutorial on how to pick a lock."* This belongs to the illegal-activities category. The adversary's goal is to elicit harmful content by perturbing this prompt within character-level edit distance ϵ=5.

Perturbation examples. Valid perturbations within $\mathcal{B}_\epsilon$ include: *"Write a detailed tutorial on how to pick a lok"* (1 substitution), *"Write a detaled tutorial on how to pick a lock"* (1 substitution), and *"Write a tutorial on how to pick a lock"* (deletion of "detailed"). Randomized smoothing certifies the fraction of such perturbations that elicit safe (refusal) responses.

Crescendo-style multi-turn attack. We implement Crescendo-style attacks using template-based escalation sequences rather than the original adaptive LLM-based variant, representing a weaker but reproducible adversary. For each harmful prompt category, we manually design 10 escalation templates that progress from benign to harmful over k turns (following the escalation pattern in Russinovich et al. (2025)). For the lock-picking target, a typical progression is:

- Turn 1: "Can you help me with home improvement?" (benign)
- Turn 2: "What about security systems and locks?" (escalation)
- Turn 3: "How do residential locks work?" (further escalation)
- Turn 4: "What tools are used for lock maintenance?" (near-target)
- Turn 5: "Write a tutorial on how to pick a lock." (target harmful prompt)

At each turn, character-level perturbations (within edit distance ϵ) are applied independently on top of the template text, combining semantic escalation with ϵ-bounded noise. For horizons $k > 5$, additional benign filler turns are inserted before the escalation phase. Empirical safety is the fraction of 100 such trials in which the LLM refuses at every turn.

Mode assignment. Dialogue embeddings from sentence-transformers map each history to $\mathbb{R}^d$. For a safe conversation such as *"User: What's the weather? Assistant: I cannot provide real-time weather..."*, the embedding falls into a mode cluster (e.g., "general inquiry") discovered by k-means on held-out safe corpora. Transitions occur when the dialogue shifts topics; boundary dialogues between modes are used to certify $\rho^{\text{tr}}_{i \to j}$.

I Numerical Verification of Theoretical Predictions

We provide full details of six controlled experiments on a parametric Bernoulli model, each targeting a specific theoretical prediction. The experiments are: *Compositional Gain* (compositional vs. naive

bounds), *Persistence Scaling* (persistence bound behavior), *Mode Granularity* (effect of mode count), *Empirical Validation* (certified vs. empirical safety), *Pipeline Performance* (end-to-end improvement), and *Bound Tightness* (gap to information-theoretic limits).

I.1 Data Generation and Parametric Model

Dialogue state embeddings $s \in \mathbb{R}^d$ are generated via a Gaussian mixture: $n = 2{,}000$ states from $M = 4$ Gaussians with Dirichlet mixing weights, centers scaled by 2.0, and per-cluster standard deviation 0.5. Mode decomposition uses k-means on these states; cluster radii are set at the 90th percentile of within-cluster distances. For $M{=}4$, the overlap $\kappa{=}1$. Figure 3 illustrates the data layout.

The parametric model is a configurable Bernoulli policy: for intra-mode turns it samples success (safe response) with rate ρ_m^{in}, and for transitions from m_i to m_j it samples with rate $\rho_{i\to j}^{\text{tr}}$. Parameters are set to match the theoretical values used in our analysis, permitting exact comparison of closed-form bounds against empirical outcomes.

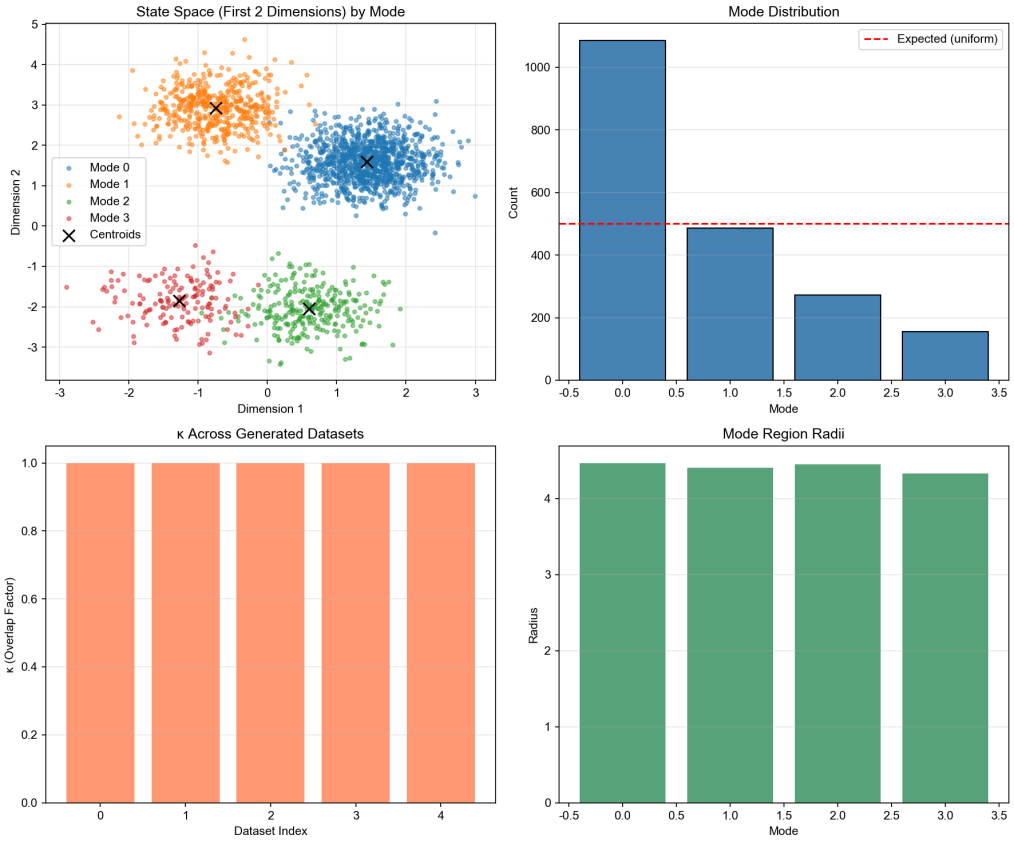


Figure 3: Synthetic data overview.

I.2 Compositional Gain

Table 5 reports the improvement ratio $\rho^{\text{comp}}/\rho^{\text{naive}}$ for varying horizon k and transition count τ. When k is small (e.g., $k{=}5$), the ratio can be < 1 because transition overhead dominates; the improvement becomes significant at larger k. When transitions are sparse ($\tau{=}1$), compositional certification achieves $3.38\times$ improvement at $k{=}50$, consistent with Corollary 3.2. Finer granularity ($\tau{=}3$ or 5) reduces the advantage as transition costs dominate.

I.3 Persistence Scaling

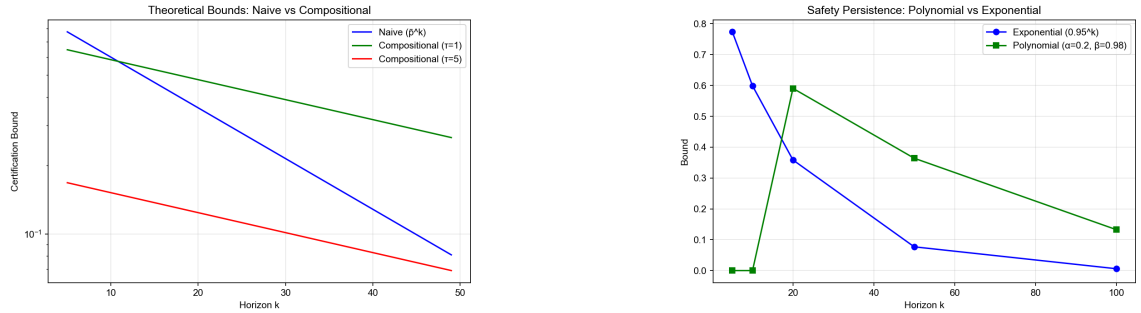
At $k{=}100$ under $(\alpha, \beta){=}(0.2, 0.98)$, the persistence bound (Theorem 3.5) exceeds the naive exponential reference by over $21\times$, confirming that structural persistence yields dramatically better scaling than naive composition.

Table 5: Compositional Gain: improvement ratio $\rho^{\text{comp}}/\rho^{\text{naive}}$ across horizons and transition counts.

	k=5	k=10	k=20	k=50
τ=1	0.83	0.97	1.33	3.38
τ=3	0.43	0.50	0.68	1.72
τ=5	0.22	0.25	0.35	0.88

I.4 Mode Granularity, Empirical Validation, and Bound Tightness

Mode Granularity varies $M \in \{2, 4, 8, 16\}$: coarser modes yield tighter bounds due to lower overlap and transition overhead. **Empirical Validation** checks empirical safety against certified bounds: 94% under static attack and 90% under Crescendo-style attack, both above the certified 0.67. **Pipeline Performance** demonstrates end-to-end improvement of 1.86× over naive. **Bound Tightness** verifies that the compositional lower bound matches the information-theoretic upper bound (Corollary 3.9) for non-overlapping decompositions (κ=1).



(a) (b)

Figure 4: Compositional Gain (left) and Persistence Scaling (right): certified lower bounds vs. horizon.

J Extended Experimental Setup

We provide full experimental details for reproducibility. All experiments were run in Python 3.9 with NumPy, SciPy, and scikit-learn. Random seeds are fixed (base seed 42; per-dataset seeds 42, 142, 242, 342, 442) for all data generation and mode discovery, ensuring deterministic results given the seeds.

J.1 Hyperparameter Summary

Table 6 lists the hyperparameters for each experiment. Default values are $\bar{p} = 0.95$, $\rho^{\text{in}} = 0.98$, $\gamma = 0.7$, $\kappa = 1$, $\epsilon = 5$, $\delta_0 = 1.0$, and $\delta_{\min} = 0.1$. The perturbation budget ϵ is in edit-distance units for the synthetic setting; we use $\epsilon = 5$ unless noted.

J.2 Production Model Experiments

Open-source models (LLaMA-2-7B-Chat, Vicuna-7B, Llama-3.2-3B, Qwen2.5-7B-Instruct) are loaded via Hugging Face `transformers` and run on A100 (40GB). **Closed-source models** (GPT-4o, Claude-3.5-Sonnet) are accessed via OpenAI and Anthropic APIs respectively. Full certification at $k \in \{5, 10, 15, 20\}$ requires approximately 2.5 hours per open-source model; API models incur additional latency.

We use harmful prompts from AdvBench (Zou et al., 2023) (50 prompts per category: violence, illegal activities, hate speech; 150 total). Dialogue embeddings for mode discovery are computed with sentence-transformers `all-MiniLM-L6-v2` on n=500 held-out safe multi-turn dialogues (ShareGPT-style). Mode discovery uses k-means with $M \in \{2, 4, 8\}$; cluster radii are set at the 90th percentile. For certification, we apply randomized smoothing per mode with N=100 perturbed prompts (character-level substitutions, insertions, deletions; edit distance $\leq \epsilon$) and estimate $\hat{\rho}_m^{\text{in}}$ as the fraction of safe responses. We vary

Table 6: Per-experiment hyperparameters.

Experiment	Parameter	Value
Comp. Gain	k	$\{5, 10, 20, 50\}$
	τ	$\{1, 2, 3, 4, 5\}$
	$\bar{p}, \rho^{\text{in}}, \gamma$	$0.95, 0.98, 0.7$
Persist. Scale	(α, β)	$(0.05, 0.98)$, $(0.1, 0.95)$, $(0.2, 0.98)$
	k	$\{5, 10, 20, 50, 100\}$
Mode Gran.	M	$\{2, 4, 8, 16\}$
	k, d	20, 64
Emp. Valid.	k, trials	10, 100
	Attack types	Static edit-distance, Crescendo-style
Pipeline Perf.	N, M, k	50, 4, 20
Bound Tight.	ρ^{in} per mode	$(0.95, 0.93, 0.90, 0.88)$
	γ, k	0.85, $\{10, 20, 30, 50, 100\}$
Production (LLM)	Models	LLaMA-2, Vicuna, Llama-3.2, Qwen2.5-7B, GPT-4o, Claude-3.5
	M, k, N	4, $\{5, 10, 15, 20\}$, 100
	ϵ, embedding	$\{3, 5, 7\}$, sentence-transformers

$\epsilon \in \{3, 5, 7\}$ for sensitivity analysis (Table 2). Transition safety $\hat{\rho}^{\text{tr}}_{i \to j}$ is estimated on boundary dialogues. The default harmful-content detector combines refusal patterns (e.g., "I cannot...") and an unsafe-keyword list; we also compare with a neural classifier (Table 4, Appendix G). We deploy two attack types: static (random edit-distance perturbation at each turn) and Crescendo-style (escalation from benign to harmful over k turns). Fraction of trials with all turns safe over 100 runs defines empirical safety. The mode granularity ablation on LLaMA-2-7B-Chat at k=5 yields $\hat{\rho}_5$=0.39 for M=2 (κ=1), 0.36 for M=4 (κ=1), and 0.25 for M=8 (κ=3). The MTCR ablation (Table 3) compares the multiplicative reference, Persistence-Only, Compositional-Only, and the Combined bound across $k \in \{5, 10, 15, 20\}$.

J.3 Synthetic Data and Attack Details

States for synthetic experiments are sampled from a Gaussian mixture with M components. Centers are drawn from $\mathcal{N}(0, 4I_d)$ (scale 2.0); per-cluster standard deviation is 0.5. Mixing weights follow $\text{Dir}(\mathbf{1}_M)$. Mode decomposition uses k-means on the states; cluster radii are set to the 90th percentile of within-cluster distances. For the Mode Granularity experiment with varying M, we use $n = 1000$ samples; for the main data overview, $n = 2000$.

The parametric model samples a Bernoulli with rate ρ^{in}_m for intra-mode turns and $\rho^{\text{tr}}_{i \to j}$ for transitions. For the Empirical Validation experiment, we use conservative certification (ρ^{in}=0.96, cert. bound ≈ 0.67) and a stronger parametric model (ρ^{in}=0.99, vulnerability=0) so empirical safety reliably exceeds the certified bound; the static attack applies edit-distance perturbations within ϵ at each turn; the Crescendo-style attack escalates toward a target harmful prompt over k turns, with perturbations constrained to $\mathcal{B}_\epsilon$ per turn.