CERTIFYING LLM SAFETY ACROSS ADAPTIVE MULTI-TURN CONVERSATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A large class of language-model safety certificates is local: it bounds the effect of perturbing one prompt. A multi-turn jailbreak is different. The attacker observes each response, adapts the next message, and succeeds if any turn drives the conversation into an unsafe state. We formulate certification for this sequential event and introduce **Multi-Turn Certified Robustness (MTCR)**, a composition layer that lifts valid one-step probability bounds to adaptive conversations. MTCR partitions safe dialogue histories into abstract states and certifies the safe probability mass transferred between them. A robust dynamic program then propagates this mass over the conversation horizon, retaining reachability and state-dependent risk that a global product discards. We further refine the abstraction with safety-margin bins, recording not only whether a trajectory remains safe but how close it moves toward the decision boundary. The resulting certificate requires neither independence across turns nor a fixed attack trajectory. Across four open-weight and two hosted language models, MTCR improves the certified lower bound over global worst-case multiplication at every evaluated horizon; on LLaMA-2-7B-Chat, the improvement grows from $1.24\times$ at five turns to $2.00\times$ at twenty turns. Controlled experiments isolate the source of this gain and recover exact survival probabilities when the abstract transition model is known. These results establish state-aware composition as a practical route from single-turn robustness to certifiable conversational safety.

1 INTRODUCTION

Language-model safety is usually tested one prompt at a time. Conversations are not. After a refusal, an attacker can reframe the request, introduce intermediate context, or exploit information revealed by the model's previous response. Crescendo and human-guided jailbreaks demonstrate how a sequence of ordinary-looking turns can gradually reach behavior that is difficult to elicit directly (Russinovich et al., 2025; Li et al., 2024). The safety event of interest is therefore sequential: every response must remain safe, even though the next input depends on the conversation so far.

Certification methods based on randomized smoothing, deletion, or combinatorial verification provide useful guarantees for a local perturbation set (Robey et al., 2023; Kumar et al., 2023; Chen et al., 2025). Applying such a certificate independently at each turn does not solve the sequential problem. The standard fallback is to multiply a global one-step lower bound, giving $\underline{p}^k$ after k turns. This is valid but wasteful. It charges every trajectory the risk of the single weakest state, including states that the trajectory cannot reach, and repeats that penalty at every turn.

We develop **Multi-Turn Certified Robustness (MTCR)** to preserve the information lost by global multiplication. The state of the sequential process is the full dialogue history. A deterministic abstraction maps each safe history to one of finitely many cells, and a local certification oracle lower-bounds the probability mass transferred from each source cell to each safe successor cell under every admissible user message. These transition bounds define a substochastic operator. Repeated application of this operator gives a lower bound on the probability that the entire conversation remains safe against an adaptive adversary. The recursion sums over all safe successor cells and therefore requires neither turn-wise independence nor a preselected mode trajectory.

The central point is simple: composition becomes tighter when it remembers where safe probability mass goes. We formalize this through a transition-matrix certificate, characterize when it dominates

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

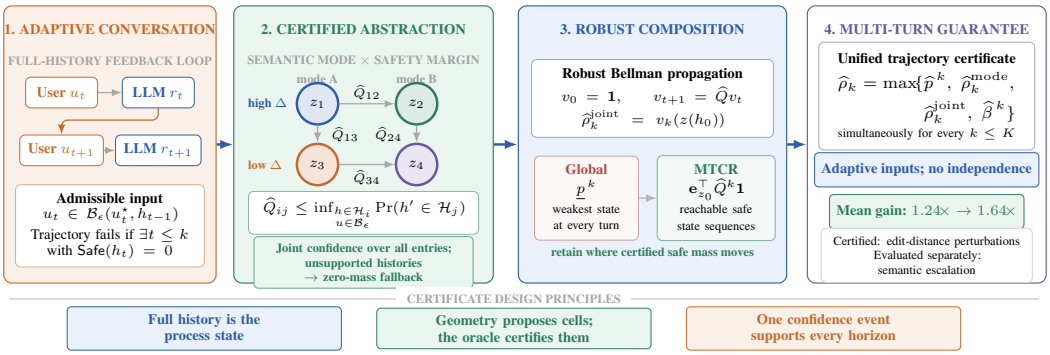


Figure 1: MTCR converts local safety certificates into a trajectory-level guarantee. An adaptive user conditions each message on the observed transcript. A deterministic mode–margin abstraction records where certified safe probability mass moves, and a robust Bellman recursion composes simultaneous transition bounds without assuming independence. The unified certificate selects the strongest jointly valid component at each horizon.

global multiplication, and show that the recursion is exact for a robust Markov abstraction whose transition uncertainty is represented without loss. We then refine each semantic cell by a safety-margin bin. This joint mode–margin state records both behavioral regime and proximity to the unsafe boundary, allowing the same certified dynamic program to exploit margin persistence without multiplying correlated events.

Our experiments evaluate the composition mechanism on six language models. MTCR improves the certified lower bound over global worst-case multiplication throughout the evaluated horizons and perturbation budgets. The gain increases with horizon when safe trajectories avoid weak regions of the state space; on LLaMA-2-7B-Chat, MTCR improves the five-turn certificate from 0.29 to 0.36 and the twenty-turn certificate from 0.01 to 0.02. Controlled experiments separate abstraction quality from local-oracle quality and verify that the method recovers the exact survival probability when supplied with an exact abstract model. We also evaluate multi-turn escalation attacks to measure behavior outside the certified edit-distance set.

The contribution is not a new guardrail or a stronger one-step detector. It is a reusable way to turn local evidence into a trajectory-level guarantee. The theory connects adaptive multi-turn safety to a robust dynamic program, while the mode–margin construction shows how semantic state and proximity to failure can enter the same certified operator. The experiments then expose the practical trade-off: richer abstractions preserve more structure, but each additional cell must be supported by enough samples to certify its outgoing transitions.

2 Related Work

Certified robustness for language models. Randomized smoothing converts local stability under noise into a probabilistic certificate (Cohen et al., 2019; Lecuyer et al., 2018). Its NLP variants certify synonym substitutions, token masking, and discrete edit operations (Ye et al., 2020; Zeng et al., 2023; Zhang et al., 2024). For jailbreak defense, SmoothLLM aggregates predictions under character-level perturbations (Robey et al., 2023); Erase-and-Check certifies harmful-prompt detection under deletion (Kumar et al., 2023); and Chen et al. derive tight worst-case bounds for a discrete perturbation model (Chen et al., 2025). These methods provide the kind of one-step interface consumed by MTCR. Our question begins after such a bound is available: how should local evidence be composed when the next prompt depends on the previous response?

Adaptive multi-turn attacks. Crescendo reaches harmful behavior through gradual semantic escalation (Russinovich et al., 2025). PAIR and TAP repeatedly refine candidate attacks with attacker models (Chao et al., 2025; Mehrotra et al., 2024); persuasion attacks and X-Teaming expand the strategy space through rhetorical tactics and coordinated agents (Zeng et al., 2024; Rahman et al., 2025). Their common feature is feedback: the transcript determines what the attacker tries next. This

invalidates evaluations that treat turns as an independent batch and motivates the history-dependent adversary in Definition 3.1.

Robust control and compositional verification. Robust MDPs optimize under transition uncertainty (Iyengar, 2005; Nilim & El Ghaoui, 2005), while state-adversarial MDPs study corrupted observations (Zhang et al., 2020; 2021). MTCR instead analyzes a reachability event in which the adversary controls the next user message and the defended model supplies the stochastic response. The proof is therefore a robust Bellman argument over a finite abstraction. This perspective is also related to compositional verification, where verified local interfaces replace a monolithic system-level analysis (Pasareanu et al., 2018; Henzinger et al., 1998), and to neural methods that exploit compositional Lipschitz or recurrent structure (Xu & Sivaranjani, 2024; Du et al., 2021; Zhang et al., 2023). The distinctive object here is a probability certificate for an adaptive conversational trajectory, rather than a reward bound or a deterministic sequence model.

3 Problem Formulation

Let $h_t = (u_1, r_1, \ldots, u_t, r_t)$ be the complete dialogue history after turn t. Using the full history as the process state avoids a Markov assumption on a learned embedding. At turn t, an adversarial user policy selects u_t after observing h_{t-1}, subject to $u_t \in \mathcal{B}_\epsilon(u_t^\star, h_{t-1})$, where $u_t^\star$ is the reference message and $\mathcal{B}_\epsilon$ is the perturbation set certified by the local oracle. The defended language-model policy π includes response generation and any randomization used by that oracle. It samples $r_t \sim \pi(\cdot \mid h_{t-1}, u_t)$, after which the state becomes $h_t = (h_{t-1}, u_t, r_t)$. A predicate $\mathsf{Safe}(h_t) \in \{0, 1\}$ determines whether the newest response is safe. We write $\mathcal{H}_+$ and $\mathcal{H}_-$ for histories with safe and unsafe newest responses, respectively.

MTCR does not run dynamic programming over the unbounded history space. A deterministic abstraction $z : \mathcal{H}_+ \to [M]$ maps each safe history to one of M cells. In the experiments, z is obtained from a dialogue embedding followed by nearest-center assignment; the theory requires only that every safe history receive exactly one label. Histories outside the certified support are assigned to a fallback cell with zero outgoing certified mass unless they are verified separately.

The abstraction and the certificate play different roles. Clustering proposes a compact state description, but it does not justify uniform behavior inside a cell. Validity comes from the local oracle, which must cover every history represented by that cell, either directly or through a certified cover of anchor neighborhoods. Refining a cell can improve tightness because it reduces within-cell variation; it can also hurt statistically because fewer calibration samples support each transition. MTCR preserves this distinction throughout: geometry chooses what information to retain, while the oracle determines what can be certified.

Definition 3.1 (k-Turn Certified Robustness)**.** Given an initial history $h_0 \in \mathcal{H}_+$, LLM policy π, perturbation budget ϵ, and horizon $k \in \mathbb{N}$, the k-turn certified robustness $\rho_k(h_0, \pi, \epsilon)$ is

$$\inf_{\nu \in \Pi_\epsilon} \mathbb{P}_{\pi,\nu}\left[\bigwedge_{t=1}^{k} \mathsf{Safe}(h_t) = 1 \,\middle|\, h_0\right], \tag{1}$$

where Π_ϵ contains all history-dependent user policies supported on the admissible perturbation set at every turn.

The infimum is taken over adaptive policies, not over a fixed sequence of prompts. For a safe history h, define the one-step robust safety probability

$$p(h, \epsilon) = \inf_{u \in \mathcal{B}_\epsilon(u^\star, h)} \mathbb{P}_{r \sim \pi(\cdot \mid h, u)}[\mathsf{Safe}(h') = 1], \tag{2}$$

where $h' = (h, u, r)$. A direct composition uses the weakest value over all safe histories.

Proposition 3.2 (Naive Multiplicative Bound)**.** *For any $h_0 \in \mathcal{H}_+$,*

$$\rho_k(h_0, \pi, \epsilon) \geq \left(\inf_{h \in \mathcal{H}_+} p(h, \epsilon)\right)^k = \underline{p}^{\,k}, \tag{3}$$

where $\underline{p}$ is the global one-step certificate.

Example 3.3 (Vacuity of Naive Bound)**.** With per-turn safety $p = 0.95$: $\rho_{10} \geq 0.60$, $\rho_{20} \geq 0.36$, $\rho_{50} \geq 0.08$. For safety-critical applications requiring $\rho_k \geq 0.9$, this permits at most $k \leq 2$ turns.

The perturbation set is inherited from the local oracle. Our production experiments use character-level edit distance. The composition theorem itself is independent of that choice and applies unchanged when the local oracle certifies a different perturbation model.

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.

4 Multi-Turn Certified Robustness

MTCR is a composition layer: it assumes access to valid one-step probability bounds and propagates them through a finite abstraction of conversation state. The abstraction may use semantic modes alone or cross them with safety-margin bins. We first give the general transition-matrix certificate and then develop persistence as a complementary structural guarantee.

4.1 Compositional Certification

Let $\mathcal{H}_i = \{h \in \mathcal{H}_+ : z(h) = i\}$ denote the histories assigned to cell i. For each ordered pair of cells, the one-step oracle returns a simultaneous lower bound

$$Q_{ij} \leq \inf_{h \in \mathcal{H}_i} \inf_{u \in \mathcal{B}_\epsilon(u^\star, h)} \mathbb{P}_{r \sim \pi(\cdot|h,u)}[h' \in \mathcal{H}_j]\,, \tag{4}$$

where $h' = (h, u, r)$. Because every destination cell contains only safe histories, the row sum lower-bounds one-step survival from the source cell. All entries used in one certificate must hold simultaneously.

In practice, a cell can be represented by certified anchors $a \in \mathcal{A}_i$. The local verifier lower-bounds each destination event on the neighborhood covered by an anchor, and the cell entry takes the minimum over the resulting cover. Any uncovered region is routed to the fallback cell rather than extrapolated from nearby samples. This interface makes MTCR agnostic to the underlying verifier: smoothing, deletion-based verification, exact enumeration, and future semantic certificates differ only in how they produce the one-step lower bounds.

Theorem 4.1 (Certified Transition-Matrix Bound)**.** *Let $Q \in [0,1]^{M \times M}$ satisfy Equation 4, and let $z(h_0)$ be the initial cell. Then, against any adaptive adversary,*

$$\rho_k(h_0, \pi, \epsilon) \geq \mathbf{e}_{z(h_0)}^\top Q^k \mathbf{1}. \tag{5}$$

The proof is a Bellman induction. If $v_t(j)$ lower-bounds survival for the next t turns from cell j, then $\sum_j Q_{ij} v_t(j)$ lower-bounds survival from cell i for $t{+}1$ turns. Crucially, the sum accounts for all safe successor cells; no independence between turns and no fixed mode trajectory is assumed.

Equivalently, the certificate expands over every safe abstract trajectory:

$$\mathbf{e}_{i_0}^\top Q^k \mathbf{1} = \sum_{i_1, \ldots, i_k \in [M]} Q_{i_0 i_1} Q_{i_1 i_2} \cdots Q_{i_{k-1} i_k}. \tag{6}$$

Global multiplication replaces every factor in every path by the same minimum. MTCR instead preserves both reachability and the risk profile of the states actually reached. This is the source of its advantage: a weak row affects the certificate only through paths that carry certified mass into that row.

Proposition 4.2 (Finite-Sample Validity)**.** *Let $\widehat{Q}$ be estimated from a calibration sample independent of evaluation, and suppose*

$$\mathbb{P}_{\text{cal}}\left[\widehat{Q}_{ij} \leq \inf_{h \in \mathcal{H}_i, u \in \mathcal{B}_\epsilon(u^\star, h)} \mathbb{P}(h' \in \mathcal{H}_j \mid h, u),\ \forall i, j\right] \geq 1 - \delta. \tag{7}$$

Then, with calibration confidence at least $1 - \delta$,

$$\rho_k(h_0, \pi, \epsilon) \geq \mathbf{e}_{z(h_0)}^\top \widehat{Q}^k \mathbf{1} \qquad \textit{for every reported } k. \tag{8}$$

The confidence correction is applied to the operator, not separately after each matrix multiplication. Once all entries are simultaneously valid, the same event supports every horizon derived from that operator. This avoids an additional union bound over k and makes the reported horizon curve a single jointly valid object.

Corollary 4.3 (Dominance over the Induced Scalar Bound)**.** *Let* $\underline{p}_Q = \min_i \sum_j Q_{ij}$*. Then*

$$\mathbf{e}_i^\top Q^k \mathbf{1} \geq \underline{p}_Q^k \tag{9}$$

for every initial cell i*.*

This comparison uses the scalar bound induced by the same matrix Q. A global one-step certificate estimated through a different statistical procedure can be larger than $\underline{p}_Q$; the unified certificate therefore retains both quantities. Once Q is available, propagation is inexpensive.

Proposition 4.4 (Certificate Evaluation)**.** *Equation 5 can be evaluated in* $O(M^2 k)$ *time by repeated matrix–vector multiplication and in* $O(M^3 \log k)$ *time by exponentiation.*

4.2 Safety Persistence

The partition can also retain information about proximity to failure. Let $\Delta : \mathcal{H} \to \mathbb{R}_{\geq 0}$ be a certified safety margin such that $\Delta(h) > 0$ implies $\mathsf{Safe}(h) = 1$. The margin can be supplied by a verified safety score or by a certified distance to the unsafe decision region; it need not be the representation used for clustering. Safety persistence describes how this quantity changes along a safe trajectory.

Definition 4.5 ((α, β)-Safety Persistence)**.** Fix an initial history h_0, horizon K, parameters $\alpha \in [0, 1)$ and $\beta \in [0, 1]$, and $\Delta_0 = \Delta(h_0) > 0$. For $t \leq K$, define the margin tube

$$\mathcal{G}_t(h_0, \alpha) = \{h \in \mathcal{H}_+ : \Delta(h) \geq (1 - \alpha)^t \Delta_0\}. \tag{10}$$

The system exhibits (α, β)**-safety persistence** along this tube if, for every $t < K$, every $h \in \mathcal{G}_t(h_0, \alpha)$, and every admissible user message u,

$$\mathbb{P}_{r \sim \pi(\cdot|h,u)}[h' \in \mathcal{G}_{t+1}(h_0, \alpha)] \geq \beta, \tag{11}$$

where $h' = (h, u, r)$.

The condition is local to a certified tube rather than uniform over every safe history. It can therefore exclude low-margin states that determine the global one-step minimum.

Theorem 4.6 (Degradation under Safety Persistence)**.** *Suppose the system exhibits* (α, β)*-safety persistence along the tube from* h_0 *for horizon* K*. Then, for every* $k \leq K$*,*

$$\rho_k(h_0, \pi, \epsilon) \geq \beta^k. \tag{12}$$

Moreover, with probability at least β^k*, the entire trajectory satisfies*

$$\Delta(h_t) \geq (1 - \alpha)^t \Delta_0 \qquad \textit{for every } t \leq k. \tag{13}$$

Corollary 4.7 (Certified Margin Horizon)**.** *Fix a certified margin threshold* $0 < \eta < \Delta_0$*. If* $\alpha > 0$*, then every horizon* $k \leq K$ *satisfying*

$$k \leq \left\lfloor \frac{\log(\eta/\Delta_0)}{\log(1 - \alpha)} \right\rfloor \tag{14}$$

retains margin at least η *throughout the conversation with probability at least* β^k*. For* $\alpha = 0$*, the same statement holds for every* $k \leq K$*.*

Rather than treat persistence as independent of semantic state, MTCR uses margin bins to refine the transition partition. The resulting operator records which modes remain safely separated from the boundary and which modes lose margin quickly.

Limits of composition. State-aware composition is exact when the abstraction preserves the robust Bellman update. Conversely, a one-state memoryless process has survival probability p^k, so exponential decay is unavoidable without verifiable state-dependent structure. Appendix C gives both results and their assumptions.

4.3 Unified Framework and Algorithm

Mode structure and margin persistence combine through partition refinement. Each joint cell records a semantic mode and a margin interval, and the one-step oracle certifies a transition matrix Q^{joint} over these cells. Theorem 4.1 then applies without an independence assumption.

Theorem 4.8 (Unified MTCR Certificate)**.** *Let m_0 and z_0 denote the mode and joint cell containing h_0. Let $\mathcal{E}$ be a calibration event on which $\underline{p}$, Q^{mode}, and Q^{joint} satisfy their respective one-step conditions and (α, β)-persistence holds through horizon K. On $\mathcal{E}$, simultaneously for every $k \le K$,*

$$\rho_k \ge \max\left\{\underline{p}^k, \mathbf{e}_{m_0}^\top (Q^{\text{mode}})^k \mathbf{1}, \mathbf{e}_{z_0}^\top (Q^{\text{joint}})^k \mathbf{1}, \beta^k\right\}. \tag{15}$$

If $\mathbb{P}_{\text{cal}}(\mathcal{E}) \ge 1 - \delta$, the simultaneous statement therefore holds with calibration confidence at least $1 - \delta$.

Each term lower-bounds the same trajectory event on the common calibration event, so their maximum is also valid. If the components are calibrated separately, their failure probabilities must be allocated so that the intersection event has the advertised confidence, for example by a union bound. Including the global product makes the reported MTCR certificate no weaker than the one-state baseline, even when a finer partition produces wider finite-sample confidence intervals. We do not multiply mode and persistence terms.

For M abstract cells, estimating a dense matrix uses $O(M^2 N)$ oracle samples and propagation costs $O(M^2 k)$. Sparse transition graphs reduce both terms to the number of retained edges.

5 Experiments

We evaluate MTCR on six language models and use a controlled transition process to isolate the behavior of the composition rule. The model study measures certificate quality across horizons, perturbation budgets, state granularities, and safety predicates. The controlled study tests exactness and the effect of abstraction sparsity.

5.1 Experimental Setup

We use 150 harmful prompts from AdvBench (Zou et al., 2023), balanced across violence, illegal activities, and hate speech. The primary baseline is the global product $\underline{p}^k$. The structured variants use a mode-only operator, a persistence certificate, or the full mode–margin partition; MTCR returns their maximum together with the global product.

The abstraction is fitted on 500 held-out safe conversations. Each full dialogue history is embedded and assigned to its nearest k-means center, producing the disjoint cells required by Theorem 4.1. Certified anchors cover the retained portion of each cell; histories outside this support enter the zero-outgoing-mass fallback cell. The joint abstraction further intersects each mode with safety-margin bins. One-sided Clopper–Pearson lower bounds use a Bonferroni allocation across every scalar, transition, and persistence quantity entering the reported maximum, yielding one 95% simultaneous calibration event. Empirical trajectory safety is measured independently over 100 attack trials and requires every response through turn k to remain safe.

5.2 Language-Model Results

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 $k \in \{5, 10, 15, 20\}$, static edit-distance perturbations and multi-turn escalation attacks, and mode granularities $M \in \{2, 4, 8\}$. Dialogue histories are embedded with all-MiniLM-L6-v2 and assigned to their nearest center. The default safety predicate combines refusal detection with harmful-content matching; a neural classifier provides an independent sensitivity check. Unless stated otherwise, $\epsilon = 5$, $M = 4$, and $N = 100$ samples are used for each local estimate. Appendix I.2 gives the complete protocol.

Table 1: Certified and observed trajectory safety (M=4, N=100, ϵ=5). Δ is the absolute certificate improvement and Gain is the ratio between MTCR and the global product. Within each model–horizon comparison, the stronger certified bound is bold. MTCR values are blue, improvements are green, and the mean block is shaded gray.

Model	k	Certified lower bound				Observed trajectory safety	
		Global	**MTCR**	Δ	Gain	Static	Escalation
Open-weight models							
LLaMA-2-7B-Chat	5	0.29	**0.36**	+0.07	1.24×	66%	55%
	10	0.08	**0.13**	+0.05	1.63×	49%	38%
	15	0.02	**0.05**	+0.03	2.50×	38%	28%
	20	0.01	**0.02**	+0.01	2.00×	30%	22%
Vicuna-7B	5	0.22	**0.28**	+0.06	1.27×	58%	46%
	10	0.05	**0.08**	+0.03	1.60×	40%	30%
Llama-3.2-3B	5	0.33	**0.40**	+0.07	1.21×	72%	62%
	10	0.11	**0.18**	+0.07	1.64×	56%	45%
Qwen2.5-7B-Instruct	5	0.35	**0.44**	+0.09	1.26×	76%	65%
	10	0.12	**0.22**	+0.10	1.83×	60%	49%
Hosted models							
GPT-4o	5	0.44	**0.55**	+0.11	1.25×	88%	78%
	10	0.20	**0.32**	+0.12	1.60×	73%	62%
Claude-3.5-Sonnet	5	0.42	**0.51**	+0.09	1.21×	85%	74%
	10	0.18	**0.28**	+0.10	1.56×	70%	58%
Mean (six models)	5	0.34	**0.42**	+0.08	1.24×	74.2%	63.3%
	10	0.12	**0.20**	+0.08	1.64×	58.0%	47.0%

Table 2: Sensitivity to perturbation budget on LLaMA-2-7B-Chat (M=4). Best values are bold; the gray row marks the default setting.

ϵ	**MTCR certificate**		Observed static safety	
	k=5	k=10	k=5	k=10
3	**0.49**	**0.19**	**75%**	**58%**
5	0.36	0.13	66%	49%
7	0.24	0.07	55%	37%

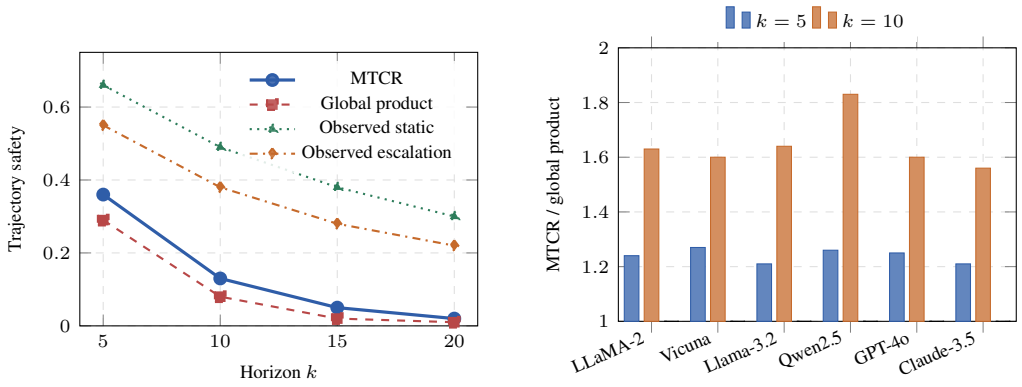
Table 1 shows a consistent horizon effect. Averaged over the six models, MTCR raises the five-turn certificate from 0.34 to 0.42 and the ten-turn certificate from 0.12 to 0.20. The mean multiplicative gain consequently increases from 1.24× to 1.64×, a 32% relative increase in the advantage of structured composition. Qwen2.5-7B-Instruct has the largest ten-turn gain (1.83×), while GPT-4o has the largest absolute improvements (+0.11 and +0.12). This pattern is exactly what the path expansion predicts: state information matters more after enough turns have accumulated for reachable and globally worst-case regions to diverge. The observed ordering of the six models is consistent across certified and empirical trajectory safety, while escalation attacks reduce observed safety by 8–12 percentage points relative to static perturbations.

We vary $M \in \{2, 4, 8\}$ on LLaMA-2-7B-Chat at k=5. The certificate is 0.39, 0.36, and 0.25, respectively. Finer partitions preserve more state information but leave fewer samples per transition entry, so simultaneous lower confidence bounds become looser. We use M=4 as the best balance for the available sample budget.

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.

Table 3: Certificate ablation on LLaMA-2-7B-Chat (M=4, ϵ=5). Best values are bold, and the proposed method is highlighted consistently in blue.

Method	k=5	k=10	k=15	k=20
Global product	0.29	0.08	0.02	0.01
Persistence only ($\hat{\beta}^k$)	0.31	0.10	0.03	0.01
Mode partition	0.34	0.11	0.04	0.01
MTCR (mode–margin)	**0.36**	**0.13**	**0.05**	**0.02**



(a) Certificate decay on LLaMA-2-7B-Chat.

(b) Relative certificate gain across models.

Figure 2: State-aware composition becomes more valuable with horizon. Left: MTCR decays more slowly than global multiplication while remaining below observed safety. Right: every model benefits at both horizons, and the gain is uniformly larger at ten turns.

Table 3 separates the information used by each certificate. Persistence uses only $\hat{\beta}$; the mode matrix uses semantic cells; MTCR uses their joint refinement. The joint state retains the strongest bound at every horizon without multiplying correlated events.

Changing the safety predicate from refusal-and-content matching to a neural harmful-content classifier leaves the ordering between certified and observed trajectory safety unchanged (Appendix F). Increasing the number of cells has a less monotone effect: it makes the state description more informative but divides the calibration budget across more transition entries. This trade-off explains why $M = 4$ outperforms $M = 8$ at the current sample size. Increasing N tightens the simultaneous lower bounds without changing the propagation rule.

5.3 Controlled Analysis

The model results combine local-oracle quality, abstraction error, and finite-sample uncertainty. We isolate the composition step with a controlled substochastic process whose abstract transition kernel is known. The matrix recursion matches the exact survival probability in this setting, as predicted by Theorem C.1. We then estimate the same operator from finite samples to measure how transition sparsity and partition size affect the certificate.

Figure 2 visualizes the same horizon effect on a production model. The certificate remains below both observed curves while decaying more slowly than global multiplication. Table 4 isolates the role of transition sparsity in the controlled process. Sparse dynamics become increasingly valuable at long horizons because the trajectory does not repeatedly visit the globally weakest cell. Dense estimated operators can be weaker at the available sample size; taking the maximum in Theorem 4.8 prevents this component from reducing the final certificate.

Table 4: Mode-only to global-product ratio in the controlled process. Column maxima are bold; green entries exceed one and therefore improve on the global product. MTCR takes the maximum of valid component certificates.

Transitions	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

6 Discussion

The gap between observed safety and the certificate has three sources: conservatism in the local oracle, the weakest-history reduction within each cell, and simultaneous finite-sample correction of the transition entries. They require different remedies. A stronger verifier improves both scalar and structured bounds, a more coherent partition reduces abstraction loss, and additional calibration data tightens the operator. Treating the gap as a single model-level failure would hide this distinction.

Partition design is therefore a statistical decision, not a search for the richest representation. Coarse cells pool evidence but merge different transition behaviors; fine cells preserve reachability and margin information but divide the sample budget across more entries. At $k = 5$, the certificates for $M \in \{2, 4, 8\}$ are 0.39, 0.36, and 0.25: the finest partition loses because its confidence penalty exceeds its structural gain. Refinement should prioritize frequently reached cells and split them only when reduced within-cell variation outweighs increased estimation uncertainty.

The matrix also exposes where uncertainty matters. The deficit $1 - \sum_j Q_{ij}$ is mass not certified as entering a safe successor, while $v_t(i)$ distinguishes cells with similar row sums but different long-horizon futures. Combined with reachable occupancy, these quantities identify where new verification samples have the highest value. They can also support conservative runtime actions when the current cell has low continuation value, without extending the certified event.

The guarantee is modular but not threat-model free. Here the oracle certifies character-level edits, whereas semantic escalation changes the reference message and is evaluated separately. A verifier covering paraphrase, topic drift, or persuasion could replace the current oracle without changing Bellman composition. Every value is also relative to an explicit safety predicate and simultaneous confidence event; the detector study preserves the composition advantage but changes the numbers because it changes the failure event.

Coverage is conservative: histories outside the certified anchor cover enter a zero-outgoing-mass fallback cell. Unsupported behavior can therefore weaken the bound but cannot inherit evidence from nearby embeddings. The finite prompt set, two predicates, $N = 100$ local trials, and template-based escalation delimit the reported numbers rather than the composition theorem. Broader guarantees require a larger certified cover, a stronger perturbation model, and calibration effort targeted by reachable transition structure.

7 Conclusion

MTCR turns valid one-step evidence into a certificate for an adaptive conversation. The key operation is not another local defense but state-aware composition: certified safe mass is propagated through reachable mode–margin cells instead of being collapsed to one worst-case scalar at every turn. This yields tighter bounds across six language models, increasingly so at longer horizons, while retaining a precise fallback when finite-sample partitions are too weak. The recursion is exact when the abstraction preserves the robust Bellman update. MTCR thus provides a clean interface through which stronger local verifiers can immediately produce stronger trajectory-level guarantees.

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Mao Ye, Chengyue Gong, and Qiang Liu. 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*, pp. 3465–3475, 2020.

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

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 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)*, pp. 14322–14350, 2024.

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

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

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

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

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

APPENDIX

A FORMAL DEFINITIONS AND ASSUMPTIONS

A.1 DIALOGUE PROCESS AND ABSTRACTION

Let $\mathcal{V}$ be a finite vocabulary and let $\mathcal{H}$ contain all finite dialogue histories. The process state is the full history $h_t = (u_1, r_1, \ldots, u_t, r_t)$. A representation $\psi : \mathcal{H} \to \mathbb{R}^d$ constructs the finite abstraction but is not assumed to be a sufficient statistic for response generation. A deterministic map $z : \mathcal{H}_+ \to [M]$ assigns each safe history to one abstract cell.

A.2 Perturbation and Agents

For a reference message $u^\star$ and metric d, the perturbation set is $\mathcal{B}_\epsilon(u^\star, h) = \{u : d(u, u^\star) \le \epsilon\}$. An adversarial policy ν maps the observed history to a distribution supported on this set. The class Π_ϵ contains all such history-dependent policies. The language-model policy $\pi(\cdot \mid h, u)$ generates the next response.

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

The next state is the concatenated history $h' = (h, u, r)$. The predicate $\mathsf{Safe} : \mathcal{H} \to \{0, 1\}$ defines $\mathcal{H}_+ = \{h : \mathsf{Safe}(h) = 1\}$ and $\mathcal{H}_- = \mathcal{H} \setminus \mathcal{H}_+$. A margin $\Delta : \mathcal{H} \to \mathbb{R}_{\ge 0}$ is safety-valid when $\Delta(h) > 0$ implies $h \in \mathcal{H}_+$. The single-turn robust safety probability is

$$p(h, \epsilon) = \inf_{u \in \mathcal{B}_\epsilon(u^\star, h)} \mathbb{P}_{r \sim \pi(\cdot|h,u)}[\mathsf{Safe}(h') = 1], \tag{16}$$

where $h' = (h, u, r)$.

A.4 Conversational Modes and Decomposition

A mode decomposition is a partition $\mathcal{H}_+ = \biguplus_{i=1}^{M} \mathcal{H}_i$, where $\mathcal{H}_i = \{h : z(h) = i\}$. Deterministic tie breaking handles cluster boundaries, and uncovered histories are routed to a fallback cell.

A.5 Intra-Mode and Transition Safety

For every ordered pair of cells, define the robust safe transition mass

$$q^\star_{ij} = \inf_{h \in \mathcal{H}_i} \inf_{u \in \mathcal{B}_\epsilon(u^\star, h)} \mathbb{P}_{r \sim \pi(\cdot|h,u)}[(h, u, r) \in \mathcal{H}_j]. \tag{17}$$

The certificate uses any simultaneous lower bounds $0 \le Q_{ij} \le q^\star_{ij}$. Such a matrix is substochastic because, for every fixed source history and input, the safe destination events are disjoint. Missing row mass represents transitions that cannot be certified as safe.

A.6 Mode Trajectory and Feasible Set

A safe trajectory induces a sequence of partition cells. MTCR does not select a single worst path; the matrix recursion sums certified probability mass over all safe successor cells at every step.

A.7 Remarks

Definition 3.1 allows the adversary to condition each message on the complete transcript. The partition is deterministic, so every safe successor contributes to exactly one destination cell. For (α, β)-persistence, α controls margin contraction and β lower-bounds the probability of the joint safety-and-margin event.

B Certificate Computation

After constructing $\hat{p}$, $\hat{Q}^{\mathrm{mode}}$, $\hat{Q}^{\mathrm{joint}}$, and $\hat{\beta}$ on one simultaneous calibration event, initialize both continuation vectors to $\mathbf{1}$. Repeated multiplication gives

$$v^{\mathrm{mode}}_t = \hat{Q}^{\mathrm{mode}} v^{\mathrm{mode}}_{t-1}, \qquad v^{\mathrm{joint}}_t = \hat{Q}^{\mathrm{joint}} v^{\mathrm{joint}}_{t-1}. \tag{18}$$

For an initial history h_0 and every $k \le K$, the reported certificate is

$$\hat{\rho}_k = \max\left\{\hat{p}^k, v^{\mathrm{mode}}_k(m(h_0)), v^{\mathrm{joint}}_k(z(h_0)), \hat{\beta}^k\right\}. \tag{19}$$

C Structural Limits of Composition

Theorem C.1 (Exactness for a Robust Markov Abstraction)**.** *Suppose there is a substochastic matrix P such that, for every nonnegative continuation vector v, every cell i, and every history $h \in \mathcal{H}_i$,*

$$\inf_{u \in \mathcal{B}_\epsilon(u^\star, h)} \mathbb{E}\left[\sum_{j=1}^{M} v_j \mathbf{1}\{h' \in \mathcal{H}_j\} \,\middle|\, h, u \right] = (Pv)_i. \tag{20}$$

Then the partition is Bellman-sufficient: for every $h_0 \in \mathcal{H}_+$, Equation 5 with $Q = P$ equals $\rho_k(h_0, \pi, \epsilon)$.

Corollary C.2 (Optimality at the Abstraction Level)**.** *Under the conditions of Theorem C.1, no method using only the same abstract kernel can return a larger valid lower bound.*

The condition is stronger than taking an entrywise infimum separately for each destination cell: the same robust Bellman update must be represented by P for every continuation vector. This distinction is necessary because different destination-wise infima may be attained by different adversarial inputs.

Theorem C.3 (Exponential Decay Is Unavoidable in General)**.** *There exist conversational systems with*

$$\rho_k(h_0, \pi, \epsilon) = p^k \tag{21}$$

for some $p < 1$, and the global multiplicative certificate is exact.

A one-state process that survives each turn with probability p, independently of the history and admissible input, gives the construction. Any slower decay therefore requires additional certified structure rather than a different algebraic manipulation of the same scalar one-step bound.

D Detailed Proofs

Let $v_t(i)$ lower-bound the probability of remaining safe for the next t turns from every history in cell i. By definition, $v_0 = \mathbf{1}$. If v_t is valid, then for every $h \in \mathcal{H}_i$ and admissible u,

$$\mathbb{P}(\text{safe for } t{+}1 \text{ turns} \mid h, u) \geq \sum_j Q_{ij} v_t(j). \tag{22}$$

Taking the infimum over h and u gives $v_{t+1} \geq Q v_t$; induction yields Theorem 4.1. Since $Q\mathbf{1} \geq \underline{p}_Q \mathbf{1}$ and Q is nonnegative, repeated multiplication proves Corollary 4.3. Proposition 4.4 follows from the matrix–vector recurrence.

For Proposition 4.2, condition on the calibration event on which every entry of $\widehat{Q}$ satisfies Equation 4. This event has probability at least $1 - \delta$ and depends only on the calibration sample. Theorem 4.1 then applies to $\widehat{Q}$ deterministically. Because the same entrywise event supports every power $\widehat{Q}^k$, no additional confidence correction is required for reporting multiple horizons from the same operator.

For persistence, let $A_t = \{h_t \in \mathcal{G}_t(h_0, \alpha)\}$. On $A_1 \cap \cdots \cap A_{t-1}$, the current history lies in $\mathcal{G}_{t-1}(h_0, \alpha)$; Definition 4.5 therefore gives $\mathbb{P}(A_t \mid A_1 \cap \cdots \cap A_{t-1}) \geq \beta$ under every adaptive policy. The chain rule yields $\mathbb{P}(\cap_{t=1}^k A_t) \geq \beta^k$. Membership in each tube is exactly the inequality $\Delta(h_t) \geq (1-\alpha)^t \Delta_0$, which proves Theorem 4.6; the threshold calculation then gives Corollary 4.7.

For Theorem C.1, the stated condition makes the robust Bellman update identical for every history in a cell and equal to multiplication by P. Starting from the cell-constant terminal vector $\mathbf{1}$, induction therefore gives the exact worst-case survival probability for every initial history. The memoryless construction in Theorem C.3 has one safe state with one-step survival p, yielding p^k. For Theorem 4.8, condition on the common calibration event. Each component then lower-bounds the same trajectory event for every $k \leq K$, so their pointwise maximum is valid simultaneously over those horizons.

E State Abstraction and Persistence Estimation

We use a deterministic geometric partition of dialogue embeddings.

Given histories $\{h^{(i)}\}_{i=1}^{n}$, we compute $\psi(h^{(i)})$, fit k-means centers, and assign every history to its nearest center with deterministic tie breaking. Joint MTCR subdivides each mode by safety-margin intervals. Any history outside the calibration support is assigned to a fallback cell whose outgoing certificate is zero unless separately verified.

For a covered cell, certification has two distinct layers. First, the local verifier supplies an oracle-specific robust transformation that is valid for every admissible edit around each anchor; this is the step that converts nominal trials into a bound uniform over $u \in \mathcal{B}_\epsilon$. Second, a one-sided Clopper–Pearson interval accounts for finite response sampling at that anchor. We take the minimum over an anchor cover of the source cell and set uncovered regions to zero. Clopper–Pearson intervals therefore quantify sampling uncertainty only; the robust verifier and the certified cover provide the required uniformity over inputs and histories.

E.1 Persistence Parameter Estimation

Persistence parameters are estimated separately within each mode.

We evaluate a prespecified grid $\mathcal{A} \subset [0, 1)$ of contraction values. For each mode and tube level t, we draw N_p certified one-step trials from histories satisfying $\Delta(h) \geq (1 - \alpha)^t \Delta_0$. The success indicator for $\alpha \in \mathcal{A}$ is

$$Y_i(\alpha) = \mathbf{1}\left[\mathsf{Safe}(h_i') = 1 \ \wedge \ \Delta(h_i') \geq (1 - \alpha)^{t+1} \Delta_0\right]. \tag{23}$$

For every mode, tube level, and α, we compute a one-sided Clopper–Pearson lower bound on $\mathbb{E}[Y_i(\alpha)]$. The confidence allocation is simultaneous over this grid and over the scalar and transition quantities used in the final maximum. For a fixed α, $\hat{\beta}$ is the minimum corrected lower bound over all reachable mode–level cells. The reported pair maximizes the resulting persistence certificate only after this joint correction. We use $N_p = 100$ trials per cell.

F Safety Detector Comparison

We compare two safety predicates. The default predicate combines refusal patterns with a harmful-content lexicon: a refusal is safe, while a non-refusal containing harmful procedural content is unsafe. The neural predicate thresholds the score of a pretrained harmful-content classifier at 0.5. Each predicate defines a separate certification problem, and all local bounds and trajectory evaluations are recomputed under that predicate. Table 5 shows that the main conclusion is unchanged.

Table 5: Sensitivity to the safety predicate on LLaMA-2-7B-Chat ($k = 5$, $\epsilon = 5$). Column maxima are bold; because the predicates define different events, they should not be read as a universal ranking. The gray row marks the default predicate.

Safety predicate	MTCR	Static	Escalation
Refusal + content	**0.36**	66%	55%
Neural classifier	0.33	**70%**	**59%**

G Illustrative Evaluation Example

We illustrate the data format, perturbation procedure, and multi-turn escalation 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.

Multi-turn escalation. For each harmful category, ten templates progress from benign context to the target request over five turns, following the escalation structure of Crescendo (Russinovich et al., 2025). A representative sequence asks first about home improvement, then about security systems, residential locks, lock-maintenance tools, and finally requests lock-picking instructions. Character-level perturbations are applied to each message. For horizons above five, neutral context turns precede the escalation sequence. A trajectory is safe only when every response is labeled safe.

Mode assignment. The representation $\psi(h) \in \mathbb{R}^d$ maps each history to its nearest k-means center. For example, a benign information request and its response typically remain in a general-inquiry cell, while a topic shift may move the history to another cell. The transition certificate is computed for every ordered source–destination pair, including a fallback cell for unsupported histories.

H Controlled Experiments

The controlled study targets three questions: when state-aware composition improves over a scalar bound, how persistence parameters affect the certified horizon, and how abstraction granularity interacts with finite-sample uncertainty. A final exact-kernel setting verifies the tightness statement in Corollary C.2.

H.1 Data Generation and Parametric Model

Synthetic representations are generated from a four-component Gaussian mixture with $n = 2{,}000$ states, Dirichlet mixing weights, centers scaled by 2.0, and per-cluster standard deviation 0.5. Nearest-center assignment produces a disjoint four-cell abstraction.

The parametric process samples safe transitions from a specified substochastic matrix. This controlled setting isolates the effect of state abstraction and permits direct comparison between the matrix certificate and observed trajectory safety.

H.2 Compositional Gain

Let τ denote the number of nonzero safe destinations retained in each source row. The ratios in Table 4 show that sparse transition structure becomes increasingly valuable at longer horizons. Ratios below one occur when entrywise confidence penalties outweigh the additional state information; the unified certificate then selects the separately estimated global product.

H.3 Persistence Scaling

At $k = 100$ and $\beta = 0.98$, the persistence certificate β^k is more than $21\times$ the global product obtained from a one-step bound of 0.95. The associated value $\alpha = 0.2$ controls the certified margin retained along those safe trajectories.

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

Increasing M trades a more informative state description for fewer samples per transition entry. In the empirical validation setting, observed trajectory safety is 94% under static perturbations and 90% under template escalation, compared with a certified value of 0.67. The structured certificate improves by $1.86\times$ over global multiplication. In the exact synthetic abstraction, it matches the abstract-process survival probability, as predicted by Corollary C.2.

I Additional Experimental Details

We fix the base random seed to 42 and use dataset seeds 42, 142, 242, 342, and 442 for data generation and mode discovery.

I.1 Hyperparameter Summary

Table 6 lists the experimental parameters. The default perturbation budget is $\epsilon = 5$ edit operations and the initial synthetic margin is $\Delta_0 = 1.0$.

Table 6: Experimental configurations.

Experiment	Parameter	Value
Composition	k	$\{5, 10, 20, 50\}$
	τ	$\{1, 2, 3, 4, 5\}$
Persistence	(α, β)	$(0.05, 0.98)$, $(0.1, 0.95)$, $(0.2, 0.98)$
	k	$\{5, 10, 20, 50, 100\}$
Granularity	M	$\{2, 4, 8, 16\}$
	k, d	20, 64
Validation	k, trials	10, 100
	Attack types	Static edit-distance, template escalation
Language models	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, representation	$\{3, 5, 7\}$, all-MiniLM-L6-v2

I.2 Production Model Experiments

The study includes four open-weight models—LLaMA-2-7B-Chat, Vicuna-7B, Llama-3.2-3B, and Qwen2.5-7B-Instruct—and two hosted models, GPT-4o and Claude-3.5-Sonnet. Every model is evaluated at the same horizons and with the same trajectory-level success criterion.

We use 150 harmful prompts from AdvBench (Zou et al., 2023), balanced across violence, illegal activities, and hate speech. Histories are represented with all-MiniLM-L6-v2; nearest-center assignment with $M \in \{2, 4, 8\}$ produces disjoint modes. Each local estimate uses $N = 100$ response trials. The oracle-specific robust transformation handles every edit in $\mathcal{B}_\epsilon$, while simultaneous one-sided lower confidence bounds account for finite response sampling. We vary $\epsilon \in \{3, 5, 7\}$. The refusal-and-content predicate is the default, and the neural predicate is used in the sensitivity study. Static attacks use the certified edit-distance set; template escalation changes the reference messages across turns.

I.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 $\mathrm{Dir}(\mathbf{1}_M)$. 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 synthetic process samples transitions from the matrix specified for each experiment. Static perturbations stay within the edit-distance set at every turn, while the escalation sequence changes the reference input as the conversation approaches a harmful target.