# ENABLING CERTIFIED ROBUSTNESS FOR LLMS AT LARGE PERTURBATION RADII VIA CORRUPTION-RATE RANDOMIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The widespread deployment of large language models (LLMs) makes jailbreak attacks that induce harmful outputs a critical safety concern. Unlike empirical defenses, randomized smoothing (RS) provides robustness guarantees against all attacks within a specified perturbation budget, but its application to LLMs often certifies only a few token modifications despite extensive sampling. We identify a fundamental bottleneck in widely used RS mechanisms based on fixed-probability masking and deletion: even when every sample receives the target label, the certifiable radius is subject to a ceiling that grows at most logarithmically with the sampling budget, severely limiting certification against large perturbations. To address this bottleneck, our method extends randomization from choosing which tokens to corrupt to choosing the corruption rate itself. By mixing lightly and heavily perturbed inputs while preserving the expected corruption rate, it increases the probability of removing multiple attack-related tokens together. We show that this change can reduce the sampling requirements for certification from exponential to polynomial in the target radius. Experiments on LLM-based harmful-content detection demonstrate substantially larger certified suffix lengths at the same sampling budget, showing that our method extends certified robustness to larger perturbation radii.

## 1 INTRODUCTION

Jailbreak attacks threaten the safety of large language models (LLMs) by causing them to produce harmful responses to requests they should refuse. For example, attackers can search for a string to append to a harmful request, or repeatedly rewrite the request, until the model follows the harmful instruction instead of rejecting it (Zou et al., 2023; Jia et al., 2024; Chao et al., 2023). To counter these attacks, researchers have proposed various empirical defenses (Jain et al., 2023; Robey et al., 2023; Zou et al., 2024; Ji et al., 2025; Cunningham et al., 2026). However, such defenses lack guarantees against all possible attacks within a specified perturbation budget, motivating us to investigate defense mechanisms that can provide theoretical robustness guarantees.

Certified defenses address this need by guaranteeing that a prediction remains unchanged under every adversarial modification within a specified budget. Randomized smoothing (RS) provides a general framework for constructing such defenses by aggregating predictions on randomly perturbed inputs (Lécuyer et al., 2019; Cohen et al., 2019). Its key insight is to use a theoretically analyzable perturbation procedure to erase the differences between an original input and its attacked version. For example, an LLM may normally refuse a harmful request but produce a harmful response after an attacker modifies it. The random perturbations introduced by RS can remove these modifications, leaving the model with the same perturbed text whether or not the attack was applied. On these shared inputs, the attack no longer affects the model's safety decision. The analyzable perturbation distribution introduced by RS allows us to bound the worst-case attack's influence on the aggregated prediction, yielding a provable budget within which the safety decision remains unchanged.

Existing RS methods applicable to LLMs typically perturb each token independently with a fixed probability by masking it, substituting it with another token, or deleting it (Huang et al., 2024; Chen et al., 2025a). However, this fixed-rate design imposes a sampling bottleneck that can restrict



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.

certification to only a few adversarial token modifications, even when predictions are aggregated over many perturbed inputs. For example, the fixed-rate masking-based RS mechanism removes the differences introduced by an attacker only if all modified tokens are masked together, so that the original and attacked inputs yield identical perturbed text. For an attack modifying $r$ tokens, this joint removal occurs with probability $\mu^r$, where $\mu$ is the fixed probability of independently masking each token. The probability of eliminating the attacker's changes therefore decreases exponentially with the attack budget, and certifying larger attacks requires exponentially more samples even if the detector correctly classifies every sampled input. More concretely, increasing the sample count from 10000 to 100000 at a masking rate of 0.1 raises the maximum certified budget from only **three** to **four** token modifications (Section 3.3). Although increasing the masking rate $\mu$ allows larger certified budgets, it also removes more of the content needed to recognize a harmful request. The central challenge is therefore to increase the probability of jointly removing adversarial changes while preserving the information required for safety detection.

The core challenge is to increase the probability of removing many adversarial changes together while preserving information for detection. Under fixed-rate masking, each token is masked with probability $\mu$, so jointly removing $r$ attacker-modified tokens has probability $\mu^r$. Increasing this exponentially small probability requires a higher masking rate, which also removes more useful content. In this paper, we propose a new RS mechanism that randomizes the corruption rate to address this problem. For each perturbed input, we draw a corruption rate $Q$ from a Beta distribution and independently perturb each token with probability $Q$, then aggregate the detector's predictions as in standard RS. This produces a mixture of heavily perturbed samples that help eliminate adversarial changes and lightly perturbed samples that retain information for detection. For example, choosing $Q \sim \mathrm{Beta}(\mu/(1-\mu), 1)$ preserves the expected corruption rate, $\mathbb{E}[Q] = \mu$, but changes the joint-removal probability from $\mu^r$ to $\mathbb{E}[Q^r] = \mu/[\mu + (1-\mu)r]$, which decays polynomially rather than exponentially with $r$. Randomizing the rate can therefore reduce the sampling requirement from exponential to polynomial without increasing the expected corruption rate, enabling certification against substantially larger attacks with the same sampling budget.

Our main contributions are summarized as follows. We first investigate the sampling bottleneck of fixed-rate smoothing and establish theoretical results showing when randomizing corruption strength reduces the sample requirement from exponential to polynomial in the target radius. We then derive certification methods for masking, uniform token substitution, and deletion, extending this design to both token replacements and variable-length edits without modifying the downstream detector. Finally, experiments on LLM-based harmful-content detection demonstrate substantially larger certified radii at matched sampling budgets, with gains observed across multiple detectors and attack settings.

## 2 Preliminaries and Related Work

### 2.1 Related Work

LLMs face a broad range of jailbreak and prompt-injection attacks (Zou et al., 2023; Chao et al., 2023; Liu et al., 2025; Greshake et al., 2023). Empirical defenses mitigate these threats through inference-time processing and post-training (Jain et al., 2023; Robey et al., 2023; Zou et al., 2024; Chen et al., 2025b; Cunningham et al., 2026), but do not guarantee robustness against all attacks within a prescribed perturbation budget. Certified defenses establish safety guarantees over a perturbation set, making the permitted edits rather than a particular attack procedure the basis of analysis. Randomized smoothing provides a model-agnostic means of obtaining such guarantees, with synonym-based formulations in text classification certifying replacements within predefined sets (Ye et al., 2020). Masking and uniform-substitution kernels lift this lexical restriction, allowing certification against arbitrary token replacements (Zeng et al., 2023; Chen et al., 2025a), with tight knapsack-based bounds available for their independent, fixed-rate forms (Chen et al., 2025a). These replacement-based certificates do not directly extend to general edit-distance attacks, where insertions and deletions change sequence length and alignment. Deletion-based procedures accommodate these changes through subsequences, which erase-and-check enumerates to recover harmful requests obscured by adversarial insertions (Kumar et al., 2024). This enumeration incurs combinatorial cost when inserted tokens occupy arbitrary positions. Its guarantee is also one-sided: it preserves the safety filter's detection of harmful requests without certifying that benign requests remain accepted under

attack. Randomized deletion avoids exhaustive enumeration by sampling subsequences and supports smoothing certificates against insertions, deletions, and substitutions (Huang et al., 2023; 2024; 2025). Across these smoothing procedures, however, broader edit coverage does not resolve the statistical cost of certifying larger attack budgets. Large-budget certification therefore remains costly: smoothing-based defenses retain fixed per-input perturbation levels, leaving sampling requirements sensitive to rapidly shrinking distributional overlap (Ye et al., 2020; Zeng et al., 2023; Chen et al., 2025a; Huang et al., 2024; 2025), whereas erase-and-check incurs combinatorial cost for arbitrary-position insertions and provides only one-sided guarantees (Kumar et al., 2024).

### 2.2 Problem Setup

We consider a smoothed detector that aggregates a base detector's decisions over randomly corrupted inputs, and seek to certify its robustness to token-level attacks. Let $x = (x_1, \ldots, x_n)$ denote the input text, where $x_i$ is the token at position $i$ and $n$ is its length. Applying a corruption mechanism to $x$ yields a random text $Z$. For a realized text $z$, let $f(z) \in [0, 1]$ be the probability that the base detector assigns it a chosen target label. If this process is deterministic, $f(z)$ is simply zero or one. The target-label probability under random corruption is

$$p(x) = \mathbb{E}[f(Z)], \tag{1}$$

where the expectation is taken over the corrupted variants of $x$. The smoothed detector assigns $x$ the target label when $p(x) \geq \tau$, with $\tau \in (0, 1)$ denoting a prescribed decision threshold. Certification establishes whether this decision remains unchanged under every modification permitted by the threat model.

To formalize the scope of this guarantee, let $\mathcal{B}_r(x)$ denote the set of inputs obtainable from $x$ under an attack budget $r$. For substitution attacks, this set allows replacements at up to $r$ token positions. When insertions and deletions are also allowed, it contains inputs within $r$ token edits in total. Beyond the permitted edit operations, the threat model specifies which part of the input the attacker can modify. For example, an attacker may be restricted to editing a suffix appended to a fixed request. In this setting, $x$ denotes the editable suffix, while the unchanged request prefix accompanies every corrupted suffix supplied to the detector.

We estimate $p(x)$ from $N$ independently corrupted samples. If $S$ samples receive the target label, these counts give a lower confidence bound $\underline{p}$ on $p(x)$ at confidence level $1 - \alpha$, where $\alpha \in (0, 1)$ is the confidence error probability. This bound accounts for finite sampling, so even unanimous predictions do not establish a target-label probability of one. Appendix B.2 gives the calculation.

To extend this statistical guarantee to adversarially modified inputs, we bound the largest decrease in target-label probability over all modifications within the prescribed attack budget. Specifically, a certificate provides a lower bound $L_r(\underline{p})$ such that, whenever $p(x) \geq \underline{p}$,

$$p(x') \geq L_r(\underline{p}) \qquad \text{for every } x' \in \mathcal{B}_r(x). \tag{2}$$

Whenever $L_r(\underline{p}) \geq \tau$, the smoothed detector therefore assigns the target label to every input in $\mathcal{B}_r(x)$. This certifies robustness against all attacks within the prescribed budget from samples of the original input alone, without evaluating the attacks individually.

## 3 Our Method

### 3.1 Proposed Procedure

The key step of our method is to vary corruption strength across samples while preserving the expected corruption rate of the corresponding fixed-rate mechanism. First, we generate $N$ independently perturbed inputs. For each sample, we draw $Q \sim \mathrm{Beta}(a, b)$ with $a = \mu b/(1 - \mu)$, where $\mu \in (0, 1)$ is the expected corruption rate and $b > 0$ controls the spread. Conditional on $Q$, each editable token is independently selected with probability $Q$ for masking, uniform replacement by a different vocabulary token, or deletion. The mask symbol lies outside the vocabulary, and any protected prefix remains unchanged.

Then, we evaluate these $N$ samples with the detector, obtaining $S$ target-label decisions. We use these counts to obtain the lower confidence bound $\underline{p}$ described in Section 2.2. Finally, we use

the fractional-knapsack framework from prior work (Chen et al., 2025a) to bound the target-label probability after an attack, with kernel probabilities averaged over $Q$. The largest radius whose resulting bound reaches $\tau$ is certified without further detector evaluations.

### 3.2 Certified Radius Computation

**Theorem 1** (Certified radius for randomized-rate smoothing)**.** *For our proposed mechanism, let* $c_r = \mathbb{E}[Q^r] = \mathrm{B}(a+r,b)/\mathrm{B}(a,b)$*, where* $\mathrm{B}(u,v) = \int_0^1 q^{u-1}(1-q)^{v-1}\,dq$ *is the beta function. Let* $\mathcal{L}(P,P';\underline{p})$ *denote the lower bound on the attacked input's target-label probability, computed below from group probabilities* $P, P'$ *under the original and attacked inputs. The three kernels admit the bounds*

$$\begin{aligned} L_r^{\mathrm{mask}}(\underline{p}) &= [\underline{p} + c_r - 1]_+, \\ L_r^{\mathrm{uniform}}(\underline{p}) &= \min_{0\le d\le \min(r,n)} \mathcal{L}(P_d, P'_d; \underline{p}), \\ L_r^{\mathrm{del}}(\underline{p}) &= \min_{\substack{s,i,j\ge 0,\ s+i+j\le r \\ s+j\le n}} \mathcal{L}(P_{s,i,j}, P'_{s,i,j}; \underline{p}), \end{aligned} \tag{3}$$

*where* $[u]_+ = \max(0,u)$*,* $n$ *is the editable input length,* $d$ *is the substitution distance, and* $(s,i,j)$ *counts substitutions, insertions, and deletions in an edit alignment. For group probabilities* $P_g, P'_g$ *and group target-label probabilities* $u_g \in [0,1]$*, we define this lower bound as*

$$\mathcal{L}(P,P';\underline{p}) = \min_{0\le u_g\le 1} \left\{ \sum_g P'_g u_g : \sum_g P_g u_g \ge \underline{p} \right\}. \tag{4}$$

*For uniform substitution with vocabulary size* $V \ge 3$*, an output differing from a length-*$n$ *input at* $t$ *positions has probability*

$$m_n(t) = \frac{\mathrm{B}(a+t, b+n-t)}{\mathrm{B}(a,b)(V-1)^t}. \tag{5}$$

*For inputs differing at* $d$ *positions, let* $i,j$ *count an output's mismatches to the original and attacked tokens within those positions, and let* $k$ *count mismatches elsewhere. The indices are integers satisfying* $0 \le i,j \le d$*,* $0 \le k \le n-d$*, and* $h = i+j-d \ge 0$*, where* $h$ *counts positions matching neither token. The groups contain* $v_{ijk}$ *outputs and have masses*

$$\begin{aligned} v_{ijk} &= \frac{d!\,(V-2)^h}{(d-i)!\,(d-j)!\,h!} \binom{n-d}{k} (V-1)^k, \\ P_{ijk} &= v_{ijk} m_n(i+k), \qquad P'_{ijk} = v_{ijk} m_n(j+k). \end{aligned} \tag{6}$$

*These masses define* $P_d = (P_{ijk})$ *and* $P'_d = (P'_{ijk})$*.*

*For deletion under an alignment* $(s,i,j)$*, the attacked length is* $m = n-j+i$ *and the number of matched occurrences is* $\ell = n-s-j$*. For integer* $k$*, the probability* $\rho_n(k)$ *of a specified pattern retaining* $k$ *occurrences, and the group masses for retaining only matched occurrences, are*

$$\begin{aligned} \rho_n(k) &= \frac{\mathrm{B}(a+n-k, b+k)}{\mathrm{B}(a,b)}, \\ P_k &= \binom{\ell}{k} \rho_n(k), \qquad P'_k = \binom{\ell}{k} \rho_m(k), \quad 0 \le k \le \ell. \end{aligned} \tag{7}$$

*We include the remaining probability mass in two residual groups, with*

$$(P_{\ell+1}, P'_{\ell+1}) = \left(1 - \sum_{k=0}^{\ell} P_k, 0\right), \qquad (P_{\ell+2}, P'_{\ell+2}) = \left(0, 1 - \sum_{k=0}^{\ell} P'_k\right). \tag{8}$$

*The resulting vectors are* $P_{s,i,j} = (P_k)_{k=0}^{\ell+2}$ *and* $P'_{s,i,j} = (P'_k)_{k=0}^{\ell+2}$*.*

*For each kernel, Equation 3 bounds the target-label probability over all allowed modifications. Whenever* $p(x) \ge \underline{p}$*,*

$$p(x') \ge L_r^{\mathrm{kernel}}(\underline{p}) \qquad \text{for every } x' \in \mathcal{B}_r(x). \tag{9}$$

*Whenever $L_r^{\mathrm{kernel}}(\underline{p}) \geq \tau$, every input in $\mathcal{B}_r(x)$ therefore retains the target label. The certified radius is*

$$\widehat{r} = \max\{r \in \mathbb{Z}_{\geq 0} : L_r^{\mathrm{kernel}}(\underline{p}) \geq \tau\}, \tag{10}$$

*with $r \leq n$ for masking and uniform substitution, which certify up to $r$ token replacements. Deletion certifies up to $r$ replacements, insertions, and deletions in total.*

*Proof sketch.* Suppose the original input has target-label probability at least $\underline{p}$, and fix an allowed attacked input $x'$. This lower bound tells us how often the target label occurs, but not which corrupted texts receive it. The worst case assigns the target label to texts that are relatively common under corruption of the original input but rare under corruption of $x'$. In Equation 4, group $g$ contributes $P_g u_g$ to the original probability and $P'_g u_g$ to the attacked probability. Thus, among groups with $P_g > 0$, assigning the label in increasing order of $P'_g/P_g$ until the original probability reaches $\underline{p}$ minimizes the attacked probability. This gives the fractional-knapsack solution. The group probabilities are obtained by averaging over $Q$; taking the minimum over allowed edit patterns then gives the bounds in Equation 3. Appendix B.7 gives the full proof, including the conservative grouping used for deletion. □

### 3.3 Theoretical Advantages of Rate Randomization

To show the certification advantage of rate randomization, we compare the radii certified by fixed-rate and randomized-rate masking at the same sample count $N$ and target-label count $S$, holding the expected corruption rate $\mu$ and decision threshold $\tau$ fixed. We consider both mechanisms with the same lower confidence bound $\underline{p}(S, N)$, computed from counts $N \geq 1$ and $0 \leq S \leq N$ as described in Section 2.2. The masking certificate bounds the decrease in this probability by the chance that at least one modified position remains unmasked. To retain target-label probability at least $\tau$ after an attack, the probability of masking all modified positions must therefore reach

$$\delta_{S,N} = 1 - \underline{p}(S, N) + \tau. \tag{11}$$

Each mechanism certifies a radius when this probability reaches $\delta_{S,N}$. We assume $\underline{p}(S, N) \geq \tau$, hence $0 < \delta_{S,N} \leq 1$.

**Theorem 2** (Certified radius for fixed-rate masking)**.** *Suppose each editable token is masked independently with probability $\mu \in (0, 1)$. For any counts $(S, N)$ satisfying the condition above, the largest radius certified among $n$ editable tokens is*

$$r_{\mathrm{fixed}}(S, N) = \min\left\{n, \left\lfloor \frac{\log(\delta_{S,N}^{-1})}{\log(\mu^{-1})} \right\rfloor\right\}. \tag{12}$$

*Proof.* The fixed-rate masking certificate from prior work (Chen et al., 2025a) is $[\underline{p}(S, N) + \mu^r - 1]_+$, where $\mu^r$ is the probability of masking all $r$ modified positions. Since $\tau > 0$, certification is equivalent to

$$[\underline{p}(S, N) + \mu^r - 1]_+ \geq \tau \quad \Longleftrightarrow \quad \mu^r \geq \delta_{S,N} \quad \Longleftrightarrow \quad r \leq \frac{\log(\delta_{S,N}^{-1})}{\log(\mu^{-1})}. \tag{13}$$

Dividing by $\log \mu < 0$ reverses the inequality. Rounding down and restricting to $r \leq n$ gives Equation 12. □

**Theorem 3** (Certified radius for randomized-rate masking)**.** *Suppose $Q \sim \mathrm{Beta}(a, 1)$, with $a = \mu/(1-\mu)$ to preserve the expected corruption rate. At the same counts $(S, N)$ and threshold, the largest certified radius is*

$$r_{\mathrm{random}}(S, N) = \min\left\{n, \left\lfloor a\left(\delta_{S,N}^{-1} - 1\right)\right\rfloor\right\}. \tag{14}$$

*Proof.* Given rate $q$, masking all $r$ modified positions has probability $q^r$. Averaging against the $\mathrm{Beta}(a, 1)$ density $aq^{a-1}$ on $(0, 1)$ gives

$$c_r = \int_0^1 q^r \, aq^{a-1} \, dq = \left[\frac{a}{a+r} q^{a+r}\right]_0^1 = \frac{a}{a+r}. \tag{15}$$

The masking certificate in Theorem 1 therefore reaches $\tau$ when

$$[\underline{p}(S,N) + c_r - 1]_+ \geq \tau \quad \Longleftrightarrow \quad \frac{a}{a+r} \geq \delta_{S,N} \quad \Longleftrightarrow \quad r \leq a(\delta_{S,N}^{-1} - 1). \tag{16}$$

Rounding down and restricting to $r \leq n$ gives Equation 14. □

At the same $(S, N)$, the two radii can be compared directly:

$$\begin{aligned} r_{\text{fixed}}(S,N) &= \min\left\{n, \left\lfloor \frac{1}{\log(\mu^{-1})} \log(\delta_{S,N}^{-1}) \right\rfloor\right\}, \\ r_{\text{random}}(S,N) &= \min\left\{n, \left\lfloor \frac{\mu}{1-\mu}\left(\delta_{S,N}^{-1} - 1\right) \right\rfloor\right\}. \end{aligned} \tag{17}$$

Before rounding and the length cap, the fixed-rate bound depends logarithmically on $\delta_{S,N}^{-1}$, whereas the randomized-rate bound depends linearly on it. The integer certified radius is never smaller under randomization: for every integer $r \geq 0$, convexity gives $\mathbb{E}[Q^r] \geq (\mathbb{E}[Q])^r = \mu^r$, so every radius passing the fixed-rate test also passes the randomized-rate test. This advantage holds at any matched $(S, N)$ with a valid certificate, without requiring every prediction to have the target label. Appendix C derives the number of samples needed to certify a given radius, including polynomial bounds for general Beta masking, deletion, and uniform substitution under the stated evidence and threshold conditions.

# 4 Experiments

## 4.1 Experimental Setup

We first evaluate suffix-only and full-input attacks, then examine corruption-rate choices and benign detection. We evaluate the 50 AdvBench requests released by I-GCG (Zou et al., 2023; Jia et al., 2024) using Vicuna-7B-v1.5, Llama-2-7B-Chat, Llama-3-8B-Instruct, and Llama Guard 3-8B (Zheng et al., 2023; Touvron et al., 2023; Dubey et al., 2024; Meta, 2024). For conversational models, we follow the safety-detection protocol used in prior work (Chen et al., 2025a); Llama Guard uses its official moderation interface. We use Llama-2 by default unless cross-model comparison is needed. We compare masking and uniform substitution with their fixed-rate counterparts (Chen et al., 2025a) to isolate the effect of randomizing the corruption rate. CERT-ED (Huang et al., 2024) provides the fixed-rate deletion control; AdaptDel (Huang et al., 2025) tests adaptation to input length.

Each fixed-rate and randomized-rate pair shares the detector, expected corruption rate, and sample budget. Masking and substitution use matched diffusion-based purification (Lou et al., 2024), whereas deletion uses the retained text directly, consistent with the respective baseline pipelines (Chen et al., 2025a; Huang et al., 2024; 2025). We set AdaptDel's reference parameter to $\mu$ while retaining its length-dependent rule, which can assign higher deletion probabilities and thus favor removal of adversarial changes. Unless otherwise specified, the expected corruption rate is $\mu = 0.1$, the Beta shape parameter is $b = 1$, the confidence error probability is $\alpha = 0.01$, and the decision threshold is $\tau = 4.6 \times 10^{-5}$. Cross-model comparisons use $N = 100000$ samples per request; experiments on the method use $N = 10000$ unless otherwise stated. We report radii averaged over all 50 requests, assigning zero to uncertified inputs. Appendix D describes radius computation and detector scoring.

**Certification scenarios.** We evaluate two scenarios: suffix-only attacks that modify only appended text, and full-input attacks that may edit any token of the request. The suffix setting reflects applications that separate fixed instructions from attacker-controlled content (Liu et al., 2024; Chen et al., 2025b) and excludes direct edits to intent-defining tokens, which can change a request's safety label after only a few substitutions (Chen et al., 2025a). Full-input certification tests robustness when attackers control the request itself, covering substitutions and variable-length edits without a protected region (Chen et al., 2025a; Huang et al., 2024).

## 4.2 Evaluation on Suffix-only Attacks

**Certification at matched corruption rates.** Table 1 compares certified radii for a 50-token suffix appended to each harmful request. Rate randomization yields substantially larger certified radii across

the tested corruption rates. For masking and substitution, the 50-token suffix sets the maximum replacement radius because these kernels preserve length; Table 2 examines larger editable regions. Deletion also certifies insertions, so the number of possible edits is not bounded by the suffix length. Its certified radius increases with the corruption rate, with randomized deletion consistently exceeding both baselines. These results show that rate randomization supports larger attack budgets when the protected request preserves the information needed for detection.

| $\mu$ | Fixed mask | Our mask | Fixed subst. | Our subst. |
|---|---|---|---|---|
| 0.1 | 3 | **50** | 3 | **50** |
| 0.25 | 5 | **50** | 5 | **50** |
| 0.5 | 10 | **50** | 10 | **50** |
| 0.75 | 26.00 | **50** | 25.70 | **50** |

| $\mu$ | CERT-ED | AdaptDel | Our deletion |
|---|---|---|---|
| 0.1 | 3 | 3 | **1913** |
| 0.25 | 5 | 5 | **6742** |
| 0.5 | 10 | 10 | **21227** |
| 0.75 | 26 | 24 | **64683** |

Table 1: Mean certified radii for 50-token suffixes on Llama-2 with $N = 10000$ and $b = 1$. Left: token-replacement radii with diffusion-based purification. Right: token-edit radii without purification, covering substitutions, insertions, and deletions. Each row compares methods at corruption rate $\mu$; bold marks the largest radius within each kernel.

**Longer editable suffixes.** Table 2 extends the evaluation to longer suffixes without purification, showing that the large replacement radii are not confined to the 50-token setting. Uniform substitution certifies more replacements than masking at intermediate suffix lengths, but neither kernel improves monotonically as the editable region grows. A possible explanation is that long sequences of `[PAD]` tokens interfere with the detector's recognition of harmful content after corruption, potentially lowering the probability of an unsafe prediction and limiting the certified radius.

Table 2: Mean certified token-replacement radii for longer suffixes on Llama-2, with $\mu = 0.1$, $b = 1$, and $N = 10000$.

| Suffix length | 100 | 200 | 400 | 800 |
|---|---|---|---|---|
| Our mask | 100.00 | 57.38 | 6.48 | 37.20 |
| Our subst. | 99.84 | 191.88 | 75.42 | 2.12 |

**Repeated requests.** As adversarial prompts become longer, evaluation on AdvBench's short requests alone may conceal how certification scales with request length (Chen et al., 2025a). We therefore repeat the protected request while keeping the editable suffix fixed. Under suffix-only attacks, however, repeating the request renders the certified radii consistent with those obtained without repetition.

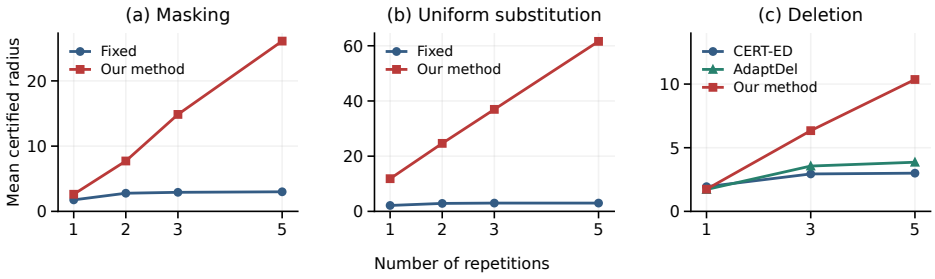


Figure 1: Mean certified radii under full-input attacks as the number of repetitions increases, using Llama-2 with $\mu = 0.1$ and $N = 10000$.

### 4.3 Evaluation on Full-Input Attacks

**Certification at matched corruption rates.** Tables 3 and 4 evaluate original, unrepeated requests across four detectors, testing whether the gains persist without protected or repeated text. Randomized

uniform substitution improves on three detectors, while masking yields a smaller improvement on average. Figure 2 further shows that randomized uniform substitution certifies a larger fraction of requests at larger target radii than its fixed-rate counterpart. In contrast, CERT-ED achieves the largest deletion radius on every detector. Together with the suffix and repetition results, this difference indicates that the realized gain depends on how well the detector recognizes harmful content after corruption.

Table 3: Mean certified token-replacement radii on unrepeated full inputs across four detectors, with $\mu = 0.1$, $N = 100000$, and matched diffusion-based purification. The final column averages the four detectors equally. Bold indicates the larger radius within each fixed-rate and randomized-rate pair, including ties.

| Method | Vicuna | Llama-2 | Llama-3 | Guard 3 | Mean |
|---|---|---|---|---|---|
| Fixed masking | 1.34 | 1.96 | **1.32** | **1.18** | 1.45 |
| Our masking | **1.60** | **2.92** | 1.12 | 0.96 | **1.65** |
| Fixed substitution | 1.68 | 2.40 | **1.26** | 1.28 | 1.66 |
| Our substitution | **5.24** | **11.94** | 0.86 | **1.46** | **4.88** |

Table 4: Mean certified token-edit radii on unrepeated full inputs across four detectors. Bold marks the largest radius for each detector.

| Method | Vicuna | Llama-2 | Llama-3 | Guard 3 |
|---|---|---|---|---|
| CERT-ED | **1.34** | **2.14** | **1.16** | **1.38** |
| AdaptDel | 1.26 | 1.78 | 1.06 | 1.24 |
| Our deletion | 1.26 | 1.84 | 0.86 | 0.82 |

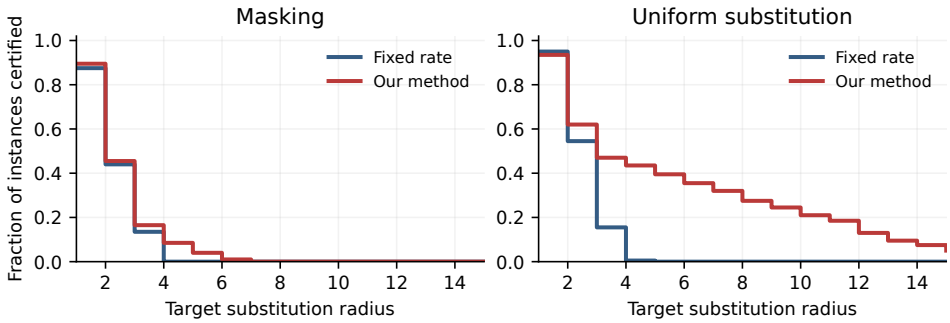


Figure 2: Certified coverage as the target substitution radius increases under full-input attacks, with $\mu = 0.1$, $N = 100000$, and diffusion-based purification.

**Repeated editable requests.** Figure 1 examines repetition when every request token is editable. The radii of all three randomized kernels increase with repetition, while the fixed-rate baselines change little.

### 4.4 Effect of the Corruption-Rate Distribution

Table 5 examines how the expected corruption rate affects full-input certification on Llama-2. Randomized substitution gives the largest gains at low rates, and masking also benefits at moderate rates. At high corruption rates, both substitution kernels certify the full input and fixed masking slightly outperforms its randomized counterpart. The advantage therefore diminishes as stronger fixed-rate corruption already yields large certificates.

To isolate the effect of distribution shape, we vary $b$ while setting $a = \mu b/(1-\mu)$ to keep the expected corruption rate fixed. Table 6 shows that reducing $b$ improves substitution and masking certificates at moderate corruption rates. Thus, allocating more probability to lightly and strongly corrupted

Table 5: Mean certified token-replacement radii at different expected corruption rates on full Llama-2 inputs, with $N = 100000$ and $b = 1$. Bold marks the larger radius within each kernel, including ties.

| $\mu$ | Fixed mask | Our mask | Fixed subst. | Our subst. |
|---|---|---|---|---|
| 0.1 | 1.96 | **2.92** | 2.40 | **11.94** |
| 0.25 | 2.14 | **3.12** | 3.20 | **12.06** |
| 0.5 | 2.74 | **3.50** | 4.90 | **12.18** |
| 0.75 | 4.00 | **4.30** | 9.28 | **12.32** |
| 0.9 | **5.28** | 5.10 | **12.32** | **12.32** |

samples improves certification without increasing expected corruption. At high corruption rates, all substitution kernels certify the full input and fixed masking remains slightly stronger, consistent with the preceding rate comparison.

Table 6: Mean certified token-replacement radii for different Beta shapes on full Llama-2 inputs, with $N = 10000$ and diffusion-based purification.

| $\mu$ | Kernel | Fixed | $b = 0.5$ | $b = 1$ | $b = 2$ | $b = 4$ |
|---|---|---|---|---|---|---|
| 0.5 | Mask | 2.60 | 3.40 | 3.36 | 3.16 | 2.84 |
| 0.5 | Subst. | 4.70 | 12.32 | 12.16 | 10.76 | 7.68 |
| 0.9 | Mask | 5.14 | 4.72 | 4.90 | 5.06 | 5.06 |
| 0.9 | Subst. | 12.32 | 12.32 | 12.32 | 12.32 | 12.32 |

### 4.5 Detection on Benign Requests

Table 7 tests whether larger certificates compromise benign detection, using 50 MT-Bench requests. Rate randomization modestly increases unsafe predictions for masking and substitution, while deletion remains comparable to or better than the baselines. Alongside the substantial radius gains, these changes support a favorable trade-off between robustness and detection performance. The frequencies concern individual corrupted samples.

Table 7: Unsafe predictions on corrupted benign requests (%; lower is better), with Llama-2, $\mu = 0.1$, and $N = 10000$. Fixed deletion denotes CERT-ED; bold marks the lowest frequency in each row.

| Kernel | Fixed rate | Our method | AdaptDel |
|---|---|---|---|
| Masking | **18.81** | 20.24 | – |
| Substitution | **25.36** | 28.35 | – |
| Deletion | 20.47 | **20.25** | 22.67 |

## 5 Conclusion

This work shows that varying corruption strength across samples can make text robustness certification more effective without increasing the expected corruption rate. Our results show that randomizing the corruption rate can certify robustness against larger attacks under the same sampling budget. These findings highlight the corruption-rate distribution as an important choice in designing certified defenses. Developing corruption mechanisms that better preserve useful information, together with models adapted to these mechanisms, is a promising direction for future work.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with experimental code development and with checking, organizing, and polishing existing proof drafts. We did not use these tools to develop the research idea or core theoretical approach, conduct the initial prototype experiments, or draft manuscript sections. AI-assisted manuscript editing was based on existing drafts. Synthetic dataset generation is not applicable to this work. Additionally, we used generative AI tools for manuscript polishing and literature search to supplement the related work.

We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023. URL https://arxiv.org/abs/2310.08419.

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*, 2025a. URL https://arxiv.org/abs/2501.19040v4.

Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. StruQ: Defending against prompt injection with structured queries. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 2383–2400, Seattle, WA, August 2025b. USENIX Association. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/chen-sizhe.

Jeremy Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320, 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

Hoagy Cunningham, Jerry Wei, Zihan Wang, Andrew Persic, Alwin Peng, Jordan Abderrachid, Raj Agarwal, Bobby Chen, Andy Dau, Alek Dimitriev, Logan Howard, Yijin Hua, Rob Gilson, Mu Lin, Christopher Liu, Vladimir Mikulik, Rohit Mittapalli, Clare O'Hara, Jin Pan, Nikhil Saxena, Alex Silverstein, Yue Song, Giulio Zhou, Jan Leike, Jared Kaplan, Ethan Perez, and Mrinank Sharma. Constitutional classifiers++: Efficient production-grade defenses against universal jailbreaks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4f697305ef1f868ad77c3c0027989a6f-Abstract-Conference.html.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *arXiv preprint arXiv:2302.12173*, 2023. URL https://arxiv.org/abs/2302.12173.

Zhuoqun Huang, Neil G Marchant, Keane Lucas, Lujo Bauer, Olga Ohrimenko, and Benjamin Rubinstein. RS-Del: Edit distance robustness certificates for sequence classifiers via randomized deletion. In *Advances in Neural Information Processing Systems*, volume 36, pp. 18676–18711, 2023. doi: 10.52202/075280-0820. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3ba82362eb0aa75487069f19fde794fe-Paper-Conference.pdf.

Zhuoqun Huang, Neil G Marchant, Olga Ohrimenko, and Benjamin I. P. Rubinstein. CERT-ED: Certifiably robust text classification for edit distance. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10813–10835, 2024. doi: 10.18653/v1/2024.findings-emnlp.635. URL https://aclanthology.org/2024.findings-emnlp.635/.

Zhuoqun Huang, Neil G. Marchant, Olga Ohrimenko, and Benjamin I. P. Rubinstein. AdaptDel: Adaptable deletion rate randomized smoothing for certified robustness. *arXiv preprint arXiv:2511.09316*, 2025. URL https://arxiv.org/abs/2511.09316.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023. URL https://arxiv.org/abs/2312.06674.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023. URL https://arxiv.org/abs/2309.00614.

Jiabao Ji, Bairu Hou, Alexander Robey, George J. Pappas, Hamed Hassani, Yang Zhang, Eric Wong, and Shiyu Chang. Defending large language models against jailbreak attacks via semantic smoothing. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 7–40, 2025. doi: 10.18653/v1/2025.ijcnlp-short.2. URL https://aclanthology.org/2025.ijcnlp-short.2/.

Xiaojun Jia, Tianyu Pang, Chao Du, Yihao Huang, Jindong Gu, Yang Liu, Xiaochun Cao, and Min Lin. Improved techniques for optimization-based jailbreaking on large language models. *arXiv preprint arXiv:2405.21018*, 2024. URL https://arxiv.org/abs/2405.21018.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. Certifying LLM safety against adversarial prompting. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2309.02705.

Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy*, 2019. URL https://arxiv.org/abs/1802.03471.

Xiaogeng Liu, Peiran Li, G. Edward Suh, Yevgeniy Vorobeychik, Zhuoqing Mao, Somesh Jha, Patrick McDaniel, Huan Sun, Bo Li, and Chaowei Xiao. AutoDAN-Turbo: A lifelong agent for strategy self-exploration to jailbreak LLMs. In *International Conference on Learning Representations*, pp. 10313–10360, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1bff3663270ba47f801e917f782d7935-Paper-Conference.pdf.

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 1831–1847, Philadelphia, PA, August 2024. USENIX Association. ISBN 978-1-939133-44-1. URL https://www.usenix.org/conference/usenixsecurity24/presentation/liu-yupei.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning*, pp. 32819–32848, 2024. URL https://proceedings.mlr.press/v235/lou24a.html.

Meta. Llama guard 3 model card, 2024. URL https://huggingface.co/meta-llama/Llama-Guard-3-8B.

Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. SmoothLLM: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023. URL https://arxiv.org/abs/2310.03684.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

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. URL https://aclanthology.org/2020.acl-main.317/.

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. URL https://aclanthology.org/2023.cl-2.5/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. URL https://arxiv.org/abs/2306.05685.

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. URL https://arxiv.org/abs/2307.15043.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In *Advances in Neural Information Processing Systems*, volume 37, pp. 83345–83373, 2024. doi: 10.52202/079017-2651. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/97ca7168c2c333df5ea61ece3b3276e1-Paper-Conference.pdf.

# A NOTATION

The notation used throughout the paper is summarized below. Symbols specific to individual derivations are defined where they first appear.

| Symbol | Meaning |
|---|---|
| $x, x', z$ | Original input, attacked input, and a randomly corrupted output. |
| $n, m$ | Lengths of the original and attacked editable sequences. |
| $f, p(x)$ | Base detector's target-label probability and its expectation over randomly corrupted inputs. |
| $\mathcal{B}_r(x), r$ | Allowed attack neighborhood and its token modification budget. |
| $Q, \mu, a, b$ | Shared random corruption rate, its expectation, and beta distribution parameters, with $a = \mu b/(1-\mu)$. |
| $M_i, c_r$ | Indicator that position $i$ is selected, and joint selection probability $c_r = \mathbb{E}[Q^r]$. |
| $N, S, \underline{p}$ | Sample count, number of target-label predictions, and lower confidence bound on $p(x)$. |
| $\alpha, \tau, L_r$ | Confidence error probability, decision threshold, and certified lower bound at radius $r$. |
| $\delta_{S,N}, \delta_N$ | Required shared mass $1 - \underline{p}(S, N) + \tau$; its all-success case is $\delta_N = \delta_{N,N} = 1 - \alpha^{1/N} + \tau$. |
| $d, s, i, j, \ell$ | Substitution distance; alignment counts of substitutions, insertions, and deletions; and matched occurrences. In uniform-substitution groups, $i, j$ instead count mismatches to the two inputs within the $d$ differing positions. |
| $P, P', P_g, P'_g$ | Original and attacked probability distributions or group vectors, with group masses $P_g, P'_g$. |
| $\mathcal{V}, V$ | Replacement vocabulary and its size $V = \lvert\mathcal{V}\rvert$. |
| $w_n(t)$ | Probability of selecting a specified set of $t$ positions among $n$ editable tokens. |
| $m_n(t), \rho_n(k)$ | Probability of one output with $t$ substitutions, and of one deletion pattern retaining $k$ occurrences. |
| $q_\star, A_V(q)$ | Input-independent uniform-substitution rate $(V-1)/V$, and the single-token affinity at rate $q$. |
| $\omega_r, \varphi_{a,b}(q)$ | Uniform-substitution overlap lower bound at radius $r$, and the density of $\mathrm{Beta}(a, b)$. |
| $\mathrm{B}, \Gamma$ | Beta and gamma functions. |

In the suffix setting, $n$ and $m$ denote the lengths of the original and attacked editable suffixes, respectively; the request prefix is fixed.

# B KERNEL DERIVATIONS AND PROOF OF THEOREM 1

Certification uses predictions on the original input to constrain predictions on inputs that have never been evaluated. The link is the probability that the corruption mechanism assigns to the same output under an original input and an attacked input. Rate randomization changes these probabilities while leaving the expected number of corrupted tokens unchanged. The derivations below identify this change, show how it determines a lower bound on the attacked input's target-label probability, and establish the three certificates in Theorem 1.

Throughout this section, $x$ denotes the original input, $x'$ an allowed modification, and $n$ the number of editable tokens in $x$. The rate follows $Q \sim \mathrm{Beta}(a, b)$ with fixed $a, b > 0$, independently of the input being corrupted. We write $f(z) \in [0, 1]$ for the target-label probability of the detector on corrupted text $z$ and $p(x)$ for its expectation under corruption of $x$. This notation includes a deterministic detector, for which $f(z)$ is zero or one, as well as fixed randomized post-processing through its conditional target-label probability. In the suffix setting, the protected prefix is attached to $z$ before evaluation.

### B.1 Shared-Rate Pattern Probabilities

Preserving the corruption probability of each token does not determine how often several tokens are corrupted together. For example, suppose that a sample uses rate $1/4$ or $3/4$ with equal probability. Each position is then selected with probability $1/2$, but two specified positions are selected together with probability $\frac{1}{2}(1/4)^2 + \frac{1}{2}(3/4)^2 = 5/16$, compared with $1/4$ under a fixed rate of $1/2$. The additional joint selection comes from sharing the rate within a sample. The Beta construction uses the same principle with a continuous range of rates.

To express this distinction for an arbitrary input, let $M_i$ indicate whether editable position $i$ is selected. Conditional on $Q = q$, the indicators $M_1, \ldots, M_n$ are independent Bernoulli variables with parameter $q$. Averaging over the rate gives

$$\mathbb{E}[M_i] = \mathbb{E}\big[\mathbb{E}[M_i \mid Q]\big] = \mathbb{E}[Q] = \mu, \qquad \mathbb{E}\left[\sum_{i=1}^{n} M_i\right] = n\mu. \tag{18}$$

Thus the expected corruption rate remains $\mu = a/(a+b)$, and the expected number of selected positions remains $n\mu$. Dependence appears only when several positions are considered together: for distinct positions, $\mathbb{E}[M_i M_j] = \mathbb{E}[Q^2]$, so their covariance is $\mathrm{Var}(Q) = \mu(1-\mu)/(a+b+1)$.

For certification, the relevant event is that every position in a specified set $A$ of $r$ modified positions is selected. Conditional independence yields

$$\mathbb{P}(M_i = 1 \text{ for every } i \in A) = \mathbb{E}\left[\prod_{i \in A} \mathbb{P}(M_i = 1 \mid Q)\right] = \mathbb{E}[Q^r]. \tag{19}$$

We denote this joint selection probability by $c_r = \mathbb{E}[Q^r]$. For integer $r \geq 1$, convexity gives $c_r \geq \mu^r$, with equality at $r = 1$. Drawing an independent rate at each position would instead give $(\mathbb{E}[Q])^r = \mu^r$. The increase therefore depends on using one shared rate, rather than merely making the rate random.

Joint selection suffices for the masking argument, but computing the other certificates also requires the probability of an entire selection pattern. A specified set of $t$ positions must be selected and all $n - t$ other positions left unchanged. Its probability is

$$\begin{aligned} w_n^{\mathrm{fixed}}(t) &= \mu^t (1-\mu)^{n-t}, \\ w_n(t) &= \mathbb{E}\big[Q^t (1-Q)^{n-t}\big] = \frac{\mathrm{B}(a+t, b+n-t)}{\mathrm{B}(a,b)}. \end{aligned} \tag{20}$$

Here $w_n(t)$ is the probability of one specified set, not the probability of selecting any $t$ positions; the latter includes a factor $\binom{n}{t}$. This distinction determines the multiplicities used below. In particular, $c_r = w_r(r)$, while uniform substitution and deletion combine pattern probabilities with the number of outputs or retained subsets represented by each group.

### B.2 Confidence Bound from Sampled Predictions

For $S$ target-label decisions among $N$ independent samples, we use the one-sided Clopper–Pearson lower bound

$$\underline{p} = \begin{cases} 0, & S = 0, \\ F^{-1}_{\mathrm{Beta}(S, N-S+1)}(\alpha), & 1 \leq S \leq N, \end{cases} \tag{21}$$

where $F^{-1}_{\mathrm{Beta}(S,N-S+1)}$ is the quantile function of the indicated Beta distribution and $\alpha \in (0,1)$ is the confidence error probability. When $S = N$, this gives $\underline{p} = \alpha^{1/N} < 1$. The bound has coverage at least $1 - \alpha$ for each fixed input and kernel. Since the target event is specified before sampling, the same confidence bound applies to all candidate radii for that input.

### B.3 Bounding the Target-Label Probability after an Attack

The confidence bound $\underline{p}$ tells us how much probability the detector assigns to its target label under the original input, but not which corrupted outputs receive that label. A certificate must remain valid

for every assignment consistent with this evidence. The least favorable assignment places target predictions on outputs that are common under the original input but rare under the attacked input. Consequently, the attacked-to-original probability ratio determines the order in which these outputs enter the optimization.

The fractional-knapsack framework from prior work (Chen et al., 2025a) implements this calculation. Its usefulness here is that outputs with a common ratio can be collected into one group: only their total probability under each input is needed. The following proposition states the exact reduction, including outputs that are possible under only one of the two inputs.

**Proposition 4** (Lower bound from grouped outputs)**.** *Let $P$ and $P'$ be probability distributions on a common finite outcome space $\Omega$. Partition $\Omega$ into groups $G_g$ such that each group either has zero $P$ mass or satisfies $P'(z) = \lambda_g P(z)$ for every $z \in G_g$, with a constant $\lambda_g \geq 0$. Set $P_g = \sum_{z \in G_g} P(z)$ and $P'_g = \sum_{z \in G_g} P'(z)$. For any $\underline{p} \in [0, 1]$,*

$$\min_{\substack{h:\Omega \to [0,1] \\ \mathbb{E}_P[h] \geq \underline{p}}} \mathbb{E}_{P'}[h] = \mathcal{L}\big((P_g)_g, (P'_g)_g; \underline{p}\big). \tag{22}$$

*Consequently, every $h$ with $\mathbb{E}_P[h] \geq \underline{p}$ satisfies $\mathbb{E}_{P'}[h] \geq \mathcal{L}((P_g)_g, (P'_g)_g; \underline{p})$.*

The auxiliary function $h$ represents a possible assignment of target-label probabilities to the outputs. Its constraint preserves the observed lower bound, and its objective measures how much of that target probability remains after the input is modified. Fractional values are allowed in the optimization even when the actual detector is deterministic; this enlarges the feasible set and therefore remains valid for a lower bound.

*Proof.* Grouping is exact because outputs with the same probability ratio have the same effect on the objective per unit of probability contributed to the constraint. To formalize this, for a group with $P_g > 0$, define its average target-label value under $P$ by

$$u_g = \frac{1}{P_g} \sum_{z \in G_g} P(z) h(z). \tag{23}$$

The constant-ratio assumption gives $\sum_{z \in G_g} P'(z) h(z) = P'_g u_g$, including $P'_g = 0$. If $P_g = 0 < P'_g$, define $u_g$ using the $P'$-weighted average instead; if both masses vanish, set $u_g = 0$. These choices preserve both expectations:

$$\mathbb{E}_P[h] = \sum_g P_g u_g, \qquad \mathbb{E}_{P'}[h] = \sum_g P'_g u_g. \tag{24}$$

Every feasible $h$ therefore produces a feasible group vector with the same objective value. Conversely, a feasible vector $(u_g)_g$ defines $h(z) = u_g$ throughout group $G_g$, preserving the same two expectations. This proves Equation 22.

The minimum is attained because the feasible group set is nonempty and compact. We may choose an optimum with $\sum_g P_g u_g = \underline{p}$: decreasing any excess probability cannot increase the nonnegative objective. Groups with $P_g = 0$ can be assigned zero, since they do not help meet the constraint. For the remaining groups, moving an equal amount of $P$-weighted assignment from a larger ratio $P'_g/P_g$ to a smaller ratio leaves the constraint unchanged and decreases the objective. An optimum therefore fills groups in increasing ratio order, with at most one partially filled group after an arbitrary ordering of ties. This recovers the established fractional-knapsack solution (Chen et al., 2025a). □

The remaining derivations construct the probability groups to which this result applies. Masking has a particularly simple shared-output structure and admits a closed form. Uniform substitution requires groups indexed by mismatch counts. Deletion additionally requires an alignment of token occurrences, because different retention patterns can produce the same text.

### B.4 Masking: Probability Mass Shared by Both Inputs

Masking can eliminate an adversarial change completely: once a changed position is replaced by the mask symbol, its original token no longer affects the output. If $x$ and $x'$ differ at $d$ positions, the two

masking distributions therefore share exactly the outputs in which all $d$ positions are masked. The probability of this event is $c_d = \mathbb{E}[Q^d]$.

To turn this observation into a bound on predictions, draw one rate $Q$ and apply the same masking decisions to both inputs, independently across positions conditional on $Q$. Each resulting text, denoted by $Z$ or $Z'$, has the required marginal distribution. On the event that all modified positions are masked, $Z = Z'$ and the detector has the same value on both. Elsewhere its value can differ by at most one, because $0 \le f \le 1$. For $d \le r$, this gives

$$p(x') = \mathbb{E}[f(Z')] \ge \mathbb{E}[f(Z)] - \mathbb{P}(Z \ne Z') \ge p(x) - (1 - c_r). \tag{25}$$

The loss $1 - c_r$ is the largest probability that this argument allows to move away from the target label. Whenever $p(x) \ge \underline{p}$, nonnegativity consequently gives $p(x') \ge [\underline{p} + c_r - 1]_+$.

**Why the bound is exact for bounded detectors.** The mask symbol lies outside the input vocabulary, so a shared output must mask every differing position. Moreover, such an output specifies the same mask pattern under both inputs and hence has the same probability, even after averaging over $Q$. At exact distance $d$, each distribution thus has a common mass $c_d$ with probability ratio one and an exclusive mass $1 - c_d$.

The least favorable detector can place its target probability on original-only outputs first. If $\underline{p} \le 1 - c_d$ and $c_d < 1$, assigning value $\underline{p}/(1 - c_d)$ to these outputs and zero elsewhere gives attacked expectation zero. Otherwise, assign value one to original-only outputs, $(\underline{p} - 1 + c_d)/c_d$ to shared outputs, and zero to attacked-only outputs. The clean expectation is exactly $\underline{p}$, while the attacked expectation is $[\underline{p} + c_d - 1]_+$. The latter construction also covers $c_d = 1$, including $d = 0$. Thus no larger model-independent bound follows from the same probability evidence over bounded detectors. Finally, $c_d$ decreases with $d$, so the least favorable distance in $0 \le d \le r \le n$ is $d = r$, giving the masking certificate in Theorem 1.

### B.5 Uniform Substitution: Grouping by Mismatch Counts

Uniform substitution can generate the same output from two inputs with different probabilities. The certificate must retain these probability ratios, rather than only count outputs shared by the two distributions. Fortunately, the probability of a particular output depends on its mismatch count with the input, not on the identities of its replacement tokens. If the output differs at $t$ of $n$ positions, that probability is $m_n(t) = w_n(t)/(V-1)^t$: the factor $w_n(t)$ selects the required positions, and each specified replacement token contributes $1/(V-1)$.

For a pair $x, x' \in \mathcal{V}^n$ differing at $d$ positions, a group must consequently determine two mismatch counts. Within the differing positions, let $i$ count mismatches to $x$ and $j$ count mismatches to $x'$. An output token there either equals the original token, equals the attacked token, or equals neither. These three possibilities occur at $d - i$, $d - j$, and $h = i + j - d$ positions, respectively. In the remaining $n - d$ positions the two inputs agree; let $k$ count their common mismatches. For vocabulary size $V = |\mathcal{V}| \ge 3$, the admissible indices satisfy

$$0 \le i, j \le d, \qquad h = i + j - d \ge 0, \qquad 0 \le k \le n - d. \tag{26}$$

Each output now belongs to exactly one group $(i, j, k)$, with total mismatch counts $i + k$ and $j + k$ under the two inputs.

The group size follows by choosing where each of the three possibilities occurs and then choosing the actual replacement tokens:

$$v_{ijk} = \frac{d!}{(d-i)!(d-j)!h!}(V-2)^h \binom{n-d}{k}(V-1)^k. \tag{27}$$

The multinomial factor allocates the $d$ changed positions; the $h$ positions matching neither input each allow $V - 2$ tokens. The last two factors select and replace the $k$ positions where the inputs originally agree. Multiplying this count by the probability of one output gives

$$\begin{aligned} P_{ijk} &= v_{ijk}\frac{w_n(i+k)}{(V-1)^{i+k}} = v_{ijk}m_n(i+k), \\ P'_{ijk} &= v_{ijk}\frac{w_n(j+k)}{(V-1)^{j+k}} = v_{ijk}m_n(j+k). \end{aligned} \tag{28}$$

All outputs in a group have the same ratio, so Proposition 4 applies without loss. Collecting these masses into $P_d, P'_d$ gives the lower bound for exactly $d$ replacements $B_d(\underline{p}) = \mathcal{L}(P_d, P'_d; \underline{p})$. Because the attacker may use fewer than $r$ changes, the certificate is the minimum over all allowed distances:

$$L_r^{\text{uniform}}(\underline{p}) = \min_{0 \le d \le \min(r,n)} B_d(\underline{p}). \tag{29}$$

**Why unchanged positions remain in the calculation.** Under a fixed rate, probability factors from positions where $x$ and $x'$ agree cancel in the ratio. Under a shared random rate, observing $k$ replacements in these positions changes the relative weight assigned to low and high rates, which also changes the probabilities of the differing positions. The index $k$ must therefore be retained. For the same reason, the rate is integrated out in $w_n$ before applying the optimization: the confidence bound constrains the mixture's target probability, not a separate target probability at each rate. Averaging fixed-rate certificates while imposing the same $\underline{p}$ at every rate would impose evidence that the samples do not provide.

### B.6 Deletion: Aligning Retained Token Occurrences

Deletion differs from the preceding kernels in two respects: attacks can change the sequence length, and a retained text need not identify the positions from which it came. For example, retaining either occurrence of a repeated token can produce the same one-token text. Directly enumerating common output strings must account for these coincidences. Instead, an edit alignment identifies a set of retained outputs whose probabilities can be computed from the lengths and edit counts alone.

Fix an alignment containing $s$ substitutions, $i$ insertions, and $j$ deletions. If the original length is $n$, the attacked length is $m = n - j + i$, and $\ell = n - s - j$ token occurrences are matched without change. Give each matched pair the same occurrence label on both sides, and use distinct labels for unmatched occurrences. A labeled output is shared exactly when it retains only matched occurrences. For a specified subset of $k$ matched occurrences, the retention probabilities are $\rho_n(k) = w_n(n-k)$ and $\rho_m(k) = w_m(m-k)$: the selected occurrences must survive and every other occurrence must be deleted. There are $\binom{\ell}{k}$ such subsets, giving

$$P_k = \binom{\ell}{k} w_n(n-k), \qquad P'_k = \binom{\ell}{k} w_m(m-k), \tag{30}$$

for $0 \le k \le \ell$. These are the shared groups in Equation 7. Within each group the ratio is constant, although its value can differ from one because the two input lengths can differ.

The remaining labeled outputs contain an unmatched occurrence and hence belong to only one input's support. They form two additional groups:

$$(P_{\ell+1}, P'_{\ell+1}) = \left(1 - \sum_{k=0}^{\ell} P_k, 0\right), \qquad (P_{\ell+2}, P'_{\ell+2}) = \left(0, 1 - \sum_{k=0}^{\ell} P'_k\right). \tag{31}$$

The indices $\ell + 1$ and $\ell + 2$ denote original-only and attacked-only outputs, respectively. Together with $k = 0, \ldots, \ell$, they complete the two probability distributions. The original-only group has ratio zero; the attacked-only group cannot contribute to the original-probability constraint.

**Why occurrence labels give a valid certificate.** The detector observes retained text, without the occurrence labels introduced above. Every detector on text is therefore also a function on labeled outputs, obtained by first discarding the labels. Optimizing over all functions on labeled outputs allows some extra assignments: two patterns that render the same text may receive different values. This can only reduce the minimum attacked probability, so the resulting bound remains valid. The labels make the calculation tractable, while the possible coincidence of rendered strings explains why this deletion certificate need not be tight on text itself.

This argument applies to every alignment whose edit counts satisfy

$$m = n - j + i, \qquad \ell = n - s - j, \qquad s, i, j \ge 0, \quad s + j \le n, \quad s + i + j \le r. \tag{32}$$

Every input reachable within $r$ edits admits such an alignment. Thus minimizing $\mathcal{L}(P_{s,i,j}, P'_{s,i,j}; \underline{p})$ over these compositions covers the whole allowed attack set, where $P_{s,i,j}, P'_{s,i,j}$ contain the groups $k = 0, \ldots, \ell + 2$. Restricting the alignment to an editable suffix leaves the argument unchanged because the same protected prefix is attached to every retained output.

**A common lower bound.** The grouped certificate distinguishes the number of matched occurrences that survive. A simpler bound follows by considering the event that all edited columns are removed. Relating this event to the groups above allows the later sampling analysis to use the same joint-removal probability as masking.

**Theorem 5** (Joint-removal bound for masking and deletion)**.** *Let $c_r = \mathbb{E}[Q^r]$ for a fixed, length-independent rate distribution. For attack budgets within the editable length in the masking case, the certificates in Section 3.2 satisfy*

$$L_r^{\mathrm{mask}}(\underline{p}) = [\underline{p} + c_r - 1]_+, \qquad L_r^{\mathrm{del}}(\underline{p}) \geq [\underline{p} + c_r - 1]_+. \tag{33}$$

*The deletion inequality holds against the full edit-distance ball. Whenever $p(x) \geq \underline{p}$, these expressions lower-bound the target-label probability of every allowed attacked input.*

*Proof.* The masking equality follows from Appendix B.4. For deletion, an alignment lets us remove edits from both inputs with one decision per edited column. Choose an alignment with $s$ substitutions, $i$ insertions, and $j$ deletions, of total cost $t = s + i + j \leq r$. A column contains either two token occurrences or one occurrence opposite a gap. Draw the same rate $Q = q$ for both inputs and independently delete each column with probability $q$, removing whichever occurrences it contains. Each token belongs to a different column within its own input, so each input still has exactly the prescribed independent deletion decisions conditional on $q$.

If all $t$ edited columns are deleted, only matched columns can survive. Their retained occurrences have the same labels and produce the same text on both sides. Hence

$$\mathbb{P}(Z = Z' \mid Q = q) \geq q^t \geq q^r, \qquad \mathbb{P}(Z = Z') \geq \mathbb{E}[Q^r] = c_r. \tag{34}$$

The construction applies to substitutions, insertions, and deletions because a column is removed in the same way whether it contains one occurrence or two. It relies on the rate distribution being identical for the original and attacked inputs.

To establish a lower bound on the grouped certificate itself, consider the occurrence-labeled distributions defined above. Write $H = \sum_g \min(P_g, P'_g)$ for their overlap, including the two exclusive groups. Within each group the probability ratio is constant, so summing the smaller group mass is the same as summing the smaller point probabilities. The joint construction above gives $H \geq c_r$: the probability of identical outcomes in any joint construction cannot exceed the distributions' overlap.

For any feasible group assignments $u_g \in [0, 1]$, their target probability can decrease by at most the total mass that is larger under the original distribution. Since both distributions sum to one, that mass is $1 - H$. Consequently,

$$\sum_g P'_g u_g \geq \sum_g P_g u_g - \sum_g (P_g - P'_g)_+ \geq \underline{p} - (1 - H) \geq \underline{p} + c_r - 1. \tag{35}$$

This holds for every feasible assignment and every alignment of cost at most $r$. Taking the minima used by the grouped certificate, and using nonnegativity, proves the deletion inequality in Equation 33. □

Thus $c_r$ determines the masking certificate and lower-bounds the shared mass used by the deletion certificate. Appendix C uses this relation to derive sufficient sample counts for both kernels.

### B.7 Proof of Theorem 1

The kernel derivations compare one original input with one attacked input. The theorem requires a single certificate for all allowed attacked inputs, and then a valid choice of radius based on finite samples. The proof combines the preceding bounds using the same lower confidence bound; no additional probability estimates are needed for individual attacks or radii.

*Proof of Theorem 1.* Fix $x$ and the corruption kernel. Let $\mathcal{E}_x = \{p(x) \geq \underline{p}\}$ denote the event that the estimated lower bound is valid. Equation 21 and independent sampling ensure $\mathbb{P}(\mathcal{E}_x) \geq 1 - \alpha$. We establish the theorem on this event, beginning with a fixed allowed input $x'$.

For masking, if $x'$ differs from $x$ at $d \le r$ positions, Appendix B.4 gives

$$p(x') \ge [\underline{p} + c_d - 1]_+ \ge [\underline{p} + c_r - 1]_+, \qquad c_r = \frac{\mathrm{B}(a+r, b)}{\mathrm{B}(a, b)}. \tag{36}$$

The second inequality follows from $Q \in [0, 1]$, which implies $c_d \ge c_r$. The right-hand side depends only on the allowed budget, so it holds for every such $x'$.

For uniform substitution, the groups in Appendix B.5 partition $\mathcal{V}^n$. Every output in group $(i, j, k)$ has probabilities $m_n(i + k)$ and $m_n(j + k)$ under the original and attacked inputs, respectively. The ratios are constant within groups, and the group masses sum to one on each side. Proposition 4, applied to the actual detector $h = f$, therefore yields

$$p(x') \ge \mathcal{L}(P_d, P'_d; \underline{p}) \ge \min_{0 \le t \le \min(r,n)} \mathcal{L}(P_t, P'_t; \underline{p}). \tag{37}$$

Here $d$ is the distance of this particular attack and $t$ ranges over all admissible integer distances. The probabilities depend on the pair only through $n$ and $d$, so the minimum covers every allowed replacement attack.

For deletion, choose an alignment from $x$ to $x'$ with composition $(s, i, j)$ of total cost at most $r$. Write $\widetilde{P}, \widetilde{P}'$ for its occurrence-labeled deletion distributions. Let $R$ discard the labels and render the retained text, and set $\widetilde{f} = f \circ R$. Since rendering recovers the specified corruption mechanism on each input,

$$\mathbb{E}_{\widetilde{P}}[\widetilde{f}] = p(x), \qquad \mathbb{E}_{\widetilde{P}'}[\widetilde{f}] = p(x'). \tag{38}$$

Thus the real detector is feasible in the optimization on labeled outcomes. The shared groups have constant ratio $\rho_m(k)/\rho_n(k)$; the two exclusive groups have the residual masses given above. These nonnegative groups exhaust both distributions, so Proposition 4 gives

$$p(x') \ge \mathcal{L}(P_{s,i,j}, P'_{s,i,j}; \underline{p}) \ge L_r^{\mathrm{del}}(\underline{p}). \tag{39}$$

The last inequality includes the chosen alignment among all feasible compositions. Distinct labeled outcomes may render the same text, but this does not affect feasibility of $\widetilde{f}$ and cannot invalidate the lower bound.

*From an attacked-probability bound to a certified radius.* All three inequalities hold on the same event $\mathcal{E}_x$. If $L_r(\underline{p}) \ge \tau$, every allowed input at that radius has $p(x') \ge \tau$ and retains the target label. The largest passing radius can therefore be selected from the same probability estimate without a union bound over attacks or tested radii. At radius zero all three bounds equal $\underline{p}$, and each bound over an attack ball is nonincreasing in its budget. Hence the procedure abstains if $\underline{p} < \tau$.

It remains to verify that a nonempty set of passing radii has a largest element. For replacement attacks, $0 \le r \le n$ makes this immediate. For deletion, the attack can change the length, but sufficiently large budgets still fail at every positive threshold. Indeed, for $r \ge n$ the feasible composition $(s, i, j) = (0, r - n, n)$ deletes all original occurrences and inserts $r - n$ new ones. There are no matched occurrences, so its only shared labeled outcome is the empty sequence, with attacked probability $c_{r-n}$. Assigning value $\underline{p}$ to every original-supported group and zero to the attacked-only group is feasible, giving

$$0 \le L_r^{\mathrm{del}}(\underline{p}) \le \underline{p}\, c_{r-n} \longrightarrow 0 \qquad (r \to \infty). \tag{40}$$

The limit follows by dominated convergence because a Beta-distributed rate is less than one almost surely. Thus the maximum in Equation 10 exists whenever certification is possible. □

## C Certified Radii and Sampling Requirements

The certificates above separate the evidence supplied by the detector from the probability loss caused by an attack. This separation allows two comparisons. At the same sample count $N$ and target-label count $S$, the confidence bound is identical, so any difference in certified radius comes from the corruption mechanism. Conversely, for a prescribed radius, the certificate determines how much statistical evidence is needed and hence how many samples can suffice. The first two proofs establish the radius comparison in Section 3.3; the subsequent analysis explains the sampling dependence for masking, deletion, and uniform substitution.

The distinction between these comparisons matters when interpreting a sampling requirement. The radius formulas allow any observed count $S$. To derive the required sample count we use the most favorable observation $S = N$, for which $\underline{p} = \alpha^{1/N}$. It gives the largest radius available at that sample budget from the stated confidence bound. We write $\delta_N = \delta_{N,N} = 1 - \alpha^{1/N} + \tau$ in this case. The corruption-rate distribution is fixed independently of the original or attacked input throughout.

### C.1 Proof of Theorem 2

With fixed-rate masking, every changed position must be masked independently at probability $\mu$. The resulting shared mass is $\mu^r$. The radius is therefore the largest number of changes for which this mass still compensates for the probability loss permitted by the certificate.

*Proof of Theorem 2.* Fix integers $N \geq 1$ and $0 \leq S \leq N$, confidence error probability $\alpha \in (0, 1)$, masking rate $\mu \in (0, 1)$, and decision threshold $\tau \in (0, 1)$. Let $\underline{p}(S, N)$ be the confidence bound in Equation 21, and assume $\underline{p}(S, N) \geq \tau$. The shared mass required for certification is

$$\delta_{S,N} = 1 - \underline{p}(S, N) + \tau \in (0, 1]. \tag{41}$$

Indeed, a shared mass of this size limits the loss to $\underline{p}(S, N) - \tau$, leaving at least the decision threshold after the attack.

Taking $Q = \mu$ in Appendix B.4 gives the tight masking bound $[\underline{p}(S, N) + \mu^r - 1]_+$, consistent with the existing certificate (Chen et al., 2025a). Since $\tau > 0$, the bound passes exactly when $\mu^r \geq \delta_{S,N}$. Solving for $r$ gives

$$\mu^r \geq \delta_{S,N} \quad \Longleftrightarrow \quad r \log \mu \geq \log \delta_{S,N} \quad \Longleftrightarrow \quad r \leq \frac{\log \delta_{S,N}}{\log \mu}. \tag{42}$$

The inequality reverses on division because $\log \mu < 0$. Its right-hand side is nonnegative, so the largest permitted integer radius, with $n$ editable tokens, is

$$r_{\text{fixed}}(S, N) = \min\left\{n, \left\lfloor \frac{\log \delta_{S,N}}{\log \mu} \right\rfloor\right\}. \tag{43}$$

Equality at the threshold is accepted, including when the ratio is an integer. At $\underline{p}(S, N) = \tau$, the required shared mass is one and the radius is zero; $n = 0$ likewise returns zero. If $\underline{p}(S, N) < \tau$, even the unchanged input fails the certification test and the procedure abstains. □

### C.2 Proof of Theorem 3

Rate randomization changes the shared mass while leaving the required mass $\delta_{S,N}$ unchanged. For the default Beta family with second parameter one, that shared mass has a simple closed form. Substituting it into the same certification condition yields the radius comparison directly.

*Proof of Theorem 3.* Fix integers $N \geq 1$ and $0 \leq S \leq N$, with $\alpha, \tau, \mu \in (0, 1)$ and editable length $n \geq 0$. Assume the confidence bound satisfies $\underline{p}(S, N) \geq \tau$, and use $\delta_{S,N} = 1 - \underline{p}(S, N) + \tau$. Set $a = \mu/(1 - \mu)$ and draw $Q \sim \text{Beta}(a, 1)$. Its density is $aq^{a-1}$ on $(0, 1)$ and its expectation is $a/(a + 1) = \mu$, matching the fixed-rate mechanism's expected corruption rate.

Conditional on $Q = q$, removing all $r$ modifications has probability $q^r$. Averaging over this density gives

$$c_r = \mathbb{E}[Q^r] = a \int_0^1 q^{a+r-1}\, dq = \frac{a}{a + r}. \tag{44}$$

The integral is finite for every nonnegative integer $r$, since $a + r > 0$. Thus the shared mass decreases inversely with the number of modifications, whereas fixed-rate masking gives $\mu^r$. Applying the masking certificate in Theorem 1 now yields

$$\begin{aligned} [\underline{p}(S, N) + c_r - 1]_+ \geq \tau \quad &\Longleftrightarrow \quad \frac{a}{a + r} \geq \delta_{S,N} \\ &\Longleftrightarrow \quad a + r \leq \frac{a}{\delta_{S,N}} \\ &\Longleftrightarrow \quad r \leq a(\delta_{S,N}^{-1} - 1). \end{aligned} \tag{45}$$

Both denominators are positive and $\delta_{S,N} \leq 1$, so these rearrangements preserve equivalence and give a nonnegative upper bound. Taking the largest integer within the editable length proves

$$r_{\text{random}}(S, N) = \min\left\{n, \left\lfloor a(\delta_{S,N}^{-1} - 1)\right\rfloor\right\}. \tag{46}$$

As before, equality is included, $\delta_{S,N} = 1$ or $n = 0$ gives radius zero, and $\underline{p}(S, N) < \tau$ leads to abstention. □

The two proofs use the same probability evidence and the same expected corruption rate. Their difference is entirely in the shared mass: $\mu^r$ versus $a/(a + r)$. More generally, $\mathbb{E}[Q^r] \geq \mu^r$ for every nonnegative integer $r$, so any radius accepted by the fixed-rate masking certificate is also accepted after rate randomization at the same $(S, N)$. This is a comparison of the certificates; the observed counts determine where a particular detector lies on these radius curves.

### C.3 Fixed-Rate Masking: Sampling Requirement

For fixed-rate masking, increasing $N$ can improve the confidence bound, but even the most favorable observation $S = N$ gives a value below one. To see how this limits the radius, hold $r$ fixed and substitute $\underline{p} = \alpha^{1/N}$ into the masking condition. Certification then requires $\alpha^{1/N} \geq 1 - \mu^r + \tau$, or equivalently

$$N \geq \frac{\log \alpha}{\log(1 - \mu^r + \tau)}, \qquad 0 < \tau < \mu^r. \tag{47}$$

The smallest integer budget is the ceiling of this quantity. If $\mu^r \leq \tau$, no finite sample count can pass: even an all-success observation cannot make the confidence bound large enough to compensate for the lost probability.

Away from this threshold boundary, the expression reveals the exponential dependence. Suppose $\tau \leq \kappa\mu^r$ for a fixed $0 < \kappa < 1$, so the threshold uses at most a fixed fraction of the shared mass. Then $\mu^r - \tau$ has the same order as $\mu^r$, and $-\log(1 - u) = \Theta(u)$ as $u \downarrow 0$ gives a budget of order $\Theta(\mu^{-r} \log(1/\alpha))$. The statistical calculation itself does not cause the exponential dependence; it inherits that dependence from the shared mass. Theorem 5 relates the grouped deletion certificate to the same removal event, allowing the sampling calculation to extend to that kernel.

### C.4 Polynomial Sampling for Masking and Deletion

A Beta mixture changes how often a sample is strongly corrupted enough to remove many edits together. For large $r$, rates in $[1 - 1/r, 1]$ give $q^r \geq (1 - 1/r)^r$, which stays bounded away from zero. Near one, the Beta density is proportional to $(1 - q)^{b-1}$, so the probability assigned to this interval has order $r^{-b}$. Thus samples capable of jointly removing many edits become less frequent only polynomially. The following result makes this intuition precise and converts it into a sample count.

**Proposition 6** (Sampling requirement). *For $Q \sim \text{Beta}(a, b)$ with fixed $a, b > 0$,*

$$c_r = \frac{\Gamma(a + b)}{\Gamma(a)} r^{-b}(1 + o(1)). \tag{48}$$

*Given $S = N$, the masking certificate reaches $\tau$ when*

$$N \geq \frac{\log \alpha}{\log(1 - c_r + \tau)}, \qquad 0 < c_r - \tau < 1. \tag{49}$$

*The same condition is sufficient for the grouped deletion certificate. In the regime $0 < \tau \leq \kappa c_r$ for a fixed $0 < \kappa < 1$, this budget has order $\Theta(r^b \log(1/\alpha))$. This order is necessary and sufficient for the masking bound and sufficient for grouped deletion; fixed-rate joint erasure instead requires $\Theta(\mu^{-r} \log(1/\alpha))$ samples in its corresponding regime.*

Here $\Gamma$ is the gamma function. The constant $\kappa$ requires the threshold to remain a fixed fraction below the available shared mass, so the comparison concerns the growth of the sample budget before certification is precluded by the threshold itself.

*Proof. How the shared mass decays.* For general Beta parameters, the recurrence $\mathrm{B}(u+1,v) = u\mathrm{B}(u,v)/(u+v)$ gives

$$c_r = \frac{\mathrm{B}(a+r,b)}{\mathrm{B}(a,b)} = \prod_{j=0}^{r-1} \frac{a+j}{a+b+j}. \tag{50}$$

This product exposes the difference from fixed-rate masking. Instead of multiplying by the same factor $\mu$ for every extra modification, its factors approach one. Taking logarithms and using $\log(1+u) = u + O(u^2)$ in the tail of the sum yields

$$\begin{aligned} \log c_r &= -\sum_{j=0}^{r-1} \log\left(1 + \frac{b}{a+j}\right) \\ &= -b\sum_{j=0}^{r-1} \frac{1}{a+j} + O(1) = -b\log r + O(1). \end{aligned} \tag{51}$$

The finitely many initial terms contribute a constant, and the remaining expansion errors have bounded sum because $\sum_{j\geq 0}(a+j)^{-2}$ converges. The first-order terms form a harmonic sum, so exponentiating gives $c_r = \Theta(r^{-b})$.

The exact leading constant follows from the integral defining the same moment:

$$c_r = \frac{1}{\mathrm{B}(a,b)} \int_0^1 q^{a+r-1}(1-q)^{b-1}\, dq \tag{52}$$

$$= \frac{\Gamma(a+b)\Gamma(a+r)}{\Gamma(a)\Gamma(a+b+r)}. \tag{53}$$

For fixed $a, b > 0$, the ratio $\Gamma(r+a)/\Gamma(r+a+b) = r^{-b}(1+o(1))$ establishes Equation 48. When $b = 1$, the expression reduces to $a/(a+r)$, agreeing with Theorem 3.

*From shared mass to a sample budget.* Under $S = N$, the lower confidence bound solves $\underline{p}^N = \alpha$. The masking certificate therefore passes exactly when $\alpha^{1/N} + c_r - 1 \geq \tau$. For $0 < c_r - \tau < 1$, taking logarithms gives Equation 49; its ceiling is the smallest integer budget. As $c_r - \tau$ decreases, $-\log(1-u) = \Theta(u)$ gives the order

$$\Theta\left(\frac{\log(1/\alpha)}{c_r - \tau}\right). \tag{54}$$

If $\tau \leq \kappa c_r$, the denominator lies between $(1-\kappa)c_r$ and $c_r$. Substituting Equation 48 proves the polynomial requirement; substituting $c_r = \mu^r$ proves the exponential fixed-rate requirement in its corresponding regime. Theorem 5 transfers the polynomial sufficient budget to grouped deletion. At any fixed positive threshold, the stated regime eventually ends as the common mass decreases; if $c_r \leq \tau$, no finite sample count can satisfy this common-mass criterion. □

The radius interpretation follows by inverting the same relation. Write $C_{a,b} = \Gamma(a+b)/\Gamma(a)$, so $c_r \sim C_{a,b} r^{-b}$. Before the length cap and rounding, $c_r \geq \delta_N$ gives radius of order $(N/\log(1/\alpha))^{1/b}$ when $\log(1/\alpha)/N$ is small and dominates $\tau$, so that $\delta_N \approx \log(1/\alpha)/N$. For $b = 1$, Theorem 3 gives the exact expression without this approximation.

### C.5 Fixed-Rate and Randomized-Rate Deletion

Deletion permits insertions and removals as well as substitutions, so its edit radius can exceed the original length. A comparison must therefore account for the entire edit ball. At a fixed deletion rate, the shared groups have a common probability ratio, which makes their fractional-knapsack optimum explicit. Randomizing the rate changes these ratios across retained lengths; the joint-removal bound then provides a sufficient sampling budget for the full grouped certificate.

**Theorem 7** (Fixed-rate deletion certificate and sampling requirement)**.** *Let $n \geq 0$ be the editable input length, $r \geq 0$ an integer edit budget, and $0 < \mu < 1$ the fixed deletion probability. Write $h_r = \min(r, n)$. The occurrence-labeled grouped certificate over the full edit ball is*

$$L^{\mathrm{del}}_{r,\mathrm{fixed}}(\underline{p}) = \mu^{r-h_r}[\underline{p} + \mu^{h_r} - 1]_+. \tag{55}$$

*For $S = N$, confidence error probability $\alpha \in (0,1)$, and threshold $\tau \in (0,1)$, its smallest sufficient integer sample count is*

$$N^{\mathrm{del}}_{\mathrm{fixed}}(r) = \left\lceil \frac{\log \alpha}{\log\left(1 - \mu^{h_r} + \tau\mu^{-(r-h_r)}\right)} \right\rceil, \qquad 0 < \tau < \mu^r. \tag{56}$$

*No finite sample count suffices when $\tau \geq \mu^r$.*

*Proof.* For an alignment $(s,i,j)$, let $t = s + i + j \leq r$ and $k = s + j \leq n$. Its shared labeled outcomes have total original mass $\mu^{s+j} = \mu^k$ and total attacked mass $\mu^{s+i}$. Every shared group has ratio $\mu^{i-j}$, because the factors for retained occurrences cancel. The optimum first assigns target probability to original-only outcomes, then to shared outcomes, giving

$$\mathcal{L}(P_{s,i,j}, P'_{s,i,j}; \underline{p}) = \mu^{i-j}[\underline{p} + \mu^k - 1]_+. \tag{57}$$

Since $i - j = t - k - j \leq r - k$ and $\mu < 1$, this expression is at least

$$\mu^{r-k}[\underline{p} + \mu^k - 1]_+ = [\mu^r - (1 - \underline{p})\mu^{r-k}]_+. \tag{58}$$

The last quantity is nonincreasing in $k$. As $k \leq h_r$, every feasible alignment gives at least the right-hand side of Equation 55. The alignment $(s,i,j) = (h_r, r - h_r, 0)$ attains it, proving equality for the grouped certificate. This equality does not assert tightness after occurrence labels are discarded.

With $\underline{p} = \alpha^{1/N}$ and $\tau > 0$, certification is equivalent to

$$\alpha^{1/N} \geq 1 - \mu^{h_r} + \tau\mu^{-(r-h_r)}. \tag{59}$$

The right-hand side lies strictly between zero and one exactly when $0 < \tau < \mu^r$. Taking logarithms and rounding up proves Equation 56. When $\tau \geq \mu^r$, the required probability is at least one, which a finite-sample confidence bound cannot reach. □

**Theorem 8** (Sampling comparison for deletion at matched expected corruption rate)**.** *Fix $a, b > 0$, $\mu = a/(a+b)$, and $0 < \kappa < 1$. For every integer $r \geq 1$, use the same threshold $0 < \tau \leq \kappa\mu^r$ and all-success observation $S = N$ for fixed-rate and randomized-rate deletion. With $c_r = \mathbb{E}[Q^r]$ for $Q \sim \mathrm{Beta}(a,b)$, the randomized grouped certificate has sufficient budget*

$$N^{\mathrm{del}}_{\mathrm{random}}(r) \leq \left\lceil \frac{\log \alpha}{\log(1 - c_r + \tau)} \right\rceil. \tag{60}$$

*Equality holds in Equation 60 when $r \leq n$. For $n \geq r$ and fixed $\alpha \in (0,1)$, these budgets satisfy*

$$\begin{aligned} N^{\mathrm{del}}_{\mathrm{fixed}}(r) &= \Theta\left(\mu^{-r}\log(1/\alpha)\right), \\ N^{\mathrm{del}}_{\mathrm{random}}(r) &= \Theta\left(r^b \log(1/\alpha)\right). \end{aligned} \tag{61}$$

*Here $N^{\mathrm{del}}_{\mathrm{random}}(r)$ denotes the smallest budget at which its grouped certificate passes under $S = N$; for $r > n$, the displayed upper bound need not be attained.*

*Proof.* Theorem 5 guarantees $L^{\mathrm{del}}_{r,\mathrm{random}}(\underline{p}) \geq [\underline{p} + c_r - 1]_+$ for every edit radius, including $r > n$. When $r \leq n$, the alignment $(s,i,j) = (r,0,0)$ has equal original and attacked shared masses $c_r$, and all shared-group ratios equal one. Its optimum is $[\underline{p} + c_r - 1]_+$, so the common-mass bound is attained by the grouped certificate in this range. Substituting $\underline{p} = \alpha^{1/N}$ gives Equation 60, with equality for $r \leq n$. Jensen's inequality gives $c_r \geq \mu^r$, so the shared threshold satisfies $\tau \leq \kappa c_r$ as well. Consequently, $c_r - \tau$ lies between $(1-\kappa)c_r$ and $c_r$, and Proposition 6 gives the polynomial order, which is exact for this grouped certificate when $r \leq n$. When $n \geq r$, Equation 56 reduces to the fixed-rate masking expression, with $\mu^r - \tau$ between $(1-\kappa)\mu^r$ and $\mu^r$. Expanding $-\log(1-u)$ at zero gives its exponential order. □

Thus the exponential-to-polynomial comparison for deletion holds at the same expected rate, threshold, and favorable observation. The order comparison uses $n \geq r$; for edits beyond the original length, Equation 56 remains exact and Equation 60 remains sufficient. At a fixed positive threshold, the fixed-rate certificate fails for every sample count once $\mu^r \leq \tau$, whereas the randomized certificate can still pass if $c_r > \tau$ and the sampling budget is sufficient.

### C.6 Fixed-Rate and Randomized-Rate Uniform Substitution

For uniform substitution, complete corruption does not erase all dependence on the input: the original token is excluded from its replacement choices. The rate that removes this dependence is instead $q_\star = (V-1)/V$, where retaining the original token and producing each alternative both have probability $1/V$. A continuous rate distribution places no mass on this exact value, but rates near it still produce similar outputs from the two inputs. The probability assigned to this neighborhood determines a sufficient sampling budget.

To quantify this similarity, define

$$A_V(q) = 2\sqrt{\frac{q(1-q)}{V-1}} + \frac{(V-2)q}{V-1}, \qquad \omega_r = \frac{1}{2}\mathbb{E}[A_V(Q)^{2r}]. \tag{62}$$

Here $A_V(q)$ measures similarity at one changed position, and $\omega_r$ will lower-bound the probability mass shared by the two output distributions for up to $r$ replacements. The following theorem converts this bound into the number of samples sufficient for certification.

**Theorem 9** (Sampling requirement for randomized-rate uniform substitution)**.** *Fix $V \geq 3$, $Q \sim \mathrm{Beta}(a,b)$ with $a,b>0$, confidence error probability $\alpha \in (0,1)$, and integers $n \geq r \geq 1$. Under $S = N$ and $0 < \tau < \omega_r$, the grouped uniform-substitution certificate reaches $\tau$ whenever*

$$N \geq \left\lceil \frac{\log \alpha}{\log(1-\omega_r+\tau)} \right\rceil. \tag{63}$$

*There exist constants $C > 0$ and $r_0 \geq 1$, depending only on $a, b, V$, such that $\omega_r \geq C/\sqrt{r}$ for $r \geq r_0$. Consequently, for $0 < \tau \leq \kappa C/\sqrt{r}$ with fixed $0 < \kappa < 1$, a sufficient integer budget is*

$$N \geq \left\lceil \frac{\sqrt{r}\log(1/\alpha)}{(1-\kappa)C} \right\rceil = O\big(\sqrt{r}\log(1/\alpha)\big). \tag{64}$$

*These budgets are sufficient for the full grouped certificate; they need not be its smallest passing sample counts.*

*Proof.* At one changed position and fixed rate $q$, the original and attacked tokens receive probabilities $1-q$ and $q/(V-1)$, in opposite order under the two inputs. Each of the other $V-2$ tokens receives $q/(V-1)$ under both. Summing the square roots of paired probabilities gives $A_V(q)$ in Equation 62. This quantity, called the affinity, multiplies across independent positions. If the inputs differ at $d$ positions, their conditional output distributions have affinity $A_V(q)^d$; positions where they agree contribute a factor of one.

To relate affinity to the probability loss allowed by certification, take any discrete distributions $P, P'$ and write $A(P,P') = \sum_z \sqrt{P(z)P'(z)}$. Factoring their probability difference and applying Cauchy–Schwarz gives

$$\mathrm{TV}(P,P') = \frac{1}{2}\sum_z \left|\sqrt{P(z)} - \sqrt{P'(z)}\right| \left(\sqrt{P(z)} + \sqrt{P'(z)}\right) \tag{65}$$

$$\leq \sqrt{1 - A(P,P')^2}, \tag{66}$$

where $\mathrm{TV}(P,P') = \frac{1}{2}\sum_z |P(z) - P'(z)|$ is total variation distance. The overlap equals $1 - \mathrm{TV}(P,P')$, so $1 - \sqrt{1-u} \geq u/2$ for $0 \leq u \leq 1$ implies an overlap of at least $A(P,P')^2/2$.

Let $K_{x,q}$ be the output distribution for input $x$ conditional on rate $q$. Since a bounded detector's expectation can decrease by at most total variation, the preceding estimate gives

$$\mathbb{E}_{K_{x',q}}[f] \geq \mathbb{E}_{K_{x,q}}[f] - 1 + \frac{1}{2}A_V(q)^{2d}. \tag{67}$$

This inequality holds at each rate without a separate confidence estimate there. We can therefore average it over the common rate law before inserting the observed lower bound. Using $d \leq r$ and $0 \leq A_V(q) \leq 1$ yields

$$p(x') \geq p(x) - 1 + \frac{1}{2}\mathbb{E}[A_V(Q)^{2d}] \geq p(x) - 1 + \omega_r. \tag{68}$$

The quantity $\omega_r$ is a computable lower bound on the overlap retained after mixing. Equation 68 holds for every bounded detector and every distance at most $r$. Proposition 4 shows that the uniform groups represent this same class of assignments exactly; minimizing over them and over allowed distances consequently gives

$$L_r^{\text{uniform}}(\underline{p}) \geq [\underline{p} - 1 + \omega_r]_+. \tag{69}$$

The argument is independent of unchanged positions and is valid for every input length $n \geq r$, including any fixed post-processing through the function $f$.

To determine how $\omega_r$ depends on the radius, write the single-position affinity as

$$A_V(q) = 1 - \left(\sqrt{1-q} - \sqrt{q/(V-1)}\right)^2. \tag{70}$$

It reaches one exactly at $q_\star = (V-1)/V$, where the two conditional distributions coincide. Near that point its decrease is quadratic:

$$A_V(q) = 1 - \frac{V^3}{4(V-1)^2}(q - q_\star)^2 + O\big((q-q_\star)^3\big). \tag{71}$$

Thus rates within order $r^{-1/2}$ of $q_\star$ have $1 - A_V(q)$ of order $1/r$, allowing $A_V(q)^{2r}$ to remain bounded away from zero. The Beta density $\varphi_{a,b}(q) = q^{a-1}(1-q)^{b-1}/\mathrm{B}(a,b)$ is continuous and positive at this interior point, so an interval of that width carries probability of order $r^{-1/2}$.

For the exact asymptotic, substitute $t = \sqrt{r}(q - q_\star)$ in the integral defining $\omega_r$. Equation 71 then gives

$$\omega_r = \frac{1}{2}\int_0^1 A_V(q)^{2r}\varphi_{a,b}(q)\,dq \tag{72}$$

$$= \varphi_{a,b}(q_\star)\frac{V-1}{V^{3/2}}\sqrt{\frac{\pi}{2r}}\,(1 + o(1)). \tag{73}$$

To justify the limit, choose a fixed small neighborhood of $q_\star$. Within it the quadratic expansion bounds the rescaled integrand by an integrable Gaussian, and the density is bounded. Outside it, $A_V(q)$ is bounded strictly below one, so the contribution is exponentially small even if the Beta density is unbounded at an endpoint. Dominated convergence on the rescaled neighborhood proves the displayed expression. Its leading coefficient is positive, yielding constants $C > 0$ and $r_0$ with $\omega_r \geq C/\sqrt{r}$ for $r \geq r_0$. Together with Equation 69, this gives

$$L_r^{\text{uniform}}(\underline{p}) \geq \left[\underline{p} - 1 + \frac{C}{\sqrt{r}}\right]_+. \tag{74}$$

The constants depend on $a, b, V$; the asymptotic is not uniform in vocabulary size.

It remains to convert this probability bound into a sample count. Under $S = N$, the confidence bound is $\underline{p} = \alpha^{1/N}$. Since $\tau > 0$, a sufficient certification condition is

$$\begin{aligned} [\alpha^{1/N} - 1 + \omega_r]_+ \geq \tau \quad &\Longleftrightarrow \quad \alpha^{1/N} \geq 1 - \omega_r + \tau \\ &\Longleftrightarrow \quad N \geq \frac{\log \alpha}{\log(1 - \omega_r + \tau)}. \end{aligned} \tag{75}$$

Here $0 < \tau < \omega_r \leq 1/2$, so the logarithm in the denominator is negative. Rounding up gives Equation 63.

For $r \geq r_0$, the inequality $e^{-t} \geq 1 - t$ makes the polynomial dependence explicit:

$$L_r^{\text{uniform}}(\alpha^{1/N}) \geq \left[\frac{C}{\sqrt{r}} - \frac{\log(1/\alpha)}{N}\right]_+. \tag{76}$$

The budget in Equation 64 makes the sampling term at most $(1-\kappa)C/\sqrt{r}$, leaving at least $\kappa C/\sqrt{r} \geq \tau$. Thus a sample count growing as $\sqrt{r}$, up to the confidence factor, suffices for randomized-rate uniform substitution in the stated threshold regime. At a fixed positive threshold, this regime ends once the available overlap is too small. □

A fixed rate away from $q_\star$ behaves differently: each changed position favors one of the two source tokens. Across many positions, these preferences make the two output distributions distinguishable with exponentially small error. The corresponding certificate therefore needs a much tighter estimate of the original target-label probability.

**Theorem 10** (Exponential requirement for fixed-rate uniform substitution)**.** *Fix a vocabulary size $V \geq 3$, a rate $0 < \mu < 1$ with $\mu \neq q_\star = (V-1)/V$, and integers $n \geq r \geq 1$. Let $\Delta = |1 - \mu - \mu/(V-1)|$ and $\varepsilon_r = \exp(-r\Delta^2/2)$. If $\varepsilon_r < \tau$, certification at radius $r$ by the exact bounded-detector optimization requires $\underline{p} > 1 - \varepsilon_r$. In particular, under $S = N$, a necessary sampling condition is*

$$N > \frac{\log \alpha}{\log(1-\varepsilon_r)} = \Theta\Big(e^{r\Delta^2/2}\log(1/\alpha)\Big)\,. \tag{77}$$

*Proof.* Choose inputs differing at $r$ positions. At each changed position, assign score $+1$ to the token more probable under the original input, $-1$ to the token more probable under the attacked input, and zero to all other tokens. The score has expectation $\Delta$ under the original distribution and $-\Delta$ under the attacked distribution. Positions are independent at the fixed rate. Hoeffding's inequality for scores in $[-1, 1]$ therefore shows that the detector returning one when the total score is nonnegative satisfies

$$p(x) \geq 1 - \varepsilon_r, \qquad p(x') \leq \varepsilon_r. \tag{78}$$

If $\underline{p} \leq 1 - \varepsilon_r$, this detector is feasible in the exact optimization, but its attacked probability is below $\tau$. Hence the optimization cannot certify this pair. Substituting $\underline{p} = \alpha^{1/N}$ into the necessary inequality $\underline{p} > 1 - \varepsilon_r$ gives the sample condition. Finally, $-\log(1-\varepsilon_r) \sim \varepsilon_r$ establishes its exponential order. □

The two results can now be compared at the same expected corruption rate and threshold.

**Corollary 11** (Sampling comparison for uniform substitution at matched expected corruption rate)**.** *Fix $a, b > 0$ and $V \geq 3$ such that $\mu = a/(a+b) \neq q_\star$, and let $C, r_0$ be the constants in Theorem 9. Choose the same threshold $\tau_r = \kappa C/\sqrt{r}$, with $0 < \kappa < 1$, for both mechanisms. For all sufficiently large $r$ with $n \geq r$, under $S = N$ the smallest passing sample counts satisfy*

$$\begin{aligned} N^{\text{uniform}}_{\text{fixed}}(r) &= \Omega\Big(e^{r\Delta^2/2}\log(1/\alpha)\Big)\,, \\ N^{\text{uniform}}_{\text{random}}(r) &= O\big(\sqrt{r}\log(1/\alpha)\big)\,. \end{aligned} \tag{79}$$

*The fixed-rate statement is a necessary lower bound, and the randomized-rate statement is a sufficient upper bound.*

*Proof.* The exponential term $\varepsilon_r$ is eventually smaller than $\tau_r$, so Theorem 10 applies. The same $\tau_r$ satisfies the premise of Theorem 9, whose sufficient budget is $\sqrt{r}\log(1/\alpha)/((1-\kappa)C)$. Both results use the same expected corruption rate and threshold. □

The excluded rate has a specific interpretation: if $\mu = q_\star$, every token is replaced or retained with equal output probability $1/V$, so the corrupted input contains no information about its source. The fixed-rate certificate then equals $\underline{p}$ at every replacement radius up to $n$. This exceptional mechanism does not face the exponential obstruction, but it cannot distinguish inputs through their editable content. It does not include the fixed rates in our experiments.

For other fixed rates, the theorem establishes an exponential necessary sample count, whereas randomizing the rate supplies a polynomial sufficient count. This asymptotic comparison does not require randomized uniform substitution to attain the largest possible radius at every finite budget. The constant in Equation 73 depends on $V$ and the Beta parameters, and the bound concerns $S = N$. Actual certificates also depend on the observed count $S$ and cannot exceed the editable length. The experiments use the full grouped optimization rather than this sufficient overlap bound.

## D Implementation Details

The empirical certificates combine the kernel bounds derived above with a lower confidence bound on the detector's unsafe probability. We describe the numerical calculation of certified radii and the detector scoring used to estimate this probability.

### D.1 Certified Radius Computation

Certification at radius $r$ requires the probability bound to hold at every attack distance up to $r$. Taking the minimum of the exact-distance bounds over this interval gives a nonincreasing bound for the entire ball. If the exact-distance bounds are themselves nonincreasing, this minimum equals the bound at distance $r$.

For uniform substitution, Equation 29 defines the ball bound as the running minimum of $B_d(\underline{p})$ over distances $d$. All uniform-substitution experiments use this grouped optimization. The masking bound is nonincreasing directly, since its common mass decreases with the radius.

For deletion with a length-independent rate distribution, let $B_t$ denote the minimum lifted bound over edit compositions of exact cost $t$. Given a composition $(s, i, j)$ of cost $t$, the composition $(s, i+1, j)$ adds one inserted occurrence. Let $T$ remove this occurrence from a labeled output whenever it is retained. The map $T$ leaves the clean distribution unchanged and maps the extended attacked distribution to the original attacked distribution: conditional on $Q$, deletion decisions at the other occurrences have the same distribution. This identity also holds after integration over $Q$.

Any bounded function $f$ on the original outputs induces a function $f \circ T$ on the extended outputs with the same clean and attacked expectations. The minimum over all functions on the extended space is therefore no greater than the original minimum. Applying this argument to a minimizing composition at cost $t$ gives $B_{t+1} \leq B_t$. Grouping outputs with equal likelihood ratios preserves the optimum, so the same monotonicity holds for the grouped bound.

The largest certified radius for fixed-rate and length-independent random-rate deletion is computed by exponential bracketing followed by integer bisection. For AdaptDel, the deletion rate depends on sequence length, so the preceding marginalization argument does not apply. We therefore take the minimum over all smaller edit budgets when evaluating its ball bound.

The deletion search is limited to $10^6$ edits; all reported suffix-deletion radii lie below this limit. A substitution radius equal to the editable sequence length certifies the full fixed-length domain. For each exact deletion budget, there are at most $(n+1)(n+2)/2$ feasible $(s, j)$ pairs and at most $n+3$ probability groups per composition. Uniform substitution likewise uses grouped probabilities in place of explicit enumeration of $V^n$ strings. Pattern probabilities and group multiplicities are evaluated using log-beta and log-binomial functions, and likelihood ratios are computed in log space for numerical stability. The radius calculations use a single clean confidence bound and require no additional detector evaluations.

### D.2 Detector Evaluation

The conversational LLMs serve as deterministic safety classifiers using the safety-classification prompt from prior work (Chen et al., 2025a). The prompt specifies that classification depends on the overall intent and explicit harmful content of the final user request, with ambiguous or nonexplicit content assigned the safe label. Using language-model label probabilities for safety classification is also supported by prior work (Inan et al., 2023). We classify each request by comparing the conditional log-likelihoods of the complete labels `[Safe]` and `[Unsafe]`. For the formatted context $c(x)$ and a label tokenized as $y = (y_1, \ldots, y_m)$, where $m$ is the label length, the score is

$$s(y \mid c(x)) = \sum_{j=1}^{m} \log P_\theta(y_j \mid c(x), y_{<j}). \tag{80}$$

Here $P_\theta$ denotes the detector's conditional next-token distribution. A request is classified as unsafe when $s(\texttt{[Unsafe]} \mid c(x)) \geq s(\texttt{[Safe]} \mid c(x))$.

For full-input repetition, the entire repeated request is editable. Masking and uniform substitution use repetition counts $1, 2, 3, 5$, while deletion uses $1, 3, 5$; all counts include the unrepeated reference.