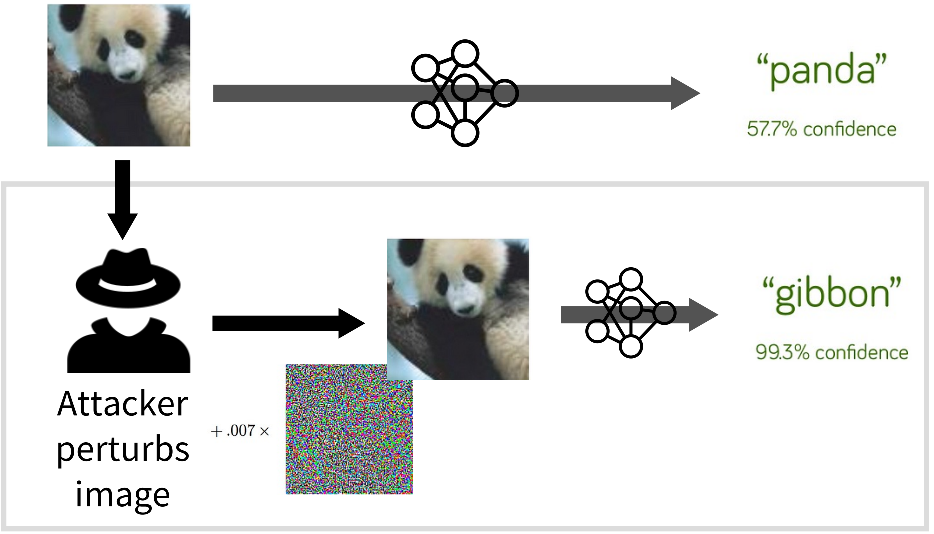
AdaptDel: Adaptable Deletion Rate Randomized Smoothing for Certified Robustness

Zhuoqun Huang, Neil G. Marchant
Olga Ohrimenko, Benjamin I.P. Rubinstein

The University of Melbourne

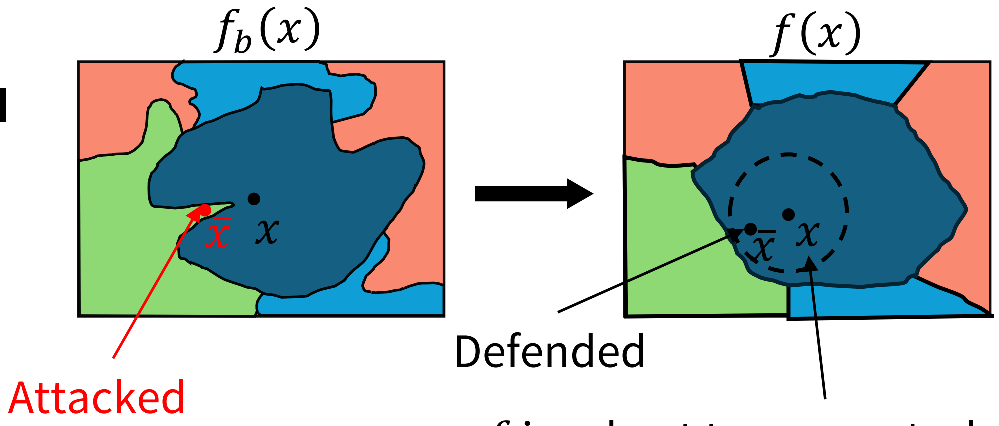
Attack



Certified Defense

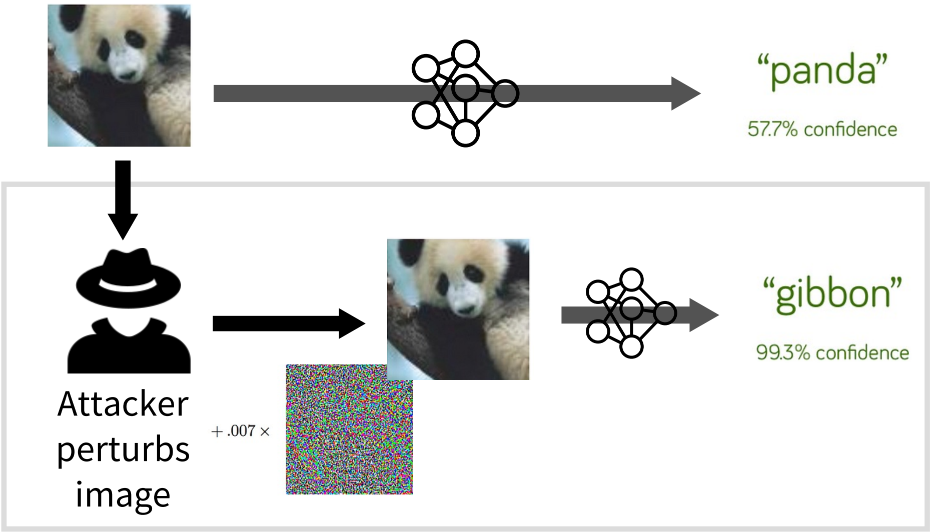
Gaussian Mech. Smoothing (Cohen et al. 2019)

$$f(x) := \arg\max_{y \in Y} \mathbb{E}_{z \sim \mathcal{N}(x, \sigma^2 \mathrm{I})} [\mathbf{1}_{f_b(z) = y}]$$



f is robust to any perturbation δ s.t. $\|\delta\|_2 \leq r = \sigma\Phi^{-1}(p_{\hat{y}}(x))$

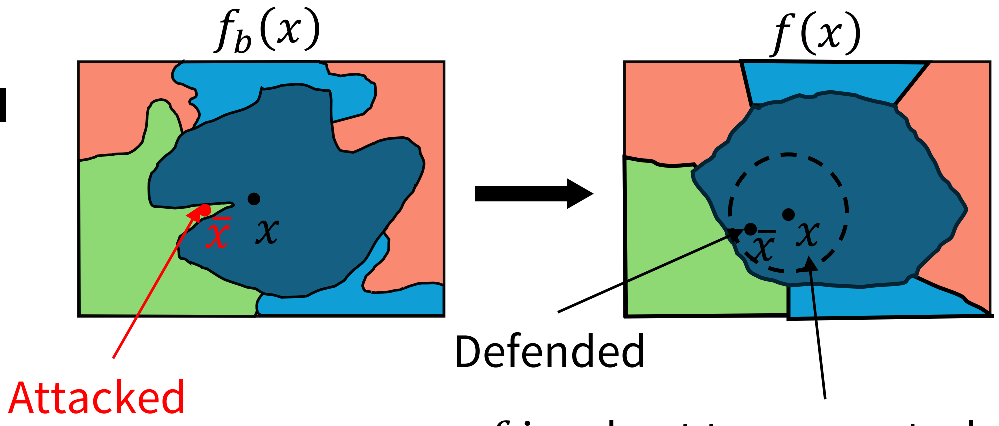
Attack



Certified Defense

Gaussian Mech. Smoothing (Cohen et al. 2019)

$$f(x) := \arg\max_{y \in Y} \mathbb{E}_{z \sim \mathcal{N}(x, \sigma^2 \mathrm{I})} [\mathbf{1}_{f_b(z) = y}]$$



f is robust to any perturbation δ s.t. $\|\delta\|_2 \leq r = \sigma\Phi^{-1}(p_{\hat{y}}(x))$

- Character-level attack
  - *South Africa's historic Soweto township marks its 100th birthday on Tuesday in a (mood) of optimism.* **[World]**
  - *South Africa's historic Soweto township marks its 100th birthday on Tuesday in a (mooP) of optimism.* **[Sci/Tech]**

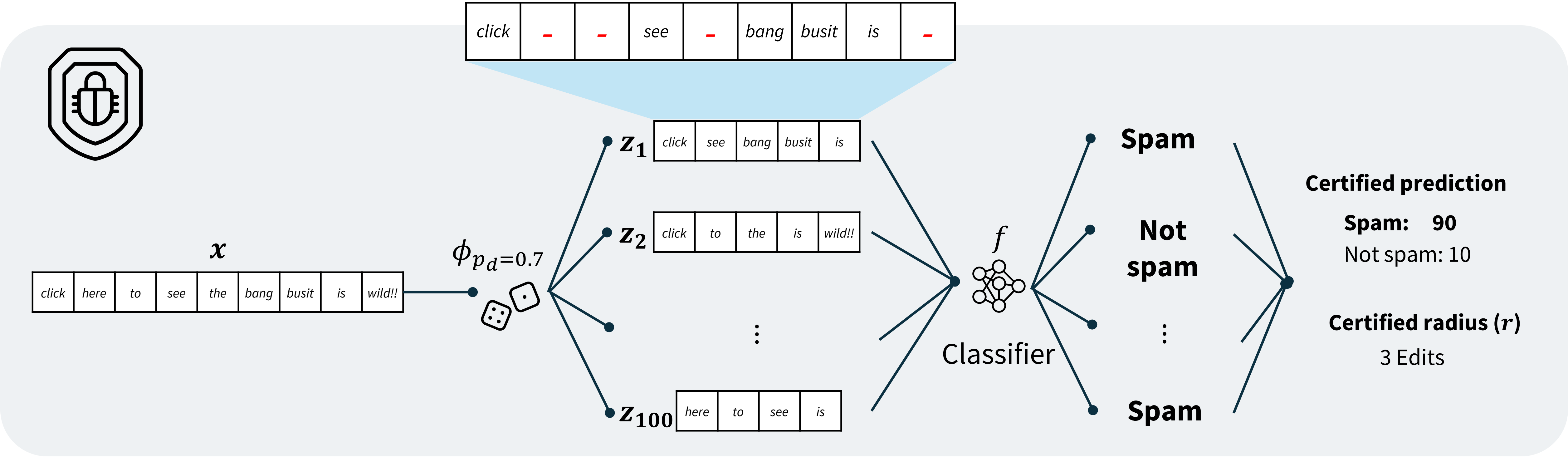
(Ebrahimi et al., ACL 2018)

- Word-level attack
  - *Super ant colony hits Australia. A (giant) 100km colony of ants could threaten (local) insect species.* **[World]**
  - *Super ant colony hits Australia (Coast). A (gigantic) 100km colony of ants could threaten ~~(local)~~ insect species.* **[Sci/Tech]**

(Li et al., NAACL 2021)

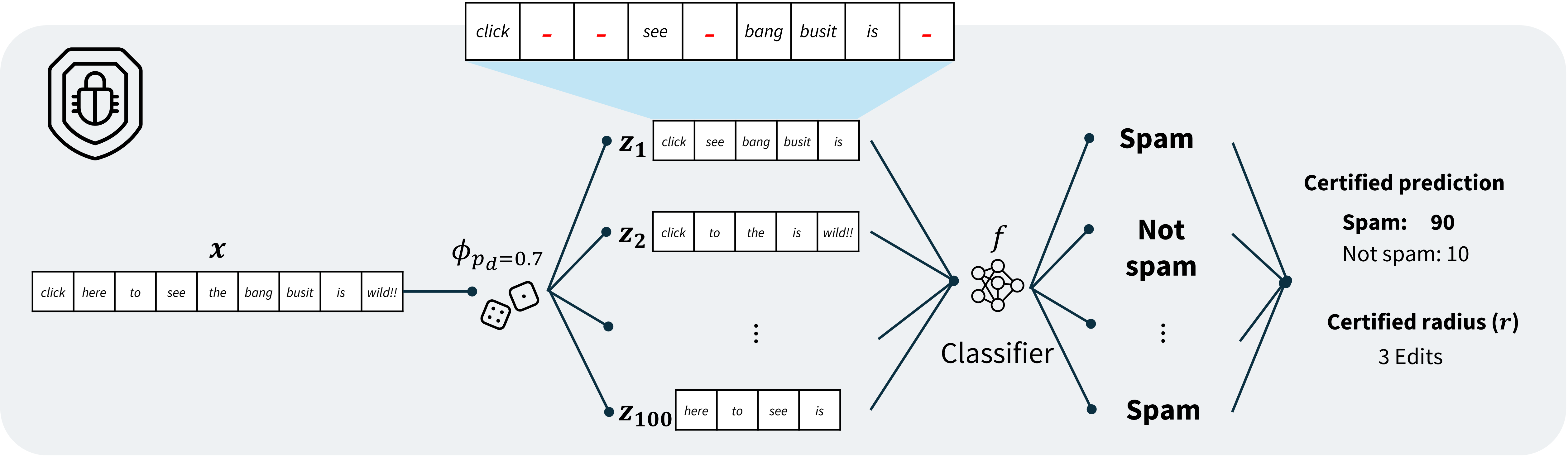
Background: CERT-ED

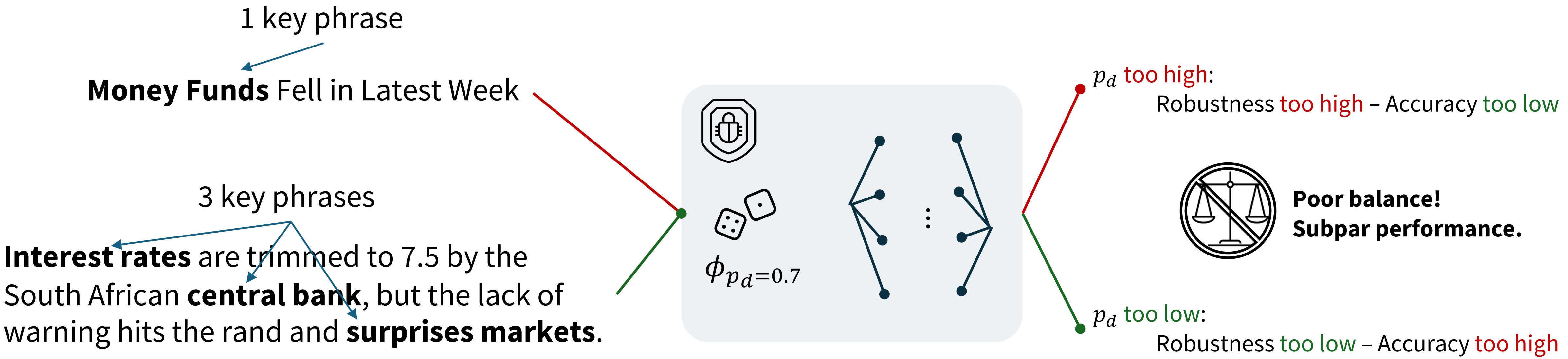
(Huang et al., EMNLP 2024)



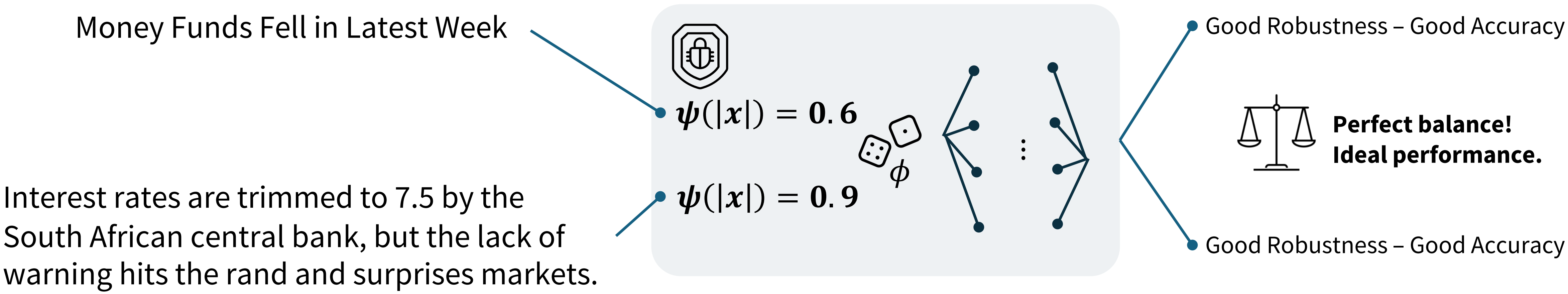
Background: CERT-ED

(Huang et al., EMNLP 2024)





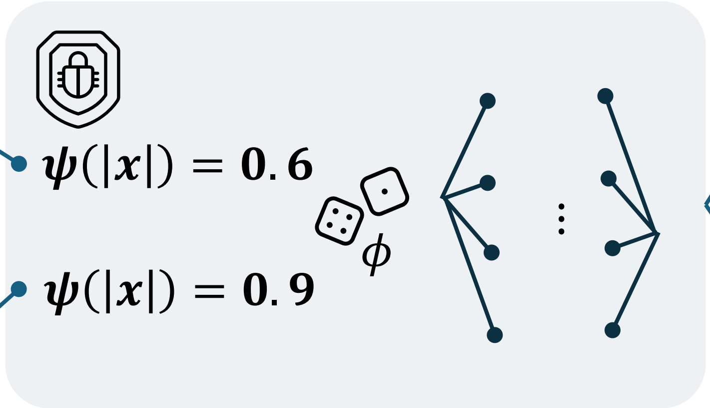
Model deletion rates as a function of the input sequence $p_{del} = \psi(|x|)$



Model deletion rates as a function of the input sequence $\boldsymbol{p_{del}} = \boldsymbol{\psi}(|\boldsymbol{x}|)$

Money Funds Fell in Latest Week

Interest rates are trimmed to 7.5 by the South African central bank, but the lack of warning hits the rand and surprises markets.



Good Robustness – Good Accuracy

Perfect balance!
Ideal performance.

Good Robustness – Good Accuracy

Algorithm 1 CERTIFY

Require: base classifier f_{b}, input sequence $\boldsymbol{x}$, predicted class y_1, length-dependent deletion probability ψ, allowed edit operations o, significance level α

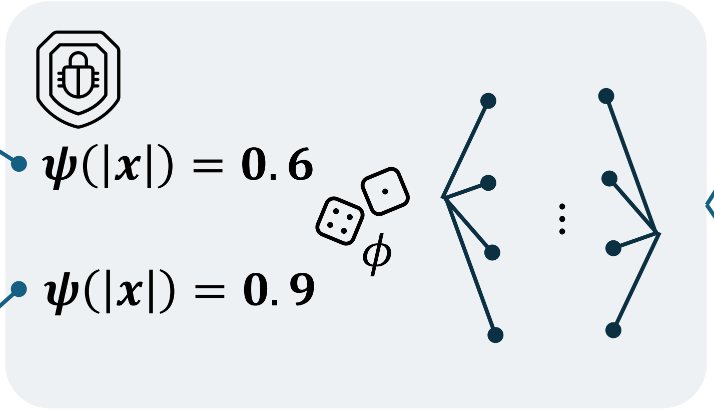
Ensure: maximum radius that can be certified

```
t_1^lb ← p̂_{y_1}^lb(x; f_b, φ_ψ, α)
t_2^ub ← max_{y≠y_1} p̂_y^ub(x; f_b, φ_ψ, α)
for r = 0 to ∞ do
    for all (n_del, n_ins, n_sub) ∈ C(o, r) do
        |x̄| ← |x| + n_ins − n_del
        t̄_1^lb ← lb(t_1^lb, x, x̄, ψ)
        t̄_2^ub ← ub(t_2^ub, x, x̄, ψ)
        if t̄_1^lb ≤ t̄_2^ub then
            return r
```

Model deletion rates as a function of the input sequence $\boldsymbol{p_{del}} = \boldsymbol{\psi}(|\boldsymbol{x}|)$

Money Funds Fell in Latest Week

Interest rates are trimmed to 7.5 by the South African central bank, but the lack of warning hits the rand and surprises markets.



- Good Robustness – Good Accuracy
- Good Robustness – Good Accuracy

Perfect balance!
Ideal performance.

Algorithm 1 CERTIFY

Require: base classifier f_{b}, input sequence $\boldsymbol{x}$, predicted class y_1, length-dependent deletion probability ψ, allowed edit operations o, significance level α

Ensure: maximum radius that can be certified

```
t_1^lb ← p̂_{y_1}^lb(x; f_b, φ_ψ, α)
t_2^ub ← max_{y≠y_1} p̂_y^ub(x; f_b, φ_ψ, α)
for r = 0 to ∞ do
    for all (n_del, n_ins, n_sub) ∈ C(o, r) do
        |x̄| ← |x| + n_ins − n_del
        t̄_1^lb ← lb(t_1^lb, x, x̄, ψ)
        t̄_2^ub ← ub(t_2^ub, x, x̄, ψ)
        if t̄_1^lb ≤ t̄_2^ub then
            return r
```

Lemma 3. *Let $\boldsymbol{x}, \bar{\boldsymbol{x}} \in \mathcal{X}$ be a pair of inputs with a longest common subsequence (LCS) $\boldsymbol{z}^\star$ and let $\mu = p_y(\boldsymbol{x}; f_{\mathrm{b}})$. Define*

$$H^* = \begin{cases} \min_{h: \sum_{i=0}^{h} \mathcal{B}_i(|\boldsymbol{z}^\star|, \psi) \geq \mu - 1 + \psi^{|\boldsymbol{x}| - |\boldsymbol{z}^\star|}} h, & \psi \geq \bar{\psi}, \\ \max_{h: \sum_{i=h}^{|\boldsymbol{z}^\star|} \mathcal{B}_i(|\boldsymbol{z}^\star|, \psi) \geq \mu - 1 + \psi^{|\boldsymbol{x}| - |\boldsymbol{z}^\star|}} h, & \psi < \bar{\psi}, \end{cases}$$

as a threshold on the number of tokens retained when editing $\boldsymbol{x}$, where $\mathcal{B}_k(n, p) := \binom{n}{k}(1-p)^k p^{n-k}$ is the Binomial pmf for n trials with success probability $1 - p$. Then there exists a lower bound $\mathrm{lb}(\mu, \boldsymbol{x}, \bar{\boldsymbol{x}}, \psi) \leq p_y(\bar{\boldsymbol{x}}; f_{\mathrm{b}})$ such that:

$$\mathrm{lb}(\mu, \boldsymbol{x}, \bar{\boldsymbol{x}}, \psi) = \frac{\bar{\psi}^{|\bar{\boldsymbol{x}}| - |\boldsymbol{z}^\star|}}{\psi^{|\boldsymbol{x}| - |\boldsymbol{z}^\star|}} \left(\sum_{i = \boldsymbol{l}(H^*+1)}^{(1-\boldsymbol{l})(H^*-1) + \boldsymbol{l}|\boldsymbol{z}^\star|} \mathcal{B}_i(|\boldsymbol{z}^\star|, \bar{\psi}) + \mathcal{B}_{H^*}(|\boldsymbol{z}^\star|, \bar{\psi}) \left\lfloor \frac{c(\mu, |\boldsymbol{x}|, |\boldsymbol{z}^\star|, \psi, H^*)}{\mathcal{B}_{H^*}(|\boldsymbol{z}^\star|, \psi)} \right\rfloor_{\binom{|\boldsymbol{z}^\star|}{H^*}^{-1}} \right),$$

where

$$c(\mu, |\boldsymbol{x}|, |\boldsymbol{z}^\star|, \psi, H^*) = \mu - 1 + \psi^{|\boldsymbol{x}| - |\boldsymbol{z}^\star|} - \sum_{i = \boldsymbol{l}(H^*+1)}^{(1-\boldsymbol{l})(H^*-1) + \boldsymbol{l}|\boldsymbol{z}^\star|} \mathcal{B}_j(|\boldsymbol{z}^\star|, \psi),$$

$\boldsymbol{l} = \mathbf{1}_{\psi < \bar{\psi}}$ is a binary indicator and $\lfloor \cdot \rfloor_v := \lfloor \frac{\cdot}{v} \rfloor v$ is a gridded flooring operation.

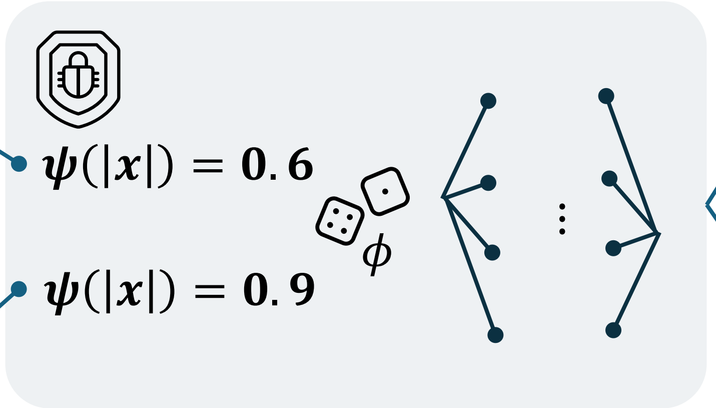
(Huang et al., NeurIPS 2023)

Sketch: Start with LCS technique, and progressively lower bound the quantity. Finally, formulate the minimization problem as knapsack problem.

Model deletion rates as a function of the input sequence $\boldsymbol{p_{del}} = \boldsymbol{\psi}(|\boldsymbol{x}|)$

Money Funds Fell in Latest Week

Interest rates are trimmed to 7.5 by the South African central bank, but the lack of warning hits the rand and surprises markets.



- Good Robustness – Good Accuracy
- **Perfect balance! Ideal performance.**
- Good Robustness – Good Accuracy

Algorithm 1 CERTIFY

Require: base classifier f_{b}, input sequence $\boldsymbol{x}$, predicted class y_1, length-dependent deletion probability ψ, allowed edit operations o, significance level α

Ensure: maximum radius that can be certified

```
t_1^lb ← p̂_{y_1}^lb(x; f_b, φ_ψ, α)
t_2^ub ← max_{y≠y_1} p̂_y^ub(x; f_b, φ_ψ, α)
for r = 0 to ∞ do
    for all (n_del, n_ins, n_sub) ∈ C(o, r) do
        |x̄| ← |x| + n_ins − n_del
        t̄_1^lb ← lb(t_1^lb, x, x̄, ψ)
        t̄_2^ub ← ub(t_2^ub, x, x̄, ψ)
        if t̄_1^lb ≤ t̄_2^ub then
            return r
```

Lemma 3. *Let $\boldsymbol{x}, \bar{\boldsymbol{x}} \in \mathcal{X}$ be a pair of inputs with a longest common subsequence (LCS) $\boldsymbol{z}^\star$ and let $\mu = p_y(\boldsymbol{x}; f_{\mathrm{b}})$. Define*

$$H^\star = \begin{cases} \min_{h: \sum_{i=0}^{h} \mathcal{B}_i(|\boldsymbol{z}^\star|, \psi) \geq \mu - 1 + \psi^{|\boldsymbol{x}| - |\boldsymbol{z}^\star|}} h, & \psi \geq \bar{\psi}, \\ \max_{h: \sum_{i=h}^{|\boldsymbol{z}^\star|} \mathcal{B}_i(|\boldsymbol{z}^\star|, \psi) \geq \mu - 1 + \psi^{|\boldsymbol{x}| - |\boldsymbol{z}^\star|}} h, & \psi < \bar{\psi}, \end{cases}$$

as a threshold on the number of tokens retained when editing $\boldsymbol{x}$, where $\mathcal{B}_k(n, p) \coloneqq \binom{n}{k}(1-p)^k p^{n-k}$ is the Binomial pmf for n trials with success probability $1 - p$. Then there exists a lower bound $\mathrm{lb}(\mu, \boldsymbol{x}, \bar{\boldsymbol{x}}, \psi) \leq p_y(\bar{\boldsymbol{x}}; f_{\mathrm{b}})$ such that:

$$\mathrm{lb}(\mu, \boldsymbol{x}, \bar{\boldsymbol{x}}, \psi) = \frac{\bar{\psi}^{|\bar{\boldsymbol{x}}| - |\boldsymbol{z}^\star|}}{\psi^{|\boldsymbol{x}| - |\boldsymbol{z}^\star|}} \left(\sum_{i=\boldsymbol{l}(H^\star+1)}^{(1-\boldsymbol{l})(H^\star-1)+\boldsymbol{l}|\boldsymbol{z}^\star|} \mathcal{B}_i(|\boldsymbol{z}^\star|, \bar{\psi}) + \mathcal{B}_{H^\star}(|\boldsymbol{z}^\star|, \bar{\psi}) \left\lfloor \frac{c(\mu, |\boldsymbol{x}|, |\boldsymbol{z}^\star|, \psi, H^\star)}{\mathcal{B}_{H^\star}(|\boldsymbol{z}^\star|, \psi)} \right\rfloor_{\binom{|\boldsymbol{z}^\star|}{H^\star}^{-1}} \right),$$

where

$$c(\mu, |\boldsymbol{x}|, |\boldsymbol{z}^\star|, \psi, H^\star) = \mu - 1 + \psi^{|\boldsymbol{x}| - |\boldsymbol{z}^\star|} - \sum_{i=\boldsymbol{l}(H^\star+1)}^{(1-\boldsymbol{l})(H^\star-1)+\boldsymbol{l}|\boldsymbol{z}^\star|} \mathcal{B}_j(|\boldsymbol{z}^\star|, \psi),$$

$\boldsymbol{l} = \mathbf{1}_{\psi < \bar{\psi}}$ is a binary indicator and $\lfloor \cdot \rfloor_v \coloneqq \lfloor \frac{\cdot}{v} \rfloor v$ is a gridded flooring operation.

(Huang et al., NeurIPS 2023)

Sketch: Start with LCS technique, and progressively lower bound the quantity. Finally, formulate the maximization problem as knapsack problem.

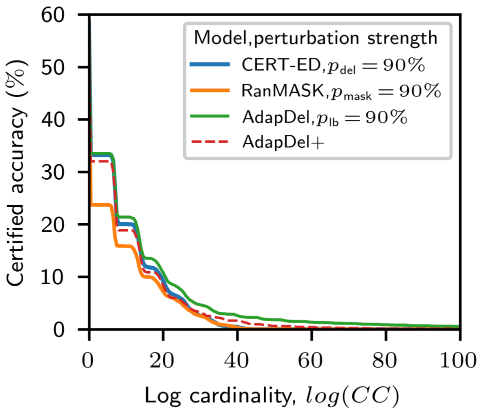
AdaptDel

$$\psi(|x|) = \max\left(p_{lb}, 1 - \frac{k}{|x|}\right)$$

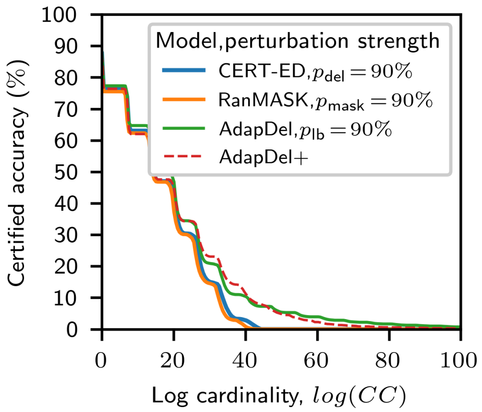
AdaptDel+

$\psi(|x|)$ calibrated automatically by length binning and golden section search

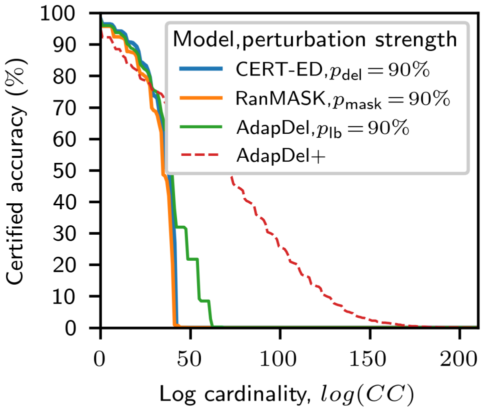
Improved certified accuracy on all datasets



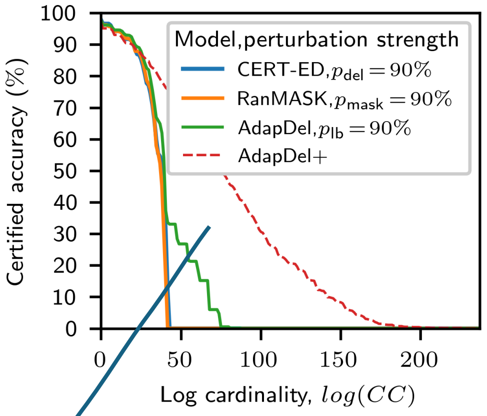
(a) Yelp



(b) IMDB

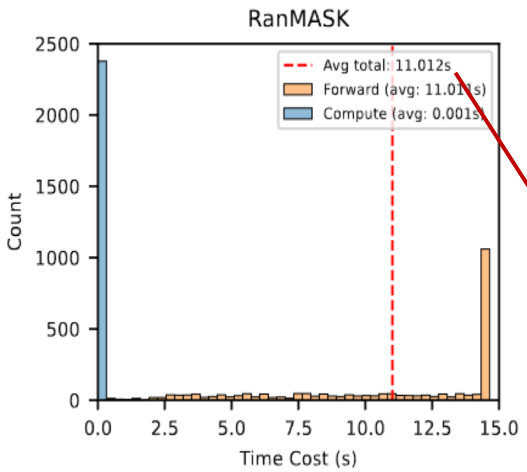


(c) LUN

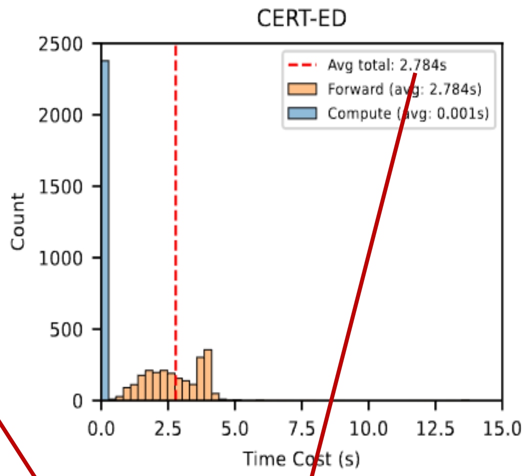


(d) Spam Assassin

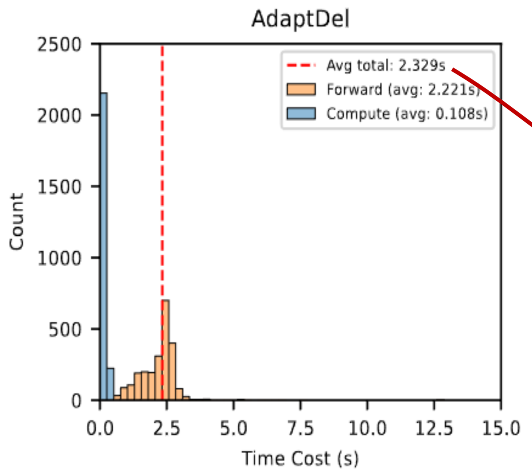
Significant improvement adapting to longer sequences



(a) RanMASK



(b) CERT-ED



(c) AdaptDel

20% and 372% faster