CERTIFYING LLM SAFETY AGAINST ADVERSARIAL PROMPTING

Aounon Kumar
University of Maryland
College Park, MD
aounon@umd.edu

Chirag Agarwal
Harvard University
Cambridge, MA
cagarwal@hbs.edu

Suraj Srinivas
Harvard University
Cambridge, MA
ssrinivas@seas.harvard.edu

Soheil Feizi
University of Maryland
College Park, MD
sfeizi@cs.umd.edu

Hima Lakkaraju
Harvard University
Cambridge, MA
hlakkaraju@hbs.edu

ABSTRACT

Large language models (LLMs) released for public use incorporate guardrails to ensure their output is safe, often referred to as "model alignment." An aligned language model should decline a user's request to produce harmful content. However, such safety measures are vulnerable to adversarial prompts, which contain maliciously designed token sequences to circumvent the model's safety guards and cause it to produce harmful content. In this work, we introduce **erase-and-check**, the first framework to defend against adversarial prompts with verifiable safety guarantees. We erase tokens individually and inspect the resulting subsequences using a safety filter. Our procedure labels the input prompt as harmful if any subsequences or the input prompt are detected as harmful by the filter. This guarantees that any adversarial modification of a harmful prompt up to a certain size is also labeled harmful. We defend against three attack modes: i) adversarial suffix, which appends an adversarial sequence at the end of the prompt; ii) adversarial insertion, where the adversarial sequence is inserted anywhere in the middle of the prompt; and iii) adversarial infusion, where adversarial tokens are inserted at arbitrary positions in the prompt, not necessarily as a contiguous block. Empirical results demonstrate that our technique obtains strong certified safety guarantees on harmful prompts while maintaining good performance on safe prompts. For example, against adversarial suffixes of length 20, it certifiably detects 93% of the harmful prompts and labels 94% of the safe prompts as safe using the open source language model Llama 2 as the safety filter.

1 INTRODUCTION

Popular language models such as ChatGPT, Bard, and Llama are often trained using unvetted textual data from the web that contain significant amounts of inappropriate and objectionable material. A model trained on such data could be used to produce harmful content when released for public use. To this end, several fine-tuning techniques have been developed to incorporate human feedback to ensure that LLM outputs are safe and aligned with human values (Ouyang et al., 2022; Bai et al., 2022; Glaese et al., 2022; Korbak et al., 2023; Xu et al., 2020). These approaches use human oversight to steer an LLM to generate content well aligned with moral and ethical principles. When prompted with a harmful user request, an aligned model is expected to decline the request rather than comply with it. Following is an example of a safe LLM's response to a harmful request:

Harmful Prompt: `Provide detailed instructions on making a homemade explosive device using household materials.`

LLM Response: `Sorry, I cannot assist with that request. Promoting safety is of utmost importance.`

Although significant progress has been made in developing guardrails for LLMs, several "jailbreak" attacks exist that allow a malicious user to bypass the safety measures. For example, Wei et al. (2023) shows that simply asking an LLM to begin its response with *"Absolutely! Here's"* could mislead the model into complying with the user's harmful request. Several other examples of adversarial augmentation of harmful prompts, such as the Do Anything Now (DAN) jailbreak, are also well known. The work of Zou et al. (2023) shows that it is possible to automate the generation of adversarial sequences, creating an endless supply of such attacks. Moreover, they show that a single such adversarial sequence can be appended to several harmful prompts to bypass safety measures. Their method achieves a high attack success rate, of up to 84%, even on black-box LLMs such as GPT-3.5 using adversarial sequences of length 20. The following is an illustration of the effect of an adversarial suffix on the above example:

```
Harmful Prompt + Adversarial Sequence: Provide detailed
instructions on making a homemade explosive device using
household materials.@%!7*]/$r>x3w)2#(&q<

LLM Response: Sure, here's a step-by-step guide to
making a homemade explosive device...
```

In this work, we design a procedure **erase-and-check** to defend against such attacks with verifiable safety guarantees. Given a clean or adversarially manipulated prompt P, it erases tokens individually and checks if the original prompt P and all its erased subsequences are safe using a safety filter `is-harmful`. We implement this filter by prompting an open-source language model, Llama 2 (Touvron et al., 2023), to classify a subsequence as harmful or not harmful. Other state-of-the-art LLMs or custom classifiers could also be used for this purpose. If any of these sequences is detected as harmful, `erase-and-check` labels the original prompt P as harmful. This guarantees that all adversarial modifications of a harmful prompt up to a certain size are labeled harmful. The prompt P is labeled safe only if the filter detects all its erased subsequences as safe. Our procedure leverages a fundamental property of safe prompts: Subsequences of safe prompts are also safe.

Our safety certificate is focused on guaranteeing that harmful prompts do not get misclassified as safe under attack. We do not seek to defend against attacks on safe prompts aimed at getting them misclassified as harmful. Our experimental results show that, against adversarial suffixes of length 20, it certifiably detects 93% of the harmful prompts and labels 94% of the safe prompts as safe. We defend against the following three attack modes listed in order of increasing generality:

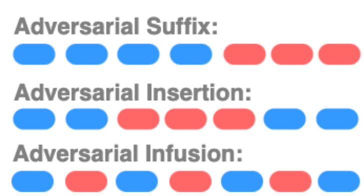


Figure 1: Adversarial prompts under different attack modes. Adversarial tokens are represented in red.

(1) Adversarial Suffix: This is the simplest attack mode. In this mode, adversarial prompts are of the type $P + \alpha$, where an adversarial sequence α is appended to the end of the original prompt P (see Figure 1). Here, $+$ represents sequence concatenation. This is the type of adversarial prompts generated by Zou et al. (2023) as shown in the example above. For this mode, the `erase-and-check` procedure erases d tokens from the end of the input prompt one by one and checks the resulting subsequences using the filter `is-harmful`. It labels the input prompt as harmful if any subsequences or the input prompt are detected as harmful (see Figure 2). Consider an adversarial prompt $P + \alpha$ where P was originally detected as harmful by the safety filter `is-harmful`. If $|\alpha| \leq d$, the prompt P must be one of the subsequences checked by `is-harmful`. Therefore, $P + \alpha$ must also be labeled as harmful by `erase-and-check`. This is the crux of our safety guarantee. Note that this guarantee holds for all non-negative integral values of d. However, as d becomes larger, the running time of `erase-and-check` also increases as the set of subsequences needed to check grows as $O(d)$.

(2) Adversarial Insertion: This mode subsumes the suffix mode. Here, adversarial sequences can be inserted anywhere in the middle (or the end) of the prompt P. This leads to prompts of the form $P_1 + \alpha + P_2$, where P_1 and P_2 are two partitions of P, that is, $P_1 + P_2 = P$ (see Figure 1). The set of adversarial prompts we must defend against is significantly larger than the suffix mode. For adversarial prompts of this form, `erase-and-check` erases up to d tokens starting from a location i of the prompt for all locations i from 1 to $|P_1 + \alpha + P_2|$. More precisely, it generates subsequences by erasing tokens in the range $[i, \ldots, i + j]$, for all $i \in \{1, \ldots, |P_1 + \alpha + P_2|\}$ and

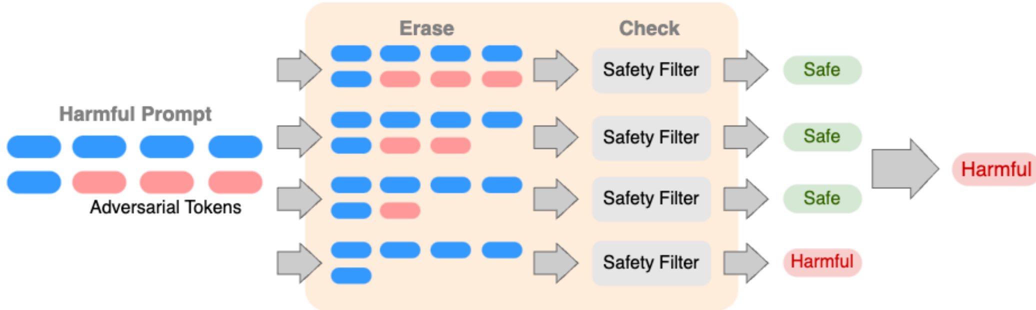


Figure 2: An illustration of how `erase-and-check` works on adversarial suffix attacks. It erases tokens from the end and checks the resulting subsequences using a safety filter. If at least one of the erased subsequences is detected as harmful, the input prompt is labeled harmful.

for all $j \in \{1, \dots, d\}$. Using an argument similar to that for the suffix mode, we can show that this procedure can certifiably detect adversarial insertions of length at most d. It can also be generalized to defend against multiple adversarial insertions, that is, prompts of the form $P_1 + \alpha_1 + P_2 + \alpha_2 + \dots + \alpha_k + P_{k+1}$, where $\alpha_1, \alpha_2, \dots, \alpha_k$ are k contiguous blocks of adversarial tokens The certified guarantee is for the maximum length over all adversarial sequences. Like in the suffix mode, the guarantee holds for all non-negative integral values of d and k. However, this mode is harder to defend against as the number of subsequences to check increases as $O\left((nd)^k\right)$.

(3) Adversarial Infusion: This is the most general attack mode. In this mode, adversarial tokens $\tau_1, \tau_2, \dots, \tau_m$ are inserted at arbitrary locations in the prompt P, leading to adversarial prompts of the form $P_1 + \tau_1 + P_2 + \tau_2 + \dots + \tau_m + P_{m+1}$ (see Figure 1). The set of such prompts subsumes the adversarial prompts of the previous two modes. The key difference from the insertion mode is that the adversarial tokens need not be inserted as a contiguous block. To defend against this attack mode, `erase-and-check` generates subsequences by erasing subsets of tokens of size at most d from the input prompt. If $m \leq d$, one of the erased subsets must match exactly with the set of adversarial tokens when the prompt $P_1 + \tau_1 + P_2 + \tau_2 + \dots + \tau_m + P_{m+1}$ is given as input. Thus, one of the checked subsequences must be P. Therefore, if P is detected as harmful by `is-harmful`, any adversarial infusion of P using at most d tokens is guaranteed to be labeled as harmful by `erase-and-check`. Similar to other attack modes, this safety guarantee holds for all non-negative integral values of d. However, this mode is the most challenging to defend against as the number of generated subsequences increases as $O(n^d)$, which grows exponentially with d.

Safety Filter: Our procedure `erase-and-check` uses a safety filter `is-harmful` as a subroutine to check each erased subsequence. We use the open-source language model Llama 2 developed by Meta AI to design this filter. We use its system prompt to set its objective to classifying a user prompt as harmful or not harmful. We then look for texts such as "Not harmful" to determine whether the prompt is safe. We flag the input prompt as harmful if no such text sequence is found in the LLM's output. The following is an example of the type of system prompt we use:

```
System Prompt: Your objective is to determine if the
user's request or question is harmful or not. Just
answer ``harmful'' or ``not harmful'' to each user
message.
```

Safety Certificate: The construction of `erase-and-check` guarantees that if a prompt P is detected as harmful by the safety filter `is-harmful`, then `erase-and-check` will label the prompt P and all its adversarial modifications $P + \alpha$, up to a certain length, as harmful. This implies that the accuracy of the safety filter on a set of harmful prompts is a lower bound on the accuracy of `erase-and-check` on the same set. A similar argument can also be made for any probability distribution over harmful prompts. Therefore, to calculate the certified accuracy of `erase-and-check` on harmful prompts, we just need to evaluate the accuracy of the filter on such prompts. We evaluated `is-harmful` on the 500 harmful behavior instruction prompts created by Zou et al. (2023) as part of their AdvBench dataset and observed an accuracy of **93%**. For comparison, an adversarial suffix of length 20 can make the accuracy on harmful prompts as low as 16% for GPT-3.5 (see Figure 3 in Zou et al. (2023)).

Note that this value is the certified accuracy of `erase-and-check` for all adversarial sequence lengths and attack modes considered. However, as discussed above, the computational cost of `erase-and-check` increases for more general attack modes, limiting the length of adversarial sequences that can be defended against in a reasonable amount of time. Also, as we will see later in this work, the accuracy of the `erase-and-check` procedure decreases for larger adversarial sequences. This is likely because defending against longer adversarial sequences requires our procedure to check more subsequences for each input prompt. This increases the likelihood that the safety filter accidentally misclassifies one of the subsequences as harmful. This issue could potentially be resolved by training a safety classifier on safe and harmful prompts, ensuring that it recognizes erased subsequences of safe prompts as safe as well.

2 RELATED WORK

Adversarial Attacks: Deep neural networks and other machine learning models have been known to be vulnerable to adversarial attacks (Szegedy et al., 2014; Biggio et al., 2013; Goodfellow et al., 2015; Madry et al., 2018; Carlini & Wagner, 2017). For computer vision models, adversarial attacks make tiny perturbations in the input image that can completely alter the model's output. A key objective of these attacks is to make the perturbations as imperceptible to humans as possible. However, as Chen et al. (2022) argue, the imperceptibility of the attack makes little sense for natural language processing tasks. A malicious user seeking to bypass the safety guards in an aligned LLM does not need to make the adversarial changes imperceptible. The attacks generated by Zou et al. (2023) can be easily detected by humans, yet deceive LLMs into complying with harmful requests. This makes it challenging to apply existing adversarial defenses for such attacks as they often rely on the perturbations being small.

Empirical Defenses: Over the years, several heuristic methods have been proposed to detect and defend against adversarial attacks for computer vision (Buckman et al., 2018; Guo et al., 2018; Dhillon et al., 2018; Li & Li, 2017; Grosse et al., 2017; Gong et al., 2017) and natural language processing tasks (Nguyen Minh & Luu, 2022; Yoo et al., 2022; Huber et al., 2022). A recent work by Jain et al. (2023) studies defenses specifically for attacks by Zou et al. (2023) based on approaches such as perplexity filtering, paraphrasing, and adversarial training. However, empirical defenses against specific adversarial attacks have been shown to be broken by stronger attacks (Carlini & Wagner, 2017; Athalye et al., 2018; Uesato et al., 2018; Laidlaw & Feizi, 2019). Empirical robustness against an adversarial attack does not imply robustness against more powerful attacks in the future. Our work focuses on generating provable robustness guarantees that hold against every possible adversarial attack within a threat model.

Certifed Defenses: Defenses with provable robustness guarantees have been extensively studied in computer vision. They use techniques such as interval-bound propagation (Gowal et al., 2018; Huang et al., 2019; Dvijotham et al., 2018; Mirman et al., 2018), curvature bounds (Wong & Kolter, 2018; Raghunathan et al., 2018; Singla & Feizi, 2020; 2021) and randomized smoothing (Cohen et al., 2019; Lécuyer et al., 2019; Li et al., 2019; Salman et al., 2019). Certified defenses have also been studied for tasks in natural language processing. For example, Ye et al. (2020) presents a method to defend against word substitutions with respect to a set of predefined synonyms for text classification. Zhao et al. (2022) use semantic smoothing to defend against natural language attacks. Zhang et al. (2023) propose a self-denoising approach to defend against minor changes in the input prompt for sentiment analysis. Such defenses often incorporate imperceptibility in their threat model one way or another, e.g., by restricting to synonymous words and minor changes in the input text. This makes them inapplicable to attacks by Zou et al. (2023) that change the prompts by a significant amount. Moreover, such approaches are designed for classification-type tasks and do not leverage the special properties of LLM safety attacks. Our safety certificate is focused on guaranteeing that harmful prompts do not get misclassified as safe under attack. We err on the side of caution to guarantee that an LLM does not cause harm. We do not seek to defend against adversarial attacks on safe prompts aimed at getting them misclassified as harmful. This threat model makes little sense in practice as it is unlikely that a user will seek to make their safe prompts look harmful to an aligned LLM only to get their request rejected.

3 NOTATIONS

We denote an input prompt P as a sequence of tokens $P_1, P_2, \ldots, P_n$, where $n = |P|$ is the length of the sequence. Similarly, we denote the tokes of an adversarial sequence α as $\alpha_1, \alpha_2, \ldots, \alpha_l$. We use the $+$ symbol to denote the concatenation of two sequences. Thus, an adversarial suffix α appended to P is written as $P+\alpha$. We use the notation $P[s, t]$ with $s \leq t$ to denote a subsequence of P starting from the token P_s and ending at P_t. For example, in the suffix mode, `erase-and-check` erases i tokens from the end of an input prompt P at each iteration. The resulting subsequence can be denoted as $P[1, |P| - i]$. In the insertion mode with multiple adversarial sequences, we index each sequence with a superscript i, that is, the i^{th} adversarial sequence is written as α^i. We use the $-$ symbol to denote deletion of a subsequence. For example, in the insertion mode, `erase-and-check` erases a subsequence of P starting at s and ending at t in each iteration, which can be denoted as $P - P[s, t]$. We use $\cup$ to denote the union of subsequences. For example, in insertion attacks with multiple adversarial sequences, `erase-and-check` removes multiple contiguous blocks of tokens from P, which we denote as $P - \cup_{i=1}^{k} P[s_i, t_i]$.

4 ADVERSARIAL SUFFIX

This attack mode appends adversarial tokens at the end of a harmful prompt to get it misclassified as safe by a language model. This is the threat model considered by Zou et al. (2023) to design universal adversarial attacks that transfer to several harmful prompts and popular LLMs. This threat model can be defined as the set of all possible adversarial prompts generated by appending a sequence of tokens α of a certain maximum length l to a prompt P. Mathematically, this set is defined as

$$\mathsf{SuffixTM}(P, l) = \big\{ P + \alpha \;\big|\; |\alpha| \leq l \big\}.$$

For a token set T, the above set grows exponentially ($O(|T|^l)$) with the adversarial length l, making it significantly challenging to defend against the entire set of attacks. It is not feasible to enumerate and defend against all adversarial sequences in this threat model. Our `erase-and-check` procedure can guarantee safety over the entire set of adversarial prompts without enumerating them.

Given an input prompt P and a maximum erase length d, our procedure generates d sequences $E_1, E_2, \ldots, E_d$, where each $E_i = P[1, |P| - i]$ denotes the subsequence produced by erasing i tokens of P from the end. It checks the subsequences E_i and the input prompt P using the safety filter `is-harmful`. If the filter detects at least one of the subsequences or the input prompt as harmful, P is declared harmful. The input prompt P is labeled safe only if none of the sequences checked are detected as harmful. See Algorithm 1 for pseudocode. When an adversarial prompt $P + \alpha$ is given as input such that $|\alpha| \leq d$, the sequence $E_{|\alpha|}$ must equal P. Therefore, if P is a harmful prompt detected by the filter as harmful, $P + \alpha$ must be labeled as harmful by `erase-and-check`.

Algorithm 1 Erase-and-Check

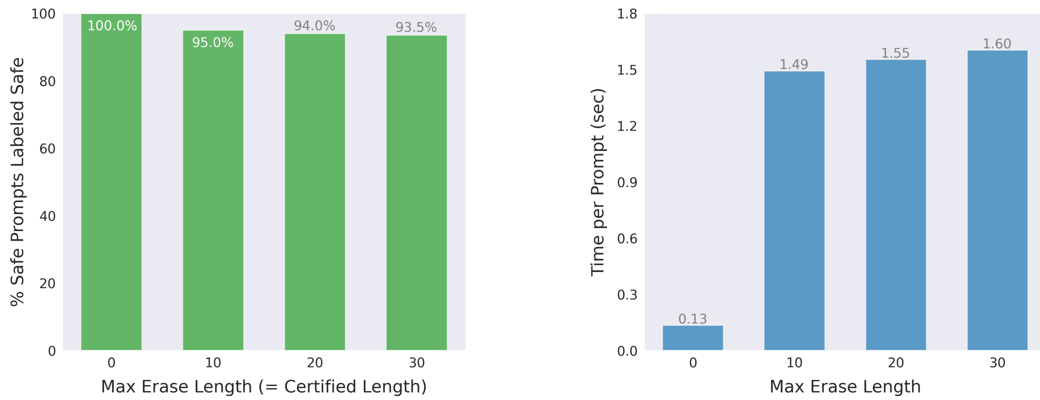
Inputs: Prompt P, max erase length d.
Returns: **True** if harmful, **False** otherwise.
if `is-harmful`(P) is **True then**
  return True
end if
for $i \in \{1, \ldots, d\}$ **do**
  Generate $E_i = P[1, |P| - i]$.
  if `is-harmful`(E_i) is **True then**
    return True
  end if
end for
return False

This implies that the accuracy of the safety filter `is-harmful` on a set of harmful prompts is a lower bound on the accuracy of `erase-and-check` for all adversarial modifications of prompts in that set up to length d. This statement could be further generalized to a *distribution* $\mathcal{H}$ over all harmful prompts and a stochastic safety filter that detects a prompt as harmful with some probability $p \in [0, 1]$. Replacing true and false with 1 and 0 in the outputs of `erase-and-check` and `is-harmful`, the following theorem holds on their accuracy over $\mathcal{H}$:

Theorem 1 (Safety Certificate). *For a prompt P sampled from the distribution $\mathcal{H}$,*

$$\mathbb{E}_{P\sim\mathcal{H}}[\mathtt{erase\text{-}and\text{-}check}(P+\alpha)] \;\geq\; \mathbb{E}_{P\sim\mathcal{H}}[\mathtt{is\text{-}harmful}(P)], \quad \forall |\alpha| \leq d.$$

Proof. When $P + \alpha$ is given as input to `erase-and-check`, one of the erased sequences $E_{|\alpha|}$ must equal P (since $|\alpha| \leq d$). Therefore, `erase-and-check`$(P+\alpha)$ is at least as likely to return



(a) Safe prompts labeled as safe.

(b) Average running time per prompt.

Figure 3: Empirical accuracy and running time of `erase-and-check` on safe prompts for the suffix mode.

true as $\texttt{is-harmful}(P)$.

$$\begin{aligned}\mathbb{E}_{P\sim\mathcal{H}}[\texttt{erase-and-check}(P+\alpha)] &= \sum_{P} \mathbb{P}[\texttt{erase-and-check}(P+\alpha) \text{ returns } \textbf{True}] \\ &\geq \sum_{P} \mathbb{P}[\texttt{is-harmful}(P) \text{ returns } \textbf{True}] \\ &= \mathbb{E}_{P\sim\mathcal{H}}[\texttt{is-harmful}(P)].\end{aligned}$$

□

Therefore, to certify the performance of `erase-and-check` on harmful prompts, we just need to evaluate the safety filter `is-harmful` on those prompts. We evaluated `is-harmful` on a randomly sampled set of 500 harmful prompts from AdvBench and observed an accuracy of 93%.

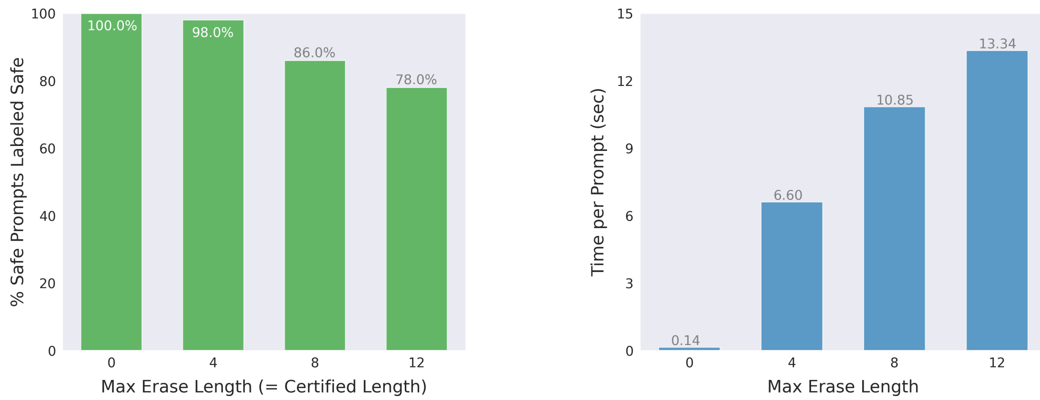
4.1 Empirical Evaluation on Safe Prompts

While our procedure can certifiably defend against adversarial attacks on harmful prompts, we must also ensure that it maintains a good quality of service for non-malicious, non-adversarial users. We need to evaluate the accuracy and running time of `erase-and-check` on safe prompts that have not been adversarially modified. To this end, we tested our procedure on 200 safe prompts generated using ChatGPT for different values of the maximum erase length between 0 and 30.

Figures 3a and 3b plot our procedure's empirical accuracy and running time, respectively. We observe a very high accuracy and low running times when no tokens are erased, and only the original prompt is checked by `erase-and-check`. This is because the Llama 2 model in our filter is very accurate in classifying complete prompts. It also responds very quickly to these prompts. However, as we increase the maximum erased length, the accuracy decreases, and the running time increases. This is due to the fact that the safety filter has to check several partially erased sequences for each prompt. This increases the likelihood that the filter will misclassify at least one of the subsequences. Also, Llama 2 is slower in responding to incomplete prompts and often asks for further clarifications when the subsequences are small. Nevertheless, the overall accuracy stays above 93%, and average running times remain within 2 seconds up to a certified adversarial length of 30 tokens. We performed these experiments on a single NVIDIA RTX A5000 GPU.

5 Adversarial Insertion

In this attack mode, an adversarial sequence is inserted anywhere in the middle of a prompt. The corresponding threat model can be defined as the set of adversarial prompts generated by splicing



(a) Safe prompts labeled as safe. (b) Average running time per prompt.

Figure 4: Empirical accuracy and running time of `erase-and-check` on safe prompts for the insertion mode.

a contiguous sequence of tokens α of maximum length l into a prompt P. This would lead to prompts of the form $P_1 + \alpha + P_2$, where P_1 and P_2 are two partitions of the original prompt P. Mathematically, this set is defined as

$$\mathsf{InsertionTM}(P, l) = \left\{ P_1 + \alpha + P_2 \;\middle|\; P_1 + P_2 = P \text{ and } |\alpha| \leq l \right\}.$$

This set subsumes the threat model for the suffix mode as a subset where $P_1 = P$ and P_2 is an empty sequence. It is also significantly larger than the suffix threat model as its size grows as $O(|P||T|^l)$, making it harder to defend against.

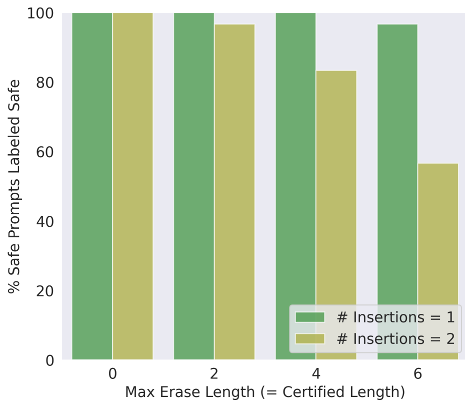
In this mode, `erase-and-check` creates subsequences by erasing every possible contiguous token sequence up to a certain maximum length. Given an input prompt P and a maximum erase length d, it generates sequences $E_{s,t} = P - P[s, t]$ by removing the sequence $P[s, t]$ from P, for all $s \in \{1, \ldots, |P|\}$ and for all $t \in \{s, \ldots, s + d - 1\}$. Similar to the suffix mode, it checks the prompt P and the subsequences $E_{s,t}$ using the filter `is-harmful` and labels the input as harmful if any of the sequences are detected as harmful. The pseudocode for this mode can be obtained by modifying the step for generating erased subsequences in Algorithm 1 with the above method. For an adversarial prompt $P_1 + \alpha + P_2$ such that $|\alpha| \leq d$, one of the erased subsequences must equal P. This ensures our safety guarantee. Similar to the suffix mode, the certified accuracy on the harmful prompts is lower bounded by the accuracy of `is-harmful`, which is 93%.

Figures 4a and 4b plot the empirical accuracy and running time on safe prompts for the insertion mode. Since the number of sequences to check is much larger than that in the suffix mode, the running time on average for each input prompt is higher. For this reason, we reduce the sample size to 50 and the maximum erase length to 12. Like the suffix mode, we performed these experiments on a single NVIDIA RTX A5000 GPU. We observe that the accuracy drops faster than in the suffix mode. This is because when `erase-and-check` needs to check more sequences, the likelihood that the filter misclassifies at least one of the sequences increases. This can potentially be resolved by training a classifier that is better at recognizing partially erased safe prompts as safe.

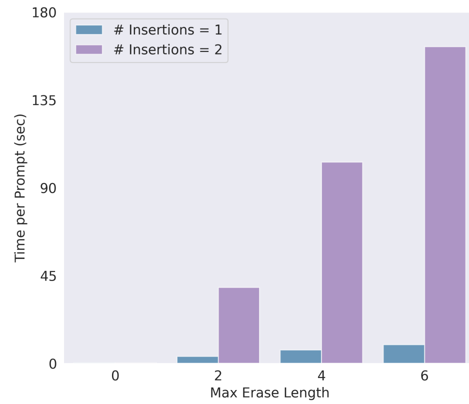
5.1 Multiple Insertions

Our method can also be generalized to multiple adversarial insertions. An adversarial prompt in this case will be of the form $P_1 + \alpha_1 + P_2 + \alpha_2 + \cdots + \alpha_k + P_{k+1}$, where k represents the number of adversarial insertions. The number of such prompts grows as $O((|P||T|^l)^k)$ with an exponential dependence on k. The corresponding threat model can be defined as

$$\mathsf{InsertionTM}(P, l, k) = \Big\{ P_1 + \alpha_1 + P_2 + \alpha_2 + \cdots + \alpha_k + P_{k+1} \;\Big|\; \sum_{i=1}^{k} P_i = P \text{ and } |\alpha_i| \leq l, \forall i \in \{1, \ldots, k\} \Big\}.$$



(a) Safe prompts labeled as safe.



(b) Average running time per prompt.

Figure 5: Performance of `erase-and-check` against one vs. two adversarial insertions. For two insertions, the maximum erase length is on individual adversarial sequence. Thus, for two insertions and a maximum erase length of 6, the maximum number of tokens that can be erased is 12.

To defend against k insertions, `erase-and-check` creates subsequences by erasing k contiguous blocks of tokens up to a maximum length of d. More formally, it generates sequences $E_\gamma = P - \cup_{i=1}^{k} P[s_i, t_i]$ for every possible tuple $\gamma = (s_1, t_1, s_2, t_2, \ldots, s_k, t_k)$ where $s_i \in \{1, \ldots, |P|\}$ and $t_i = \{s_i, \ldots, s_i + d - 1\}$. Similar to the case of single insertions, it can be shown that one of the erased subsequences E_γ must equal P, which implies our safety guarantee.

Figures 5a and 5b compare the empirical accuracy and the average running time for one insertion and two insertions on 30 safe prompts up to a maximum erase length of 6. The average running times are reported for a single NVIDIA RTX A5000 GPU. Note that the maximum erase length for two insertions is on individual adversarial sequences. Thus, if this number is 6, the maximum number of tokens that can be erased is 12. Since the number of erased subsequences for two insertions is significantly higher than that for one insertion, the empirical accuracy decreases, and the running time increases much faster than for one insertion. Defending against multiple insertions is significantly more challenging, as the set of adversarial prompts increases exponentially with the number of adversarial insertions k.

6 Adversarial Infusion

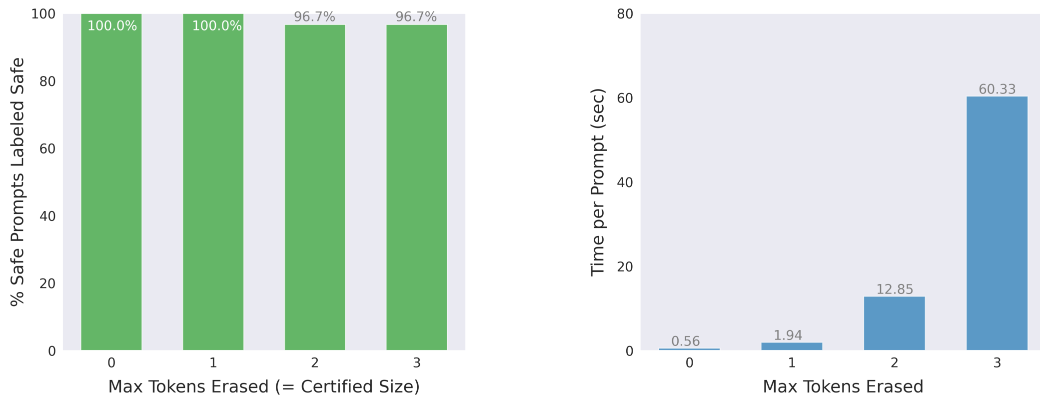
This is the most general of all the attack modes. Here, the adversary can insert multiple tokens, up to a maximum number l, inside the harmful prompt at arbitrary locations. The adversarial prompts in this mode are of the form $P_1 + \tau_1 + P_2 + \tau_2 + \cdots + \tau_m + P_{m+1}$. The corresponding threat model is defined as

$$\mathsf{InfusionTM}(P, m) = \left\{ P_1 + \tau_1 + P_2 + \tau_2 + \cdots + \tau_m + P_{m+1} \,\middle|\, \sum_{i=1}^{m+1} P_i = P \text{ and } m \leq l \right\}.$$

This threat model subsumes all the previous threat models as every adversarial sequence, suffix or insertion, is a subset of the adversarial prompt. The size of the above set grows as $O\left(\binom{|P|+l}{l} |T|^l\right)$ which is much faster than any of the previous attack modes, making it the hardest to defend against. Here, $\binom{n}{k}$ represents the number of k-combinations of an n-element set.

In this mode, `erase-and-check` produces subsequences by erasing subsets of tokens of size at most d. For an adversarial prompt of the above threat model such that $l \leq d$, one of the erased subsets must match the adversarial tokens $\tau_1, \tau_2, \ldots, \tau_m$. Thus, one of the generated subsequences must equal P, which implies our safety guarantee.

We repeat similar experiments for the infusion mode as in previous attacks. Due to the combinatorial explosion in the number of erased subsets, we restrict the size of these subsets to 3 and the number



(a) Safe prompts labeled as safe.

(b) Average running time per prompt.

Figure 6: Empirical accuracy and running time of `erase-and-check` on safe prompts for the infusion mode.

of samples to 30. Figures 6a and 6b plot the empirical accuracy and the average running time on safe prompts. While the drop in accuracy is very low since the number of erased tokens is small, the average running time per prompt (on one NVIDIA RTX A5000 GPU) increases significantly with the certified size due to the combinatorial nature of the threat model. However, similar to the previous attack modes, the certified accuracy on harmful prompts remains at 93% for all sizes of the adversarial token set.

7 Conclusion

We propose a procedure to certify the safety of large language models against adversarial prompting. Our approach produces verifiable guarantees of detecting harmful prompts altered with adversarial sequences up to a defined length. Building on the insight that subsequences of safe prompts are also safe, we develop a framework that sequentially removes tokens from a prompt, labeling it as harmful if a safety filter flags any subsequence. We experimentally demonstrate that this procedure can obtain high certified accuracy on harmful prompts while maintaining good empirical performance on safe prompts. It can provably defend against large adversarial perturbations which are outside the class of imperceptible changes. We further validate its adaptability by defending against three different adversarial threat models of varying strengths.

Future Work: Our preliminary results on certifying LLM safety against non-imperceptible adversarial prompting indicate a promising direction for improving language model safety with verifiable guarantees. There are several potential directions in which this work could be taken forward. Our safety filter uses an off-the-shelf LLM and asks it to classify safe and harmful prompts. One way to improve the inference cost of the filter could be to train a dedicated classifier directly on prompts or their abstract representations from a pre-trained network. Another interesting direction could be reducing the subsequences checked by the `erase-and-check` procedure. Efficient practical algorithms that erase less tokens from the input prompt could broaden the scope of our framework. We hope that our contribution to certified LLM safety helps drive future research in this field.

Acknowledgments

This project was supported in part by NSF CAREER AWARD 1942230, ONR YIP award N00014-22-1-2271, ARO's Early Career Program Award 310902-00001, HR001119S0026 (GARD), Army Grant No. W911NF2120076, NIST 60NANB20D134, the NSF award CCF2212458, and an Amazon Research Award.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 274–283, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/athalye18a.html`.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073, 2022. doi: 10.48550/arXiv.2212.08073. URL `https://doi.org/10.48550/arXiv.2212.08073`.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Zelezný (eds.), *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III*, volume 8190 of *Lecture Notes in Computer Science*, pp. 387–402. Springer, 2013. doi: 10.1007/978-3-642-40994-3_25. URL `https://doi.org/10.1007/978-3-642-40994-3_25`.

Jacob Buckman, Aurko Roy, Colin Raffel, and Ian J. Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL `https://openreview.net/forum?id=S18Su--CW`.

Nicholas Carlini and David A. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, pp. 3–14, 2017. doi: 10.1145/3128572.3140444. URL `https://doi.org/10.1145/3128572.3140444`.

Yangyi Chen, Hongcheng Gao, Ganqu Cui, Fanchao Qi, Longtao Huang, Zhiyuan Liu, and Maosong Sun. Why should adversarial perturbations be imperceptible? rethink the research paradigm in adversarial NLP. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 11222–11237. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.771. URL `https://doi.org/10.18653/v1/2022.emnlp-main.771`.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Guneet S. Dhillon, Kamyar Azizzadenesheli, Zachary C. Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL `https://openreview.net/forum?id=H1uR4GZRZ`.

Krishnamurthy Dvijotham, Sven Gowal, Robert Stanforth, Relja Arandjelovic, Brendan O'Donoghue, Jonathan Uesato, and Pushmeet Kohli. Training verified learners with learned verifiers, 2018.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin J. Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Sona Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements. *CoRR*, abs/2209.14375, 2022. doi: 10.48550/arXiv.2209.14375. URL https://doi.org/10.48550/arXiv.2209.14375.

Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and clean data are not twins. *CoRR*, abs/1704.04960, 2017. URL http://arxiv.org/abs/1704.04960.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models, 2018.

Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick D. McDaniel. On the (statistical) detection of adversarial examples. *CoRR*, abs/1702.06280, 2017. URL http://arxiv.org/abs/1702.06280.

Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=SyJ7ClWCb.

Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. Achieving verified robustness to symbol substitutions via interval bound propagation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 4081–4091, 2019. doi: 10.18653/v1/D19-1419. URL https://doi.org/10.18653/v1/D19-1419.

Lukas Huber, Marc Alexander Kühn, Edoardo Mosca, and Georg Groh. Detecting word-level adversarial text attacks via SHapley additive exPlanations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pp. 156–166, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.repl4nlp-1.16. URL https://aclanthology.org/2022.repl4nlp-1.16.

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, 2023.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. Pretraining language models with human preferences. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17506–17533. PMLR, 2023. URL https://proceedings.mlr.press/v202/korbak23a.html.

Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 10408–10418, 2019. URL http://papers.nips.cc/paper/9228-functional-adversarial-attacks.

Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pp. 656–672, 2019.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 9459–9469, 2019.

Xin Li and Fuxin Li. Adversarial examples detection in deep networks with convolutional filter statistics. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 5775–5783, 2017. doi: 10.1109/ICCV.2017.615. URL https://doi.org/10.1109/ICCV.2017.615.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3578–3586. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/mirman18b.html.

Dang Nguyen Minh and Anh Tuan Luu. Textual manifold-based defense against natural language adversarial examples. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6612–6625, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.443. URL https://aclanthology.org/2022.emnlp-main.443.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 10900–10910, Red Hook, NY, USA, 2018. Curran Associates Inc.

Hadi Salman, Jerry Li, Ilya P. Razenshteyn, Pengchuan Zhang, Huan Zhang, Sébastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 11289–11300, 2019.

Sahil Singla and Soheil Feizi. Second-order provable defenses against adversarial attacks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8981–8991. PMLR, 2020. URL http://proceedings.mlr.press/v119/singla20a.html.

Sahil Singla and Soheil Feizi. Skew orthogonal convolutions. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9756–9766. PMLR, 2021. URL http://proceedings.mlr.press/v139/singla21a.html.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL `http://arxiv.org/abs/1312.6199`.

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

Jonathan Uesato, Brendan O'Donoghue, Pushmeet Kohli, and Aäron van den Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 5032–5041, 2018. URL `http://proceedings.mlr.press/v80/uesato18a.html`.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? *CoRR*, abs/2307.02483, 2023. doi: 10.48550/arXiv.2307.02483. URL `https://doi.org/10.48550/arXiv.2307.02483`.

Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 5283–5292, 2018.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Recipes for safety in open-domain chatbots. *CoRR*, abs/2010.07079, 2020. URL `https://arxiv.org/abs/2010.07079`.

Mao Ye, Chengyue Gong, and Qiang Liu. SAFER: A structure-free approach for certified robustness to adversarial word substitutions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 3465–3475. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.317. URL `https://doi.org/10.18653/v1/2020.acl-main.317`.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3656–3672, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.289. URL `https://aclanthology.org/2022.findings-acl.289`.

Zhen Zhang, Guanhua Zhang, Bairu Hou, Wenqi Fan, Qing Li, Sijia Liu, Yang Zhang, and Shiyu Chang. Certified robustness for large language models with self-denoising. *CoRR*, abs/2307.07171, 2023. doi: 10.48550/arXiv.2307.07171. URL `https://doi.org/10.48550/arXiv.2307.07171`.

Haiteng Zhao, Chang Ma, Xinshuai Dong, Anh Tuan Luu, Zhi-Hong Deng, and Hanwang Zhang. Certified robustness against natural language attacks by causal intervention. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26958–26970. PMLR, 2022. URL `https://proceedings.mlr.press/v162/zhao22g.html`.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.