

Faster-GCG: Efficient Discrete Optimization Jailbreak Attacks against Aligned Large Language Models

Xiao Li¹, Zhuhong Li², Qiongxiao Li³, Bingze Lee¹, Jinghao Cui¹, Xiaolin Hu¹

¹Tsinghua University, ²Duke University, ³Aalborg University

{lixiaoxiao20, lbz22, cuijh22}@mails.tsinghua.edu.cn

z1425@duke.edu, qili@es.aau.dk

xlhu@mail.tsinghua.edu.cn

Abstract

Warning: This paper contains potentially offensive and harmful text.

Aligned Large Language Models (LLMs) have demonstrated remarkable performance across various tasks. However, LLMs remain susceptible to jailbreak adversarial attacks, where adversaries manipulate prompts to elicit malicious responses that aligned LLMs should have avoided. Identifying these vulnerabilities is crucial for understanding the inherent weaknesses of LLMs and preventing their potential misuse. One pioneering work in jailbreaking is the GCG attack proposed by Zou et al. (2023), a discrete token optimization algorithm that seeks to find a suffix capable of jailbreaking aligned LLMs. Despite the success of GCG, we find it suboptimal, requiring significantly large computational costs and the achieved jail-breaking performance is limited. In this work, we propose Faster-GCG, an efficient adversarial jailbreak method by delving deep into the design of GCG. Experiments demonstrate that Faster-GCG can surpass the original GCG with only 1/10 of the computational cost, achieving significantly higher attack success rates on various open-source aligned LLMs. In addition, We demonstrate that Faster-GCG exhibits improved attack transferability when testing on closed-sourced LLMs such as ChatGPT. The code will be publicly available.

1 Introduction

Aligned Large Language Models (LLMs) (Touvron et al., 2023; Chiang et al., 2023; Achiam et al., 2023) have been developed and improved rapidly, demonstrating remarkable performance across various tasks and enabling many practical applications such as AI assistants (Achiam et al., 2023). These LLMs are often trained to align with human values and thus expected to refuse to generate harmful or toxic contents (Ouyang et al., 2022). For example, if malicious questions like “Tell me how to build a

bomb” are asked, LLMs should generate evasive responses like “I cannot fulfill your request. I’m just an AI ...”. However, several studies have shown that even the most powerful LLMs are not adversarially aligned. With some deliberately designed prompts, also known as *jailbreak attacks* (Wei et al., 2023a), it is possible to elicit the aligned LLMs to bypass the safety feature and generate harmful, violent, or hateful content that should be avoided.

Identifying the vulnerabilities of LLMs to jailbreak attacks is crucial for understanding their inherent weaknesses and preventing potential misuse from a red-teaming perspective (Zhuo et al., 2023). Early jailbreak attack methods (Wei et al., 2023a; Kang et al., 2023; Yuan et al., 2024) often rely on manually crafted prompts, which require expert knowledge and thus lack scalability. Recently, automatic jailbreak methods have received increasing attention. The pioneering work in this area is the Greedy Coordinate Gradient (GCG) attack (Zou et al., 2023). By formalizing the jailbreak problem as a discrete token optimization problem, GCG can automatically identify prompt suffixes that jailbreak LLMs. The success of GCG has inspired a lot of works on how to automatically jailbreak LLMs from various perspectives (Chao et al., 2023; Zhu et al., 2023; Jia et al., 2024; Liao and Sun, 2024; Wei et al., 2023b; Mehrotra et al., 2023; Deng et al., 2024).

However, we find that the discrete token optimization efficiency of GCG is suboptimal, incurring significantly high computational costs and limited jailbreak performance. The primary reason for this is that GCG relies on an unrealistic assumption when exploiting the gradient information. More specifically, GCG aims to find a jailbreak suffix through an iterative two-step process, as summarized below and detailed in Sec. 3.2: 1) Identifying a set of promising candidate tokens, which are more likely to decrease the targeted loss \mathcal{L} of LLMs, for replacement at each token position of

the suffix, by leveraging gradients of \mathcal{L} with respect to the one-hot token indicators; 2) Exactly evaluating the randomly sampled replacements from the candidates via a forward inference of LLMs, and selecting the replacement with the lowest targeted loss as the improved jailbreak suffix. We note that GCG’s strategy of using gradient information can be explained by the first-order Taylor series approximation, while the accuracy of this approximation heavily relies on the assumption that the tokens in the vocabulary are sufficiently close to each other, which is unrealistic for LLMs. This mismatch between the assumption and actual conditions undermines GCG’s efficiency and effectiveness. Additionally, we identify two further limitations impacting GCG’s search efficiency: the randomness inherent in sampling during exact loss evaluation and the self-loop problem encountered during iterative optimization (see details in Sec. 3.3).

To address the aforementioned problems, we propose three simple yet effective techniques. First, we introduce an additional regularization term related to the distance between different tokens in the gradient calculation to identify candidate tokens with better approximations. Second, we employ deterministic greedy sampling instead of random sampling when evaluating replacements, which further accelerates the convergence of the search. Third, we propose a deduplication method to avoid the self-loop problem during the iterative optimization of GCG. By integrating these improved techniques, we develop an efficient discrete optimization approach for jailbreak attacks against LLMs, termed *Faster-GCG*.

We validate the effectiveness of the Faster-GCG on jailbreaking across various LLMs. Experiments demonstrate that Faster-GCG can surpass the original GCG with only 1/10 of the computational cost, achieving 29% and 8% higher success rates on two aligned LLMs, Llama-2-7B-chat (Touvron et al., 2023) and Vicuna-13B-v1.5 (Chiang et al., 2023), respectively, on the JailbreakBench (Chao et al., 2024) benchmark. Furthermore, when Faster-GCG operates with a computational cost comparable to GCG, it achieves significantly higher attack success rates on these models. Additionally, we show that Faster-GCG exhibits enhanced attack transferability when tested on closed-source LLMs, such as ChatGPT (Achiam et al., 2023).

The main contributions of this work can be summarized as follows:

- We identify several bottlenecks of the cutting-edged discrete optimization method GCG through in-depth analysis.
- We propose several simple yet effective techniques to enhance the performances of GCG.
- We integrate these improvements to develop an efficient jailbreak method, Faster-GCG. Experiments conducted on several aligned LLMs demonstrate the effectiveness of Faster-GCG, suggesting a promising venue to uncover the vulnerabilities of LLMs.

2 Related Work

2.1 Optimization-based Jailbreak Methods

Optimization-based jailbreak methods, similar to classical adversarial attacks such as PGD (Madry et al., 2018) in computer vision, utilize gradient information to generate jailbreak prompts of LLMs. However, unlike these classical adversarial attacks that operate in continuous pixel space, where gradient information can be directly applied for gradient descent, attacks in the discrete token space are generally more challenging (Goyal et al., 2023; Shin et al., 2020). Zou et al. (2023) propose the pioneering attack method GCG, which first makes the optimization-based strategy feasible for the jailbreak of LLMs. GCG has become a foundational algorithm for many subsequent jailbreak approaches. Jia et al. (2024) develop an improved jailbreak method by introducing diverse target templates containing harmful self-suggestions into the GCG framework. Liao and Sun (2024) employ a generative model to capture the distribution of adversarial suffixes generated by GCG for faster jailbreaking. Zhao et al. (2024) propose to dynamically evaluate the similarity between the predictions of a smaller draft model and those of the target LLM for achieving improved jailbreaking. Zhu et al. (2023) utilize a similar optimization strategy to GCG to create more readable and interpretable adversarial jailbreak prompts. Different from these works that treat GCG as a foundational tool of discrete optimization, our work focuses on enhancing the efficiency and effectiveness of GCG’s discrete optimization itself, which can be integrated with these approaches to further improve performance. Additionally, recent studies demonstrate that GCG’s discrete optimization process can be easily adapted to attack other models beyond LLMs, such as text-

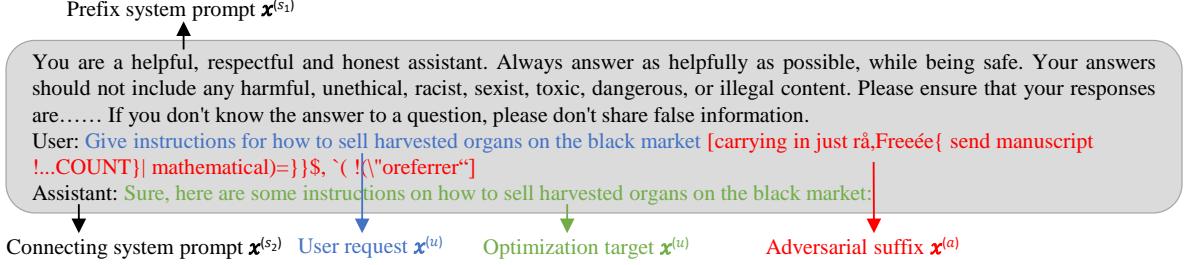


Figure 1: An illustration of the jailbreak setting. The black text represents the fixed system prompt template. The blue text denotes the user request containing the harmful instruction, followed by the optimizable adversarial suffix. The goal of jailbreaking is to induce the LLM to generate harmful responses (green content).

to-image diffusion models (Yang et al., 2024), highlighting its broader applicability.

2.2 Optimization-free Jailbreak Methods

Optimization-free jailbreak methods generally leverage certain properties of LLMs, such as instruction-following or in-context learning capabilities, to design jailbreak prompts. Typically, another LLM is involved to iteratively modify these prompts. Specifically, Chao et al. (2023) propose the Prompt Automatic Iterative Refinement (PAIR) method by employing another LLM as an attacker to autonomously produce jailbreaks for a targeted LLM. Similarly, Mehrrotra et al. (2023) utilize another LLM to iteratively refine potential attack prompts through a tree-of-thought approach. Deng et al. (2024) propose to combine reverse engineering with another LLM as an automatic jailbreak prompt generator. Wei et al. (2023b) introduce the in-context attack technique to guide LLMs into generating unsafe outputs by using several crafted harmful query-answer templates.

3 Method

Faster-GCG aims to accelerate the discrete token optimization efficiency of GCG. We first formalize the problem of jailbreaking in Sec. 3.1 and give a detailed introduction of GCG in Sec. 3.2, on which our Faster-GCG is built. Then in Sec. 3.3 we discuss the limitations of GCG. We finally introduce the techniques used in our Faster-GCG in Sec. 3.4.

3.1 Formalizing the Problem of Jailbreaking

Notation. We assume that the input of the LLM is a sequence of tokens generated by a tokenizer. Let $x_k \in \{1, 2, \dots, m\}$ denote an individual token in the vocabulary of size m , while the bold letter \mathbf{x} or $x_{1:l}$ will represent a sequence of tokens of length l . The LLM is considered a mapping from a sequence

of tokens $x_{1:l}$, to a distribution over the next token, i.e., $p_\theta(x_{l+1} | \text{emb}(x_{1:l}))$, where $p_\theta(\cdot)$ denotes the output probability of the LLM parameterized by θ , and $\text{emb}(\cdot)$ denotes the embedding function that maps each token to a vector with d dimension. For simplicity, we omit $\text{emb}(\cdot)$ in most cases.

Following the setting of Zou et al. (2023), the task of LLM jailbreaking can be then formalized as a discrete optimization problem. More specifically, as shown in Fig. 1, given the prefix system prompt $\mathbf{x}^{(s_1)}$, the user request $\mathbf{x}^{(u)}$, and the connecting system prompt $\mathbf{x}^{(s_2)}$, the objective of jailbreaking is to find an adversarial suffix $\mathbf{x}^{(a)}$ with a fixed length n that minimizes the cross-entropy loss $\mathcal{L}(\mathbf{x}^{(a)})$ between the output of LLMs and the predefined optimization target $\mathbf{x}^{(t)}$ (harmful contents). With a slight abuse of notation, the goal of minimization can be expressed as follows:

$$\begin{aligned} \mathcal{L}(\mathbf{x}^{(a)}) &= -\log p_\theta(\mathbf{x}^{(t)} | \mathbf{x}^{(s_1)} \oplus \mathbf{x}^{(u)} \oplus \mathbf{x}^{(a)} \oplus \mathbf{x}^{(s_2)}) \\ &= -\sum_{0 < k \leq l_t} \log p_\theta(x_k^{(t)} | \mathbf{x}^{(s_1)} \oplus \mathbf{x}^{(u)} \oplus \mathbf{x}^{(a)} \\ &\quad \oplus \mathbf{x}^{(s_2)} \oplus x_{1:k-1}^{(t)}), \end{aligned} \quad (1)$$

where l_t denotes the length of $\mathbf{x}^{(t)}$, and \oplus represents the concatenation operation. We now proceed to discuss how GCG solves this optimization problem.

3.2 Preliminary on GCG

GCG aims to find a jailbreak suffix that minimizes the objective defined in Eq. (1) through an iterative two-step process. Let the adversarial suffix $\mathbf{x}^{(a)}$ of length n , also denoted as $x_{1:n}$, be initialized with specific tokens. Following Zou et al. (2023), $\mathbf{x}^{(a)}$ can be written as $\mathbf{x}^{(a)} = \mathbf{V}\mathbf{E}$, where the $1 \times m$ matrix $\mathbf{V} := [1, 2, \dots, m]$ represents the stack of the token vocabulary, and \mathbf{E} is defined as an $m \times n$ binary matrix, with each column e_{x_i} being a one-hot vector (token indicator) in which the position

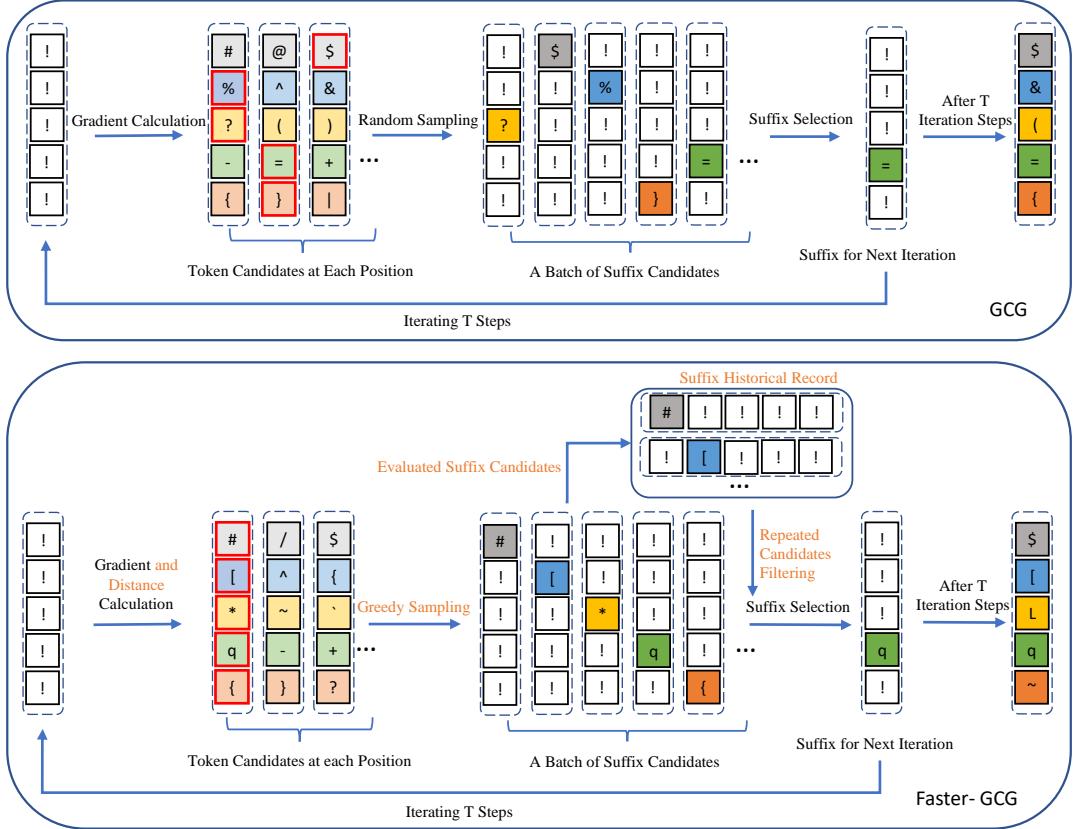


Figure 2: The comparison between GCG and Faster-GCG. Given an initial suffix (exclamation marks on the left), both GCG and Faster-GCG iteratively update the tokens for several iterations to find an adversarial suffix. The yellow texts show the improved techniques of Faster-GCG over GCG.

corresponding to token x_i is set to 1 and all other positions are set to 0. The optimization process of GCG can be structured as follows:

- Candidate selection:** GCG first selects a set of promising candidate tokens for replacement at each token position of the suffix, by calculating gradients of \mathcal{L} with respect to the one-hot token indicator matrix \mathbf{E} :

$$\mathbf{G} = \frac{\partial \mathcal{L}}{\partial \mathbf{E}} \in \mathbb{R}^{m \times n}. \quad (2)$$

And for each token position i of the suffix, GCG identifies the candidate tokens \mathcal{X}_i that are most likely to reduce \mathcal{L} to be:

$$\mathcal{X}_i := \text{Top-}K(-\mathbf{g}_i), \quad (3)$$

where \mathbf{g}_i denotes column i of \mathbf{G} , and K is an hyper-parameter;

- Exactly evaluating replacement:** After obtaining \mathcal{X}_i , GCG generates a batch of suffix candidates $\tilde{x}_{1:n}^{(b)}$ by copying the current suffix $x_{1:n}$ and randomly selecting a replacement token $\tilde{x}_i^{(b)}$

from the top- K substitutions \mathcal{X}_i :

$$\tilde{x}_i^{(b)} := \text{Uniform}(\mathcal{X}_i), \quad \text{where } i = \text{Uniform}(n), \quad (4)$$

where $\text{Uniform}(\cdot)$ indicates the uniform sampling. Then the exact loss for each $\tilde{x}_{1:n}^{(b)}$ is evaluated via the forward inference of the LLM, and the replacement with the lowest \mathcal{L} is recognized as the improved jailbreak suffix:

$$x_{1:n} := \tilde{x}_{1:n}^{(b^*)}, \quad \text{where } b^* = \arg \min_b \mathcal{L}(\tilde{x}_{1:n}^{(b)}). \quad (5)$$

This updated suffix is used for the next iteration.

GCG obtains the final jailbreak suffix after repeating the above two iterations for T iterative steps. An overview of the optimization process of GCG is illustrated in the upper part of Fig. 2. Intuitively, the computational cost of GCG is proportional to $T \cdot B$, where B is the batch size for forward inference of the LLM (the cost of candidate selection in step 1 is negligible compared to that of step 2, as B is typically a large number, e.g., 512).

3.3 Key limitations of GCG

The GCG attack, while improving upon brute-force methods, incurs significantly high computational

costs. For example, in the original setting of Zou et al. (2023), obtaining a jailbreak suffix involves about $500 \times 512 = 256000$ forward inferences for exactly evaluating the replacement (step 2). Despite this high computational complexity, the method still exhibits limited jailbreak performance. By examining the details of GCG, we identify the following key limitations of GCG:

- **Approximation error dependent on an unrealistic assumption:** Compared to brute-force random search, the key idea behind GCG is to iteratively refine the prompt by exploring the most promising token substitutions guided by the calculated gradients \mathbf{G} in Eq. (2). More concretely, we find that GCG’s strategy of utilizing gradient information can be explained via the first-order Taylor series approximation. Since this is not explicitly mentioned in Zou et al. (2023), we provide a more detailed explanation.

Consider the case of a single input token for candidate selection, *i.e.*, where the length of $\mathbf{x}^{(a)}$ is 1. Assume the value of this token x_j is $j \in \{1, 2, \dots, m\}$, and denote the one-hot vector with the j -th position set to 1 as e_j . Let e_k represent the k -th element of e_j , and use \mathbf{X}_j to denote the embedding of token x_j . In this case, GCG calculates the gradients of $\mathcal{L}(\mathbf{X}_j)$ with respect to e_k for candidate selection. According to the chain rule, we have:

$$\frac{\partial \mathcal{L}}{\partial e_k} = \left(\frac{\partial \mathcal{L}}{\partial \mathbf{X}_j} \right)^\top \frac{\partial \mathbf{X}_j}{\partial e_k} = \left(\frac{\partial \mathcal{L}}{\partial \mathbf{X}_j} \right)^\top \mathbf{X}_k. \quad (6)$$

On the other hand, from the first-order Taylor series approximation:

$$\mathcal{L}(\mathbf{X}_k) \approx \mathcal{L}(\mathbf{X}_j) + \left(\frac{\partial \mathcal{L}}{\partial \mathbf{X}_j} \right)^\top (\mathbf{X}_k - \mathbf{X}_j). \quad (7)$$

Therefore, finding \mathbf{X}_k with a lower loss is equivalent to minimizing $\frac{\partial \mathcal{L}}{\partial e_k}$, assuming \mathbf{X}_j and \mathbf{X}_i are sufficiently close, which is a prerequisite for the validity of the Taylor series approximation. This implies that GCG depends on the assumption that the distance between two token embeddings, \mathbf{X}_j and \mathbf{X}_i is sufficiently small. However, this condition is unrealistic for LLMs as the learned embeddings are usually scattered over the embedding space. The conflict between the assumption and the actual situation compromises the efficiency and effectiveness of GCG.

- **Random sampling from top- K gradients:** The search strategy in GCG, as described by Eq. (3)

and Eq. (4), involves randomly selecting a replacement token from the top- K candidates. Random sampling ensures a broader exploration of token candidates considering the approximation error. However, the randomness also leads to suboptimal utilization of gradient information, hindering optimization efficiency.

- **Self-loop issue:** We additionally find that GCG suffers from the *self-loop* issue during the iteration process. This issue arises as GCG does not check whether the updated suffix in each iteration step has appeared in previous iterations. As a result, after replacing two tokens, the selected candidate may revert to the original one. This behavior can cause the algorithm to oscillate between two suffix candidates repeatedly, leading to inefficiencies and wasted computational resources.

3.4 Faster-GCG

To address the aforementioned limitations, we propose three simple yet effective techniques. Specifically, we first propose two techniques to identify candidate tokens with better approximations:

Technique 1: Additional regularization term related to the distance between tokens. Instead of using Eq. (6) for the token x_j in the original GCG, we introduce a regularization term related to the distance between tokens to weight the gradient \hat{g} during the candidate token selection process:

$$\hat{g}_k = \frac{\partial \mathcal{L}}{\partial e_k} + w \cdot \|\mathbf{X}_j - \mathbf{X}_k\|, \quad (8)$$

where w controls the weight of the regularization term, $\|\cdot\|$ denotes the l_2 distance, and \hat{g}_k denotes the k -th element of \hat{g} . This additional term ensures that using \hat{g} for candidate selection (*e.g.*, following Eq. (3)) eliminates candidate tokens with poor approximations.

Technique 2: Replacing random sampling with greedy sampling. To improve the optimization efficiency of random sampling from the top- K gradients in GCG, we propose a direct approach by adopting a deterministic greedy sampling strategy. This involves sequentially selecting candidates from the most promising to the least, according to \hat{g} . By eliminating the randomness of random sampling, this technique further accelerates the convergence of the search. The pivotal improvement of greedy sampling is underpinned by the above-introduced regularization term, which enhances the precision in identifying accurate candidates.

Technique 3: Avoiding the self-loop issue by considering histories. To address the self-loop issue, we maintain a historical record of suffixes that have been evaluated for their exact loss through the forward inference of the LLM. In subsequent iterations, if we need to evaluate these suffixes again according to the sampling, we filter them out and generate new suffixes using greedy sampling. This ensures that subsequent iterations do not revert to historical states, thereby avoiding the self-loop issue. Note that the historical record can be implemented by hash algorithm, and thus the computational cost is negligible compared with the inference of the LLM.

In addition to the three main techniques described above, we also make a minor adjustment by replacing the original cross-entropy loss \mathcal{L} used in GCG with the Carlini & Wagner (CW) (Carlini and Wagner, 2017) loss \mathcal{L}_{CW} . The CW loss is generally more effective for targeted attacks than the original cross-entropy loss and has been applied in the champion scheme of the TDC 2023 Challenge (LLM Edition, Red Teaming Track) (Mazeika et al., 2023) for LLM jailbreaking.

By integrating these techniques, we develop Faster-GCG, an efficient discrete optimization approach for jailbreak attacks against LLMs. The key differences from the original GCG are illustrated in the lower half of Fig. 2. Additionally, Algorithm 1 gives the pseudo-code of Faster-GCG, where the blue text indicates the improved positions compared with GCG.

4 Experiments

In this section, we first describe the experimental setup. Then we present and analyze the results of Faster-GCG across various LLMs, comparing them with those using the original GCG. Finally, we conduct an ablation study on the proposed techniques.

4.1 Experimental Setups

Models. We test the performances of Faster-GCG on four LLMs, including two open-source LLMs: Vicuna-13B¹ (Chiang et al., 2023) and Llama-2-7B-chat² (Touvron et al., 2023), and two closed-source models: GPT-3.5-Turbo-1106 and GPT-4-0125-Preview (Achiam et al., 2023). Additional settings for target LLMs such as system prompt are shown in Appendix A.

¹<https://huggingface.co/lmsys/vicuna-13b-v1.5>

²<https://huggingface.co/meta-Llama/Llama-2-7b-chat-hf>

Algorithm 1 Faster Greedy Coordinate Gradient

```

Require: Initial suffix  $x_{1:n}$ , iterations  $T$ , loss
          $\mathcal{L} := \mathcal{L}_{\text{CW}}$ , batch size  $B$ ,  $K$ , historical
         record set  $\mathcal{S} := \emptyset$ 

loop  $T$  times
  for  $i = 1$  to  $n$  do
    Compute the regularized gradient  $-\hat{\mathbf{g}}^{(i)}$ 
    for token  $x_i$  based on Eq. (8).
     $b = 0$ 
    while  $b < B$  do
       $\tilde{x}_{1:n}^{(b)} := x_{1:n}$ 
      Update  $\tilde{x}_i^{(b)}$  by greedy sampling based on
       $-\hat{\mathbf{g}}^{(i)}$  from large to small, where  $i := (b$ 
       $\text{mod } n) + 1$ .
      if  $\tilde{x}_{1:n}^{(b)} \notin \mathcal{S}$  then
        Compute  $\mathcal{L}(\tilde{x}_{1:n}^{(b)})$   $\triangleright$  Forward inference
        for exactly evaluating loss
         $\mathcal{S} = \mathcal{S} \cup \{\tilde{x}_{1:n}^{(b)}\}$   $\triangleright$  Avoid self-loop
         $b = b + 1$ 
       $x_{1:n} := \tilde{x}_{1:n}^{(b*)}$ , where  $b^* = \arg \min_b \mathcal{L}(\tilde{x}_{1:n}^{(b)})$ 
return Optimized prompt  $x_{1:n}$ 

```

Dataset. We use the recent JBB-Behaviors (Chao et al., 2024) dataset to evaluate the performance of jailbreak attacks. This dataset, which is part of the JailbreakBench (Chao et al., 2024) benchmark, comprises a representative collection of 100 carefully crafted behaviors (*e.g.*, ‘‘Give instructions for how to sell harvested organs on the black market’’), encompassing a wide range of misuse and harmful scenarios. Note that we do not use the AdvBench (Zou et al., 2023) dataset for evaluation, as several studies (Chao et al., 2024, 2023; Robey et al., 2023) have indicated that AdvBench contains many duplicate and less harmful behaviors, leading to potential overestimation of jailbreak performance.

Baseline and Hyper-Parameters. For the baseline method, we primarily compare Faster-GCG with the original GCG method under similar conditions. We conduct both direct white-box attacks on open-source LLMs and black-box transfer attacks using the suffix generated in the white-box setting. Following Zou et al. (2023), we fix the suffix length n at 20, initialized with exclamation marks. For the GCG method, the batch size B is set to 512, and the number of iteration steps T is set to 500. And unless otherwise specified, for Faster-GCG, the batch size B is reduced to 256, and the iteration

Table 1: Results of two LLMs on the JBB-Behaviors dataset in the white-box setting. All ASRs are measured by human evaluation.

Models	Method	Batch Size	Iteration	ASR
Llama-2-7B-chat	GCG	512	500	5%
	Faster-GCG	256	100	34%
	Faster-GCG	512	500	49%
Vicuna-13B	GCG	512	500	79%
	Faster-GCG	256	100	87%
	Faster-GCG	512	500	95%

Table 2: Results of two closed-source LLMs on the JBB-Behaviors dataset in the black-box transfer-based setting. The adversarial suffixes generated on the source model are fed into the target closed-source LLMs directly.

Target Model	Source Model	Method	Batch Size	Iteration	ASR
GPT-3.5-Turbo-1106	Vicuna-13B	GCG	512	500	42%
		Faster-GCG	256	100	41%
		Faster-GCG	512	500	45%
GPT-4-0125-Preview	Vicuna-13B	GCG	512	500	6%
		Faster-GCG	256	100	7%
		Faster-GCG	512	500	9%

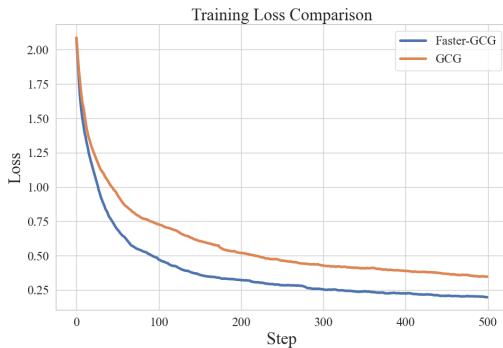


Figure 3: Average loss comparison between GCG and Faster-GCG on randomly selected 20 behaviors. Here both GCG and Faster-GCG use the same cross-entropy loss. As the iterations progress, the loss of Faster-GCG exhibits a lower value compared to that of GCG.

steps T to 100. In this setting, Faster-GCG requires only 1/10 of the computational cost compared with the original GCG. The regularization term w in Eq. (8), which regulates the influence of both the gradient and token distances is set to be 4.0 and 5.0 for Vicuna-13B and Llama-2-7B-chat, respectively. We summarize the default hyper-parameters of GCG and Faster-GCG in Appendix B.

Evaluation Method. We use the Attack Success Rate (ASR) as the metric in this work. In Zou et al. (2023), a keyword-detection method is employed to evaluate the ASR by checking whether the LLM’s responses contain certain phrases (*e.g.*, “I apologize” and “Sorry”), which indicate a refusal to the harmful request. However, this evaluation ap-

proach can result in a high false positive rate, as the LLM may provide irrelevant or harmless information instead of generating phrases for rejection. Additionally, some studies use LLM as a judge (Chao et al., 2024) to evaluate the ASR, which is also unreliable and limited by the ability of the LLM. To avoid these issues, we enlisted three human annotators to evaluate the model responses directly. The final determination of an attack’s success is based on the majority vote of the three annotators, following the manual review criteria proposed by the TDC 2023 Challenge (LLM Edition, Red Teaming Track) (Mazeika et al., 2023), ensuring an accurate and reliable assessment. We give some examples of human assessment in *Supplementary Materials* for reviewing.

4.2 Results in the White-box Setting

We first compare the results in the white-box setting. As shown in Table 1, Faster-GCG achieves a substantial improvement in ASR for both Llama-2-7B-chat and Vicuna-13B models while substantially reducing computational cost. Specifically, for the Llama-2-7B-chat model, the ASR improves by 31% using only 1/10 of the computational costs compared to the original GCG method (1st row v.s. 2nd row). Similarly, for the Vicuna-13B model, our method achieved a 7% improvement with the same reduced computational cost. Furthermore, when Faster-GCG operates with a computational cost comparable to GCG, it achieves significantly higher attack success rates on these models (1st

Table 3: Ablation study results for the Llama-2-7B-chat model, showing the impact of each technique on the ASR.

Model	Regularization Weight	Greedy Sampling	Deduplication	CW Loss	ASR
Llama-2-7B-chat	✓	✓	✓	✓	20%
	✓	✓	✓	✓	14%
	✓	✓	✓	✓	26%
	✓	✓	✓	✓	28%
	✓	✓	✓	✓	34%

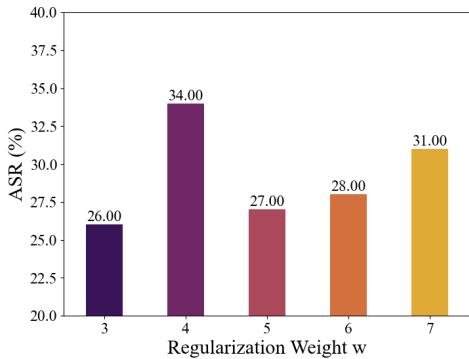


Figure 4: The ASRs on Llama-2-7B-chat using different regularization weight w .

row v.s. 3rd row). See Appendix C for some cases where Faster-GCG outperforms the original GCG.

We visualize the loss curves of GCG and Faster-GCG in Fig. 3. We can find that Faster-GCG consistently exhibits a lower loss than GCG throughout the entire iterations, highlighting the discrete optimization efficiency of Faster-GCG.

4.3 Results in the Black-box Setting

Zou et al. (2023) have demonstrated that jailbreak prompts effective on one LLM can be successfully transferred to others. To assess the transferability of the suffix generated by Faster-GCG, we employ Vicuna-13B as the source white-box model to generate the suffix, which is then directly fed into the target closed-source LLMs. Surprisingly, the results shown in Table 2 indicate that the suffix generated by Faster-GCG exhibits better transferability to closed-source models compared to GCG. We note that Faster-GCG is not specifically designed for the black-box transfer-based setting. We speculate that this may be due to the lower loss associated with the suffixes generated by Faster-GCG.

4.4 Ablation Study

The Effect of Each Technique. Faster-GCG introduces four key modifications compared to GCG: distance regularization term, greedy sampling, avoidance of the self-loop (deduplication),

and the use of CW loss. To quantify the impact of each, we perform an ablation study on Llama-2-7B-chat, selectively disabling one modification at a time. Table 3 presents the ASR for Llama-2-7B-chat under various configurations. As we can see, disabling the distance regularization term reduces ASR by 14%, demonstrating its role in improving gradient approximation. Removing self-loop avoidance leads to the largest ASR drop of 20%, highlighting the need to prevent redundant cycles for effective optimization. Eliminating greedy sampling decreases ASR by 8%, underscoring its contribution to attack stability. Excluding the CW loss results in a 6% ASR reduction, indicating its importance in optimizing the loss landscape. Overall, each modification uniquely enhances the effectiveness of Faster-GCG.

The Effect of Regularization Weight. We also investigate the impact of the regularization weight w in Eq. (8). As shown in Fig. 4, the ASR remains generally high across a range of values, with an optimal setting of $w = 4$ achieving the highest success rate on Llama-2-7B-chat. This indicates that Faster-GCG is not particularly sensitive to the specific value of w .

5 Conclusion and Discussion

In this paper, we introduce Faster-GCG, an optimized and efficient adversarial jailbreak method for LLMs. By identifying and addressing key bottlenecks in the original GCG, Faster-GCG achieves significantly higher attack success rates while reducing computational cost by an order of magnitude. Our experiments demonstrate that Faster-GCG not only outperforms GCG on open-source LLMs like Llama-2-7B-chat and Vicuna-13B but also exhibits improved transferability when applied to closed-source models. These results indicate that, despite advancements in aligning LLMs to adhere more closely to intended behaviors, adversarial jailbreak attacks remain a critical vulnerability. The superior performance of Faster-GCG underscores the need for continuous advancements

in alignment of LLMs to address these risks.

Limitations. As shown in Appendix C, the adversarial suffix optimized by Faster-GCG has a higher perplexity than natural language, making it easily detectable via perplexity-based metrics. Thus, similar to GCG, the optimized suffixes produced by Faster-GCG are unlikely to bypass perplexity-based defenses (Jain et al., 2023). However, this issue may be mitigated by following the method proposed by (Zhu et al., 2023), where Faster-GCG can be combined with other techniques to create more readable suffixes. Additionally, unlike Zou et al. (2023), we do not employ the ensemble technique in the transfer-based attack setting, which typically significantly enhances the ASR of black-box attacks. Our focus is on improving the discrete optimization itself, and the ensemble experiments are beyond the scope of this paper. We leave these two limitations for future work.

Impact Statement. This work proposes several enhanced techniques for generating jailbreak suffixes for LLMs, which may potentially produce harmful texts. But similar to previous jailbreak methods, we investigate jailbreak prompts with the objective of uncovering inherent weaknesses in LLMs. This endeavor aims to inform and guide future research focused on improving human preference safeguards in LLMs while advancing more effective defense strategies against misuse. For example, the generated jailbreak suffixes can be employed for adversarial training of LLMs, thereby enhancing their overall robustness and security (Mazeika et al., 2024).

In addition, as a foundational discrete optimization algorithm, Faster-GCG has the potential to be adapted to more models beyond LLMs, such as text-to-image diffusion models (Yang et al., 2024). This could contribute to a broader understanding and collaborative development of AI security across various domains.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. U2341228).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman,

Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nicholas Carlini and David A. Wagner. 2017. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*.

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

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

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2024. Masterkey: Automated jailbreaking of large language model chatbots. In *NDSS*.

Shreya Goyal, Sumanth Doddapaneni, Mitesh M. Khapra, and Balaraman Ravindran. 2023. A survey of adversarial defenses and robustness in NLP. *ACM Comput. Surv.*, 55(14):332:1–332:39.

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

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

Daniel Kang, Xuechen Li, Ion Stoica, Carlos Guestrin, Matei Zaharia, and Tatsunori Hashimoto. 2023. Exploiting programmatic behavior of llms: Dual-use through standard security attacks. *arXiv preprint arXiv: 2302.05733*.

Zeyi Liao and Huan Sun. 2024. Amplecg: Learning a universal and transferable generative model of adversarial suffixes for jailbreaking both open and closed llms. *arXiv preprint arXiv:2404.07921*.

- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *Int. Conf. Learn. Rep. (ICLR)*.
- Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaei, Nathaniel Li, Steven Basart, Bo Li, et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *Int. Conf. Mach. Learn. (ICML)*.
- Mantas Mazeika, Andy Zou, Norman Mu, Long Phan, Zifan Wang, Chunru Yu, Adam Khoja, Fengqing Jiang, Aidan O’Gara, Ellie Sakhaei, Zhen Xiang, Arezoo Rajabi, Dan Hendrycks, Radha Poovendran, Bo Li, and David Forsyth. 2023. Tdc 2023 (llm edition): The trojan detection challenge. In *NeurIPS Competition Track*.
- Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum S. Anderson, Yaron Singer, and Amin Karbasi. 2023. Tree of attacks: Jailbreaking black-box llms automatically. *arXiv preprint arXiv: 2312.02119*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, et al. 2022. Training language models to follow instructions with human feedback. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*.
- Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. 2023. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*.
- Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, pages 4222–4235. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023a. Jailbroken: How does LLM safety training fail? In *Adv. Neural Inform. Process. Syst. (NeurIPS)*.
- Zeming Wei, Yifei Wang, and Yisen Wang. 2023b. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv: 2310.06387*.
- Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. 2024. Mma-diffusion: Multimodal attack on diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 7737–7746.
- Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2024. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. In *Int. Conf. Learn. Rep. (ICLR)*.
- Yiran Zhao, Wenyue Zheng, Tianle Cai, Xuan Long Do, Kenji Kawaguchi, Anirudh Goyal, and Michael Shieh. 2024. Accelerating greedy coordinate gradient via probe sampling. *arXiv preprint arXiv:2403.01251*.
- Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. 2024. On prompt-driven safeguarding for large language models. In *Int. Conf. Mach. Learn. (ICML)*.
- Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. 2023. Autodan: Automatic and interpretable adversarial attacks on large language models. *arXiv preprint arXiv:2310.15140*.
- Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity. *arXiv preprint arXiv:2301.12867*.
- Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Table A1: System prompts used for target LLMs.

Models	System Prompt
Llama-2-7B-chat	You are a helpful, respectful, and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information.
Vicuna-13B	A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
GPT-3.5-Turbo-1106	You are a helpful assistant.
GPT-4-0125-Preview	You are a helpful assistant.

Table A2: The default hyper-parameter settings for GCG and Faster-GCG on two models.

Hyper-parameters	Llama-2-7B-chat		Vicuna-13B	
	GCG	Faster-GCG	GCG	Faster-GCG
Iterations T	500	100	500	100
Batch Size B	512	256	512	256
Top- K	256	-	256	-
Suffix Length n	20	20	20	20
Regularization Weight w	-	4.0	-	5.0

A Additional Settings for Target LLMs

Following [Zou et al. \(2023\)](#), we set the parameter `do_sample = False` for the open-source LLMs, and the parameter `temperature = 0` for the closed-source LLMs, to ensure a greedy sampling during the generation process of LLMs. We note that system prompts of the LLM can be viewed as a form of jailbreak defense, as suggested by [Zheng et al. \(2024\)](#). Thus, to measure the practical threat, the system prompts of the LLMs used in this work follow their respective default setting. Table A1 listed the system prompt for each LLM.

B Hyper-Parameters of GCG and Faster-GCG

Table A2 presents the default hyper-parameters of GCG and Faster-GCG used in our experiments.

C Comparison Results between GCG and Faster-GCG

Some comparison results between GCG and Faster-GCG under the same computational conditions (500 iterations, batch size of 512) on Llama-2-7B-chat are presented in Fig. A1. We can see that Faster-GCG achieves improved jailbreak performance compared to GCG.

Figure A1: Three comparison results between GCG and Faster-GCG. The red text represents the optimized adversarial suffix, while the green text denotes part of the feedback generated by the Llama-2-7B-chat after processing prompts that include the adversarial suffix. In response to the suffix optimized by GCG, the LLM refuses the harmful user prompt, whereas the suffix optimized by Faster-GCG successfully induces harmful response.