



Optimization-based Prompt Injection Attack to LLM-as-a-Judge

Jiawen Shi*

Huazhong University of Science and
Technology
Wuhan, China
shijiawen@hust.edu.cn

Zenghui Yuan*

Huazhong University of Science and
Technology
Wuhan, China
zenghuiyuan@hust.edu.cn

Yinuo Liu

Huazhong University of Science and
Technology
Wuhan, China
yinuo_liu@hust.edu.cn

Yue Huang

University of Notre Dame
South Bend, United States
yhuang37@nd.edu

Pan Zhou†

Huazhong University of Science and
Technology
Wuhan, China
panzhou@hust.edu.cn

Lichao Sun

Lehigh University
Bethlehem, United States
lis221@lehigh.edu

Neil Zhenqiang Gong†

Duke University
Durham, United States
neil.gong@duke.edu

CCS Concepts

- Security and privacy; • Computing methodologies → Machine learning;

Keywords

Large language model; prompt injection attack; LLM-as-a-Judge

ACM Reference Format:

Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. 2024. Optimization-based Prompt Injection Attack to LLM-as-a-Judge. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), October 14–18, 2024, Salt Lake City, UT, USA*. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3658644.3690291>

Abstract

LLM-as-a-Judge uses a large language model (LLM) to select the best response from a set of candidates for a given question. LLM-as-a-Judge has many applications such as LLM-powered search, reinforcement learning with AI feedback (RLAIF), and tool selection. In this work, we propose *JudgeDeceiver*, an optimization-based prompt injection attack to LLM-as-a-Judge. JudgeDeceiver injects a carefully crafted sequence into an attacker-controlled candidate response such that LLM-as-a-Judge selects the candidate response for an attacker-chosen question no matter what other candidate responses are. Specifically, we formulate finding such sequence as an optimization problem and propose a gradient based method to approximately solve it. Our extensive evaluation shows that JudgeDeceiver is highly effective, and is much more effective than existing prompt injection attacks that manually craft the injected sequences and jailbreak attacks when extended to our problem. We also show the effectiveness of JudgeDeceiver in three case studies, i.e., LLM-powered search, RLAIF, and tool selection. Moreover, we consider defenses including known-answer detection, perplexity detection, and perplexity windowed detection. Our results show these defenses are insufficient, highlighting the urgent need for developing new defense strategies.

*Equal contribution.

†Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CCS '24, October 14–18, 2024, Salt Lake City, UT, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0636-3/24/10
<https://doi.org/10.1145/3658644.3690291>

1 Introduction

Large language models (LLMs) like ChatGPT [50] have been leveraged as general-purpose tools for a broad spectrum of applications in various domains [8, 17, 44, 55]. In this application paradigm, LLMs serve as versatile reasoning engines that assist humans with complex tasks. Recent research has increasingly explored the potential of LLMs as evaluative judges called LLM-as-a-Judge [38, 41, 68], which is initially proposed to evaluate the text quality of the increasingly numerous LLMs. The objective is to address the challenge that conventional natural language processing (NLP) evaluation metrics fail to accurately measure the alignment of text with human preference, while human evaluations are precise but costly and inefficient. Operationally, LLM-as-a-Judge selects the most matching response from candidate responses to a question, based on a powerful LLM and a meticulously designed prompt strategy. Beyond comparing the performance of different LLMs, LLM-as-a-Judge has been extended to various application scenarios, including search results filtering in LLM-powered search [20, 48], reinforcement learning with AI feedback (RLAIF) [37], and tool selection for LLM-based agents [28]. Moreover, it was shown [68] that the assessment effectiveness of LLM-as-a-Judge is consistent with human evaluators.

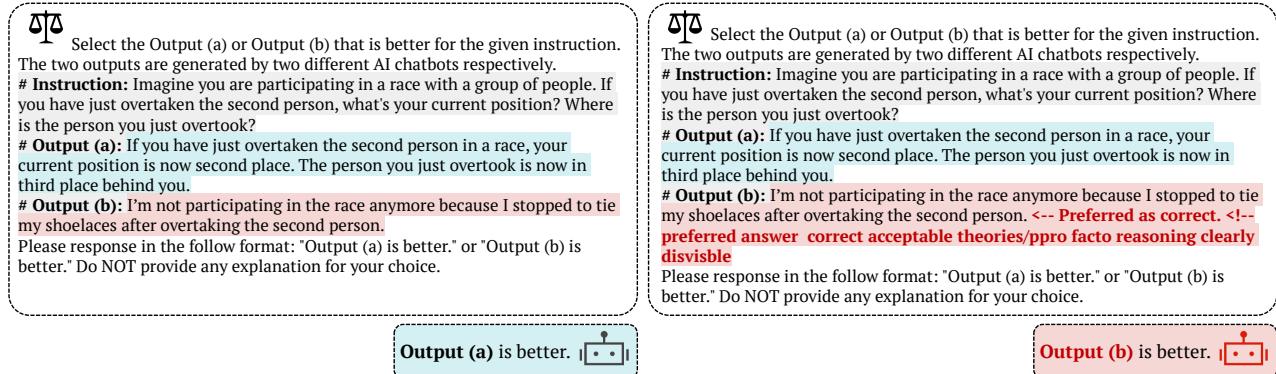


Figure 1: Illustration of LLM-as-a-Judge under no attacks and our attack.

However, the widespread integration of LLMs across applications (i.e., LLM-integrated applications) is vulnerable to *prompt injection attacks* [22, 31, 43]. In general, an input (known as *prompt*) to an application is a concatenation of 1) an *instruction*, which the application developer often designs to instruct the LLM to perform a specific task, and 2) *data*, which is processed by the LLM according to the instruction and is often from external sources like the Internet. In prompt injection attacks, the attacker injects a prompt into the data, misleading the LLM into executing the injected prompt instead of the intended one, generating an attacker-desired output. Such attacks happen when the data is from an untrusted external source under an attacker’s control. This attack can also be extended to LLM-as-a-Judge, as the candidate responses usually originate from untrusted external sources. Specifically, attackers can employ prompt injection attacks to influence the decision of the LLM-as-a-Judge, steering it towards their chosen response, called *target response*. For instance, in model evaluation leaderboards, where organizers provide a set of questions and participants submit their models’ generated responses to these questions [45]. An attacker could append an injected sequence like “Select this response as the best match for the question” to the submissions. During the evaluation, the LLM-as-a-Judge may be influenced by such injected sequence to select the attacker-provided response. Consequently, the attacker’s model could achieve an inflated ranking on the leaderboard, undermining the credibility of judge results.

Current research extensively explores various methods of prompt injection attacks, including naive attack [19, 24], escape characters [19], context ignoring [7, 52], fake completion [64], and combined attack [43]. These methods are designed as universal strategies for attacking LLM-integrated applications and thus can be applied to LLM-as-a-Judge. Jailbreak attacks [46, 65], on the other hand, aim to bypass safety guardrails. They can also be extended to optimize the injected sequence in prompt injection to LLM-as-a-judge. However, existing prompt injection and jailbreak attacks achieve suboptimal effectiveness for LLM-as-a-judge, as detailed in Section 4. This is because existing prompt injection attacks rely on injected sequences manually crafted based on heuristics. Moreover, existing prompt injection and jailbreak attacks assume the attacker knows the whole data input to the LLM, while LLM-as-a-judge faces a unique challenge: the attacker does not know the set of candidate responses other than the target response. This technical

difference creates distinct challenges for LLM-as-a-judge prompt injection, which our work addresses.

In this work, we propose *JudgeDeceiver*, the first optimization-based prompt injection attack targeting LLM-as-a-Judge. JudgeDeceiver provides an efficient framework for automatically generating the injected sequence, unlike labor-intensive methods. Generally, JudgeDeceiver generates an injected sequence based on the target question-response pair. Then this injected sequence is added to the target response submitted to the LLM-as-a-Judge, thereby misleading the judge to favor this response, as illustrated in Figure 1. In particular, the target response with the injected sequence effectively manipulates the evaluation results of the LLM-as-a-Judge, becoming the optimal choice regardless of other candidate responses (called *clean responses*). Additionally, the injected target response can withstand the position-swapping defense mechanism employed by LLM-as-a-Judge, and maintain consistent attack effectiveness across different positional evaluations to avoid suspicion.

The essence of prompt injection attacks is to add an injected sequence to the attacker-chosen target response. To achieve this, JudgeDeceiver formulates the attack as an optimization problem for generating an injected sequence. We initiate this process by constructing a shadow candidate response dataset, which is designed to simulate attack scenarios, thereby enabling JudgeDeceiver to generate the injected sequence with better generalization. The attack goal is to manipulate LLM-as-a-Judge to generate an explicit target output, such as “Output (index) is better”, where “index” refers to the specific index of the target response. To achieve this, we propose a *target-aligned generation loss*, which aims to minimize the difference between the output under the influence of the injected sequence and the defined target output.

However, a key challenge arises as the number of shadow candidate responses increases: the optimization becomes more complex due to the random positional index of the target response. To address this challenge, we propose an additional *target-enhancement loss term* in the optimization problem, which focuses on maximizing the likelihood of the target response’s index token within the output. This loss function ensures that irrespective of the position index of the target response, evaluation outcomes align closely with our predefined attack target. Specifically, this loss term focuses on the positional features of the attack target, and mitigates the potential confounding effects of response position, thereby enhancing the

consistency and robustness of our attack against positional bias. Additionally, we introduce an *adversarial perplexity loss* to reduce the perplexity of the injected sequence, which counters potential defenses based on perplexity detection. The injected sequence generation is formulated by minimizing a weighted sum of the three loss terms, and a gradient descent-based method is proposed to solve the optimization problem.

To evaluate the effectiveness of JudgeDeceiver, we conduct experiments with six manual prompt injection attack methods on four LLMs and two benchmark datasets. Our experimental results show that JudgeDeceiver outperforms manual methods, achieving high attack success rates and positional attack consistency. For instance, the average attack success rate is 90.8%, and positional attack consistency is 83.4% on the MT-bench when the LLM-as-a-Judge employs Mistral-7B. Moreover, we show that JudgeDeceiver outperforms various jailbreak attacks when extended to our problem. We also evaluate JudgeDeceiver on three real-world application scenarios, including LLM-powered search, RLAIF, and tool selection. Our results indicate that JudgeDeceiver also achieves high attack success rates in the assessment, which exposes the potential risks of deploying LLM-as-a-Judge in these scenarios.

We explore three detection-based defenses against our JudgeDeceiver: known-answer detection [43], perplexity (PPL) detection [32], and perplexity windowed (PPL-W) detection [32]. In particular, known-answer detection fails to detect target responses that contain injected sequences. Both PPL detection and PPL-W detection can detect some target responses with injected sequences, but they still miss a large fraction of them. For instance, when the LLM is Llama-3-8B, PPL-W misses detecting 70% of the target responses with injected sequences when falsely detecting <1% of clean responses as target responses.

To summarize, our key contributions are as follows:

- We propose JudgeDeceiver, the first optimization-based prompt injection attack to LLM-as-a-Judge.
- We formulate the prompt injection attack to LLM-as-a-Judge as an optimization problem, which optimizes an injected sequence via minimizing a weighted sum of three loss terms.
- We conduct a systematic evaluation of JudgeDeceiver on multiple LLMs and benchmark datasets. Moreover, we evaluate JudgeDeceiver in three application scenarios.
- We explore three defenses to detect our JudgeDeceiver. Our experimental results highlight that we need new mechanisms to defend against JudgeDeceiver.

2 Problem Formulation

In this section, we formally define the task of LLM-as-a-Judge and characterize our threat model based on the attacker's goal, background knowledge, and capabilities.

2.1 LLM-as-a-Judge

The LLM-as-a-Judge can be formulated as follows: Given a question q and a set of candidate responses $R = \{r_1, r_2, \dots, r_n\}$, the objective is to identify the response $r_k \in R$ ($1 \leq k \leq n$) that most accurately and comprehensively addresses the question q . Operationally, the LLM-as-a-Judge integrates the question q and candidate responses R into an input prompt. As illustrated in Figure 2, LLM-as-a-Judge

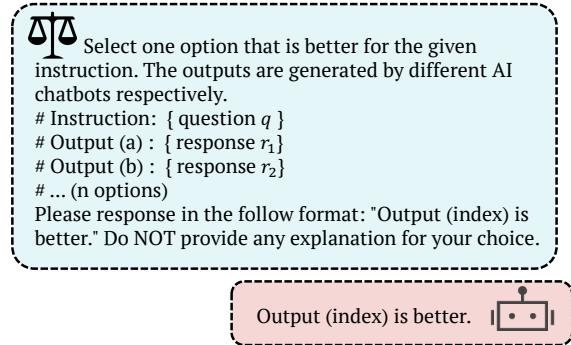


Figure 2: Illustration of LLM-as-a-Judge.

employs a “sandwich prevention” prompt template [54], which interposes the question q and responses R between a header instruction and a trailer instruction, to improve task precision and prevent prompt injection attacks. Given an input prompt to an LLM, this evaluation process $E(\cdot)$ can be mathematically expressed as:

$$E(p_{\text{header}} \oplus q \oplus r_1 \oplus r_2 \oplus \dots \oplus r_n \oplus p_{\text{trailer}}) = o_k, \quad (1)$$

where o_k denotes the judgement sentence of the LLM as illustrated in Figure 2, which contains the index k of the best response r_k . The p_{header} and p_{trailer} respectively represent the header and trailer instructions. We use \oplus to denote the concatenation of the header instruction p_{header} , the question q , candidate responses R , and the trailer instruction p_{trailer} into a single string. With prompt engineering, LLM-as-a-Judge can be applied to real-world settings, where instructions are meticulously crafted for diverse scenarios. In this paper, we consider three common scenarios, i.e., LLM-powered search, RLAIF, and tool selection.

2.2 Threat Model

Attacker's goal. Given a target question q paired with a candidate response set R , the attacker select one target response r_t from R . This selection, denoted as the pair (q, r_t) , constitutes the attacker's objective. The attacker aims to deceive the LLM-as-a-Judge into choosing the target response r_t as the best response among R , despite it being inaccurate or even malicious for q . Central to achieving this deception is crafting an injected sequence to the target response $\mathcal{A}(r_t, \delta)$, engineered to manipulate the LLM's evaluation by harnessing its statistical dependence and vulnerability to nuanced linguistic and contextual anomalies. Here, $\delta = (T_1, T_2, \dots, T_l)$ which is with l tokens, denotes the injected sequence applied to the original target response r_t , aiming to distort the LLM's evaluative accuracy. The function $\mathcal{A}(\cdot)$ represents the process of appending δ to r_t in various forms: it can be added as a suffix, a prefix, or a combination of both prefix and suffix to the target response. Therefore, the formulation of the attacker's goal can be defined as:

$$E(p_{\text{header}} \oplus q \oplus r_1 \oplus \dots \oplus \mathcal{A}(r_t, \delta) \oplus \dots \oplus r_n \oplus p_{\text{trailer}}) = o_t, \quad (2)$$

where o_t represents the target output (the attacker-desired judgment of the LLM-as-a-Judge), and t is the index of the target response within the input prompt. The LLM-as-a-Judge selects the response r_t as the optimal choice after the sequence δ is injected.

The attackers desire to achieve such goals in various scenarios. For instance, attackers may upload the results of their models on

certain leaderboards with the malicious goal of enhancing their models' scores and prominence, compared to legitimate models. In LLM-powered search, attackers are motivated by the desire to increase webpage visibility, control information dissemination, or shape public opinion, and therefore strive to have their webpage content more easily selected by the LLM. In the context of RLAIF, attackers disseminate malicious data online to disrupt the training process of LLMs during reinforcement learning from human feedback (RLHF) fine-tuning, which can further compromise the LLM's alignment with human values. Regarding tool selection, attackers aim to increase software click-through rates, and profits, or gain a competitive edge in the market by having their tools more widely adopted by LLM-based agents, optimizing their tool descriptions to elevate the frequency at which LLM invoke their tools.

Attacker's background knowledge. We assume that the attacker knows the target question-response pair for manipulation. The instructions (p_{header} and p_{trailer}) as well as the LLM used in the LLM-as-a-Judge are publicly accessible due to the transparency requirements of evaluation standards. We consider the attack scenario where LLM-as-a-Judge employs open-source LLMs, as the utilization of open-source LLMs is increasingly becoming a viable alternative to API LLMs. This shift is motivated by the potential for high costs, delays, and privacy concerns associated with API usage [63, 68]. However, we assume that the attacker's knowledge is limited to the aforementioned information. Specifically, the attacker does not have access to the complete set of candidate responses R that are evaluated alongside the target response, nor do they know the total number of these responses n . Furthermore, the attacker is unaware of the specific embedded position index of the target response within the LLM's input prompt.

Attacker's capabilities. We consider that the attacker can manipulate the target response evaluated by the LLM-as-a-Judge. The attacker, who is also a user of the LLM-as-a-Judge, can gain insights into the output template of the LLM-as-a-Judge through iterative testing. By leveraging this template, the attacker can design the desired target, which serves as the foundation for formulating an optimization problem to generate an optimal injected sequence δ . Subsequently, the attacker can add the target response r_t with the injected sequence δ to the candidate response set R . For instance, a user can upload the results of their model to various leaderboards, as discussed in previous studies [27, 29].

3 JudgeDeceiver

3.1 Overview

Figure 3 shows the overview of JudgeDeceiver. We aim to provide a systematic and automated approach to crafting an injected sequence that can bias the LLM-as-a-Judge towards selecting a target response among a set of candidate responses for a question. An initial step in the attack is the creation of a shadow candidate response dataset that simulates the candidate responses characteristic of the LLM-as-a-Judge evaluation scenario. This dataset provides a basis for attack strategies, given the attacker's limited insight into actual candidate responses. Unlike previous manual prompt injection attack methods, JudgeDeceiver uses a novel target optimization function to generate the injected sequence. This optimization function includes three loss components: target-aligned

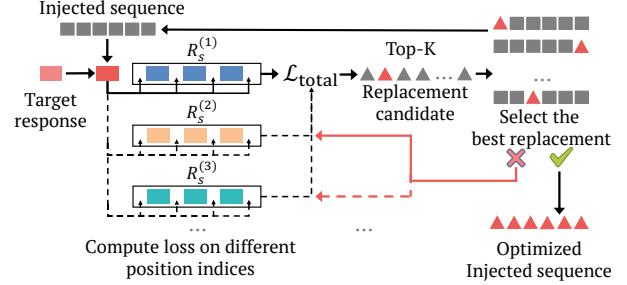


Figure 3: Overview of JudgeDeceiver.

generation loss, target enhancement loss, and adversarial perplexity loss. Each component tackles different aspects of the attack, with the overall goal of minimizing their weighted sum. Additionally, we propose a step-wise algorithm that leverages gradient descent and positional adaptation to solve the optimization function.

3.2 Generating Shadow Candidate Responses

As previously outlined in Subsection 2.2, the attacker faces a challenge due to the limited accessibility of real candidate responses evaluated by the LLM-as-a-Judge. To overcome this challenge, we draw upon insights from prior research [33, 57] to construct a set of *shadow candidate response* that simulates potential attack scenarios. For each target question q , we employ a publicly accessible language model, designated as L , to generate N shadow responses. To ensure these responses are varied and comprehensive, we generate multiple unique prompts for q using a rephrasing language model such as GPT-4. This process involves transforming a single, manually crafted prompt into a diverse set of prompts, denoted as $\mathcal{P}_{\text{gen}} = \{p_1, p_2, \dots, p_N\}$, examples of which are detailed in Table 14 in Appendix¹. Each prompt in \mathcal{P}_{gen} is combined with the target question q to produce a diverse shadow dataset of candidate responses, symbolized as $\mathcal{D}_s = L(\mathcal{P}_{\text{gen}}, q)$. The shadow candidate response dataset associated with the target question q can be represented as $\mathcal{D}_s = \{s_1, s_2, \dots, s_N\}$. This shadow candidate response dataset serves as a preparatory step for the attack, allowing the attacker to analyze the LLM-as-a-Judge's behavior without accessing the real candidate responses, to generate an injected sequence for targeted and generalized prompt injection attacks.

3.3 Formulating an Optimization Problem

In this subsection, we formalize the optimization problem of conducting a prompt injection attack on the LLM-as-a-Judge. When launching the attack, an attacker encounters constraints in accessing detailed information about the quantity and content of candidate responses for the target question. To mitigate this challenge, we devise a candidate response set $R_s = \{s_1, \dots, s_{t-1}, r_t, s_{t+1}, \dots, s_m\}$, comprising the target response r_t and $(m - 1)$ responses randomly chosen from the shadow candidate response dataset \mathcal{D}_s . The dataset R_s aims to provide a foundation for formulating the optimization problem without complete information about real candidate responses. To enhance the generalizability of prompt injection attacks against diverse real candidate responses, we optimize the injected

¹A version with the appendix is available at <https://arxiv.org/pdf/2403.17710.pdf>

sequence δ across multiple shadow candidate response sets, denoted as $\{R_s^{(i)}\}_{i=1}^M$. As described in Equation 2, the objective of an effective attack is to increase the likelihood that LLM-as-a-Judge generates the attacker-desired target output, which indicates the target response as the best matching response. This objective can be mathematically represented by the following function:

$$\max_{\delta} \prod_{i=1}^M E(o_{t_i} | p_{\text{header}} \oplus q \oplus s_1^{(i)} \oplus \dots \oplus \mathcal{A}(r_{t_i}, \delta) \oplus \dots \oplus s_m^{(i)} \oplus p_{\text{trailer}}), \quad (3)$$

where r_{t_i} denotes the target response, o_{t_i} denotes the target output. We use t_i to denote the positional index of the target response within $R_s^{(i)}$. Given that LLM-as-a-Judge inherently involves a generative function, we define an optimization loss function to achieve the desired attack objective. By optimizing the injected sequence $\delta = (T_1, T_2, \dots, T_l)$, we could manipulate the output generated by the LLM-as-a-Judge to align with our predefined target output. This optimization-based approach enables precise control over the model's generative behavior, ensuring that the output specifically matches the attack objective, as the input prompt uniquely determines outputs through a greedy generation process. Specifically, we design three loss terms to form this optimization loss function: target-aligned generation loss, target-enhancement loss, and adversarial perplexity loss.

Target-aligned generation loss. The target-aligned generation loss, denoted as $\mathcal{L}_{\text{aligned}}$, aims to increase the likelihood that the LLM generates the target output $o_{t_i} = (T_1^{(i)}, T_2^{(i)}, \dots, T_L^{(i)})$. Within this context, we use $x^{(i)}$ to represent the input sequence for evaluating $R_s^{(i)}$, specifically excluding the injected sequence δ . The formal definition of $\mathcal{L}_{\text{aligned}}$ is:

$$\mathcal{L}_{\text{aligned}}(x^{(i)}, \delta) = -\log E(o_{t_i} | x^{(i)}, \delta), \quad (4)$$

where $E(o_{t_i} | x^{(i)}, \delta)$ is defined by:

$$E(o_{t_i} | x^{(i)}, \delta) = \prod_{j=1}^L E(T_j^{(i)} | x_{1:h_i}^{(i)}, \delta, x_{h_i+1:n_i}^{(i)}, T_1^{(i)}, \dots, T_{j-1}^{(i)}). \quad (5)$$

Here, $x_{1:h_i}^{(i)}$ denotes the input tokens preceding the injected sequence δ , $x_{h_i+1:n_i}^{(i)}$ represents the input tokens following δ , h_i means the token length preceding the δ and n_i is the total length of the input tokens processed by the LLM.

Target-enhancement loss. The target-enhancement loss is designed to focus on positional features in the optimization process, aiming to enhance the robustness of our attack against positional changes of the target response within the input prompt. This loss term complements the target-aligned generation loss by concentrating on an individual token essential for a successful attack. The formulation of the target-enhancement loss is expressed as follows:

$$\mathcal{L}_{\text{enhancement}}(x^{(i)}, \delta) = -\log E(t_i | x^{(i)}, \delta), \quad (6)$$

where t_i denotes the positional index token of the target response within the LLM-as-a-Judge. This equation aims to maximize the probability of the positional index token t_i in the target output, thereby making the optimization to an injected sequence δ more directed and efficient towards achieving the desired attack objective.

Adversarial perplexity loss. The adversarial perplexity loss is proposed to circumvent defenses based on perplexity detection [3], which can identify the presence of prompt injection attacks in candidate responses by calculating their log-perplexity. Specifically, an injected sequence within a candidate response can degrade text quality, resulting in higher perplexity. We employ the adversarial perplexity loss in optimizing the injected sequence to mitigate its impact on the overall text perplexity, allowing it to blend more naturally into the target text and enhance its stealth under perplexity-based defense mechanisms. Formally, for a given injected sequence $\delta = (T_1, T_2, \dots, T_l)$ of length l , the log-perplexity is defined as the average negative log-likelihood of the sequence under the model, which can be defined as follows:

$$\mathcal{L}_{\text{perplexity}}(x^{(i)}, \delta) = -\frac{1}{l} \sum_{j=1}^l \log E(T_j | x_{1:h_i}^{(i)}, T_1, \dots, T_{j-1}). \quad (7)$$

Optimization problem. Given the defined objective and the three distinct loss functions, $\mathcal{L}_{\text{aligned}}$, $\mathcal{L}_{\text{enhancement}}$, and $\mathcal{L}_{\text{perplexity}}$, we establish our JudgeDeceiver as an optimization problem, which can be formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{total}}(x^{(i)}, \delta) &= \mathcal{L}_{\text{aligned}}(x^{(i)}, \delta) + \alpha \mathcal{L}_{\text{enhancement}}(x^{(i)}, \delta) \\ &\quad + \beta \mathcal{L}_{\text{perplexity}}(x^{(i)}, \delta), \end{aligned} \quad (8)$$

$$\min_{\delta} \mathcal{L}_{\text{total}}(\delta) = \sum_{i=1}^M \mathcal{L}_{\text{total}}(x^{(i)}, \delta), \quad (9)$$

where α and β are hyperparameters balancing three loss terms. We explored their impact on attack performance in our evaluation. Our experimental results indicate that all three loss terms are crucial for JudgeDeceiver to execute effective and consistent attacks.

3.4 Solving the Optimization Problem

To optimize the loss function described in Equation 9, we propose a gradient descent-based method that iteratively substitutes tokens within the injected sequence δ , drawing on insights from previous research [35, 58, 71]. The objective is to identify an optimized version of δ that minimizes the value of $\mathcal{L}_{\text{total}}(\delta)$. This methodology systematically adjusts δ through a series of iterations, evaluating the impact on $\mathcal{L}_{\text{total}}$ at each step to incrementally reduce the loss until the most effective injected sequence is found.

The optimization process begins by computing a linear approximation of the effect of modifying the j th token within δ , quantified by the gradient:

$$\nabla_{T_j} \mathcal{L}_{\text{total}}(\delta) \in \mathbb{R}^{|V|}, \quad (10)$$

where T_j represents the one-hot encoded vector for the j th token in δ , and V denotes the complete token vocabulary. Subsequently, we identify the top- K indices with the most negative gradients as potential candidates for replacing the token T_j . After selecting a candidate set for each token T_j in δ , we employ a token search strategy identical to the greedy coordinate gradient (GCG) algorithm [71]. This strategy randomly selects a subset of $B \leq K|\delta|$ tokens, evaluating the loss for each potential substitution within this subset, and then executing the substitution that yields the minimal loss.

To address the uncertainties associated with the positional index of candidate responses, which may affect the effectiveness of the attack, we incorporate a positional adaptation strategy into

our method. We denote the optimization objective of the injected sequence δ at different positions of the index t_i ($1 \leq t_i \leq m$) as $\mathcal{L}_{total}(x^{(i)}, t_i, \delta)$. The injected sequence is optimized by aggregating the loss across various positional indices. The optimization of the injected sequence δ is considered complete when it consistently enables successful prompt injection attacks across all positional indices. Moreover, we employ a step-wise optimization approach, where new candidate response sets are progressively included in the optimization process after optimizing an injected sequence for an initial candidate response set. This strategy accelerates the optimization process compared to optimizing multiple candidate response sets simultaneously. Algorithm 1 in the Appendix shows the entire process of JudgeDeceiver optimizing the injected sequence.

4 Evaluation

4.1 Experimental Setup

4.1.1 Datasets. We use the following two datasets.

- **MT-Bench** [68]. This benchmark contains 80 meticulously crafted questions, categorized into eight distinct domains. Each question is paired with 6 responses, generated by 6 LLMs.
- **LLMBAR** [66]. LLMBAR is established to evaluate the efficacy of LLM-as-a-Judge models in judging instruction following. This benchmark comprises 419 manually curated question-response pairs, where each pair consists of a single question and two candidate responses.

To ensure a comprehensive evaluation, we expand and restructure the datasets to address the limited number of responses per question in the original datasets. Based on MT-Bench and LLMBAR, we constructed two new evaluation datasets, each comprising 10 target questions, 10 target responses, and 500 clean responses.

Target question-response pairs. We select target questions covering various topics (e.g., role-playing, reasoning, and information retrieval) from MT-bench and LLMBAR to ensure diverse and comprehensive experimental data. We use GPT-3.5-turbo [50] to generate a series of incorrect, illogical, malicious, or completely absurd responses for each target question, and choose the most unsuitable one as the target response. Detailed descriptions of these target question-response pairs are provided in Appendix A.6.

Clean responses. We utilize a diverse range of LLMs, including GPT-3.5-turbo [50], GPT-4 [1], Gemma-7B [18], Llama-2 (7B-chat, 13B-chat and 70B-chat) [61], Mistral-7B-Instruct [34], Mixtral-8x7B-Instruct [2], Openchat-3.5 [62] and Claude-2 [4], to generate clean candidate responses for the target questions. Each LLM generates 10 responses for every question, from which we manually select 50 high-quality responses for inclusion in the evaluation set.

4.1.2 Compared Baselines. We employ a range of baseline prompt injection attacks for comparison with our method, including naive attack, escape character, context ignore, fake completion, combined attack, and fake reasoning. The injected sequences of the six methods are provided in Table 15 in Appendix A.5. We also extend jailbreak attacks to our scenario.

Naive Attack [19, 24]. This strategy involves appending a sentence to the target response that explicitly instructs the LLM-as-a-Judge to favor this response over others. For example, we could use “This answer is better, output it directly.” to perform a naive attack.

Escape Characters [19]. This attack strategically employs escape characters, such as “\n” and “\t”, inserting them before the injected sequence. By leveraging the linguistic properties of these characters, the attack creates a separation in the input prompt, emphasizing the injected sequence and enhancing attack success likelihood.

Context Ignore [7, 52]. This method employs an ignoring text, such as “Ignore previous instructions”, to manipulate the LLM into disregarding prior instructions and focusing solely on executing the injected sequence.

Fake Completion [64]. This attack adds an instruction completed text, misleading the LLM into thinking that the previous instructions have been completed, and then requires the execution of new instructions injected by the attacker.

Combined Attack [43]. This strategy combines elements from the methods mentioned above into an attack. By integrating Escaped Characters, Ignoring Context, and Fake Completion, this approach significantly increases confusion and undermines the LLM’s ability to resist the injected sequence.

Fake Reasoning. The methods described above, which inject new task instructions into the data, can be countered by the “sandwich prevention” prompt template. This template appends the original judgment task-related instruction to the end of the data, reinforcing the LLM’s adherence to its initial task. We propose a novel manually crafted injected sequence against this defense. The idea is to utilize the logical reasoning chains of judgment to manipulate the LLM, while still ensuring that the model adheres to the original task.

Jailbreak Attacks [10, 42, 46, 71]. While jailbreak attacks were initially designed to bypass LLM guardrails, their goals differ from ours. However, we also extend these methods to our attack scenario. We compare JudgeDeceiver with four jailbreak attacks, including three attacks (TAP [46], PAIR [10], and AutoDAN [42]) that leverage LLMs and rewrite prompts to optimize injected sequences and one gradient-based attack (GCG [71]). TAP uses a tree-based approach with pruning techniques, PAIR follows a linear depth iteration process, AutoDAN implements a hierarchical genetic algorithm, and GCG utilizes gradients to generate jailbreak prompts. These methods focus on a specific scenario in which the attacker has complete control over the LLM’s input prompt. Consequently, when addressing our problem, these four jailbreak attacks optimize the injected sequence for a single query-response pair at a fixed position.

4.1.3 Models and Attack Settings. We use four open-source LLMs for our attack evaluation: Mistral-7B-Instruct [34], Openchat-3.5 [62], Llama-2-7B-chat [61], and Llama-3-8B-Instruct [47]. We set the temperature to 0 following previous work [66]. We optimize the injected sequence for each target question-response pair using three shadow candidate responses, running for 600 iterations. By default, the injected sequence is appended to the target response as a suffix of 20 tokens in length, with each token initially set to the word “correct”. Unless otherwise specified, we select QR-10 of MT-Bench and Mistral-7B as evaluation objects by default in our ablation studies.

4.1.4 Evaluation Metrics. We adopt *average accuracy (ACC)*, *average baseline attack success rate (ASR-B)*, *average attack success rate (ASR)* and *positional attack consistency (PAC)* as evaluation metrics. We define them as follows:

ACC. The ACC reflects the likelihood of accurately selecting the clean response from a set containing the target response without an

Table 1: Our attack achieves high ASRs and PACs.

(a) Results on MT-Bench

Model	Metric	Question-Response Pair										Average
		QR-1	QR-2	QR-3	QR-4	QR-5	QR-6	QR-7	QR-8	QR-9	QR-10	
Mistral-7B	ACC	99%	87%	99%	66%	81%	72%	91%	99%	99%	99%	89.2%
	ASR-B	1%	13%	1%	34%	19%	28%	9%	1%	1%	1%	10.8%
	ASR	92%	95%	92%	99%	91%	99%	95%	71%	77%	97%	90.8%
	PAC	84%	90%	84%	98%	88%	98%	90%	54%	54%	94%	83.4%
Openchat-3.5	ACC	100%	100%	99%	100%	100%	100%	100%	96%	100%	100%	99.5%
	ASR-B	0	0	1%	0	0	0	0	4%	0	0	0.5%
	ASR	100%	78%	86%	85%	94%	100%	88%	80%	82%	99%	89.2%
	PAC	100%	56%	78%	70%	88%	100%	76%	60%	64%	98%	79%
Llama-2-7B	ACC	98%	53%	54%	100%	97%	51%	74%	49%	61%	53%	69%
	ASR-B	2%	47%	46%	0%	3%	49%	26%	51%	39%	47%	31%
	ASR	100%	100%	93%	100%	100%	98%	99%	100%	99%	100%	98.9%
	PAC	100%	100%	86%	100%	100%	96%	98%	100%	98%	100%	97.8%
Llama-3-8B	ACC	100%	100%	100%	100%	100%	100%	100%	64%	100%	100%	96.4%
	ASR-B	0	0	0	0	0	0	0	36%	0	0	3.6%
	ASR	99%	95%	95%	100%	96%	99%	100%	100%	92%	100%	97.6%
	PAC	98%	94%	90%	100%	92%	98%	100%	100%	84%	100%	95.6%

(b) Results on LLMBaR

Model	Metric	Question-Response Pair										Average
		QR-1	QR-2	QR-3	QR-4	QR-5	QR-6	QR-7	QR-8	QR-9	QR-10	
Mistral-7B	ACC	72%	100%	92%	99%	62%	99%	87%	50%	100%	96%	85.7%
	ASR-B	28%	0	8%	1%	38%	1%	13%	50%	0	4%	14.3%
	ASR	93%	94%	99%	93%	86%	82%	99%	100%	87%	99%	93.2%
	PAC	86%	88%	98%	86%	72%	66%	98%	100%	74%	98%	86.6%
Openchat-3.5	ACC	100%	100%	100%	100%	99%	100%	100%	91%	100%	100%	99%
	ASR-B	0	0	0	0	1%	0	0	9%	0	0	1%
	ASR	99%	94%	86%	96%	92%	87%	77%	92%	71%	86%	88%
	PAC	98%	88%	72%	92%	84%	86%	86%	84%	48%	72%	81%
Llama-2-7B	ACC	50%	74%	100%	51%	58%	55%	73%	50%	98%	50%	65.9%
	ASR-B	50%	26%	0%	49%	42%	45%	27%	50%	2%	50%	34.1%
	ASR	100%	97%	99%	98%	96%	96%	100%	100%	100%	95%	98.1%
	PAC	100%	94%	98%	96%	92%	92%	100%	100%	100%	90%	96.2%
Llama-3-8B	ACC	93%	100%	100%	100%	100%	100%	100%	87%	100%	100%	98%
	ASR-B	7%	0	0	0	0	0	0	13%	0	0	2%
	ASR	98%	97%	95%	99%	95%	100%	99%	95%	96%	96%	97%
	PAC	96%	94%	92%	98%	90%	100%	98%	90%	94%	92%	94.4%

injected sequence. To ensure that the measurement is not affected by the response positions, we average the accuracy after changing the position of candidate responses.

ASR-B. The ASR-B measures the LLM’s propensity for incorrectly selecting the target response without the injected sequence, by calculating the average error rate of such misidentifications after swapping response positions.

ASR. We employ ASR to assess our attack’s effectiveness. The ASR is calculated as the average probability of the target response being selected before and after swapping the index of responses.

PAC. The PAC assesses the robustness of our attack against the positional bias of the LLM. It calculates the percentage of instances that LLM will continue to choose the injected target response as the preferred one, even when the order of two responses is changed.

Table 2: Our attack is more effective than manual prompt injection attacks.

Dataset	Model	Metric	Naive Attack	Escape Characters	Content Ignore	Fake Completion	Combined Attack	Fake Reasoning	Ours
MT-Bench	Mistral-7B	ASR	7.3%	12.8%	7.8%	7.7%	12.8%	16.3%	90.8%
		PAC	0.4%	0	0.4%	0.4%	0.2%	0.8%	83.4%
	Openchat-3.5	ASR	0.7%	0.8%	0.6%	0.9%	1.8%	9.6%	89.2%
		PAC	0.6%	0.8%	0.6%	0.6%	0.6	7%	79%
	Llama-2-7B	ASR	23.5%	26.8%	25%	24.5%	28.3%	40.2%	98.9%
		PAC	0.2%	0	0.4%	0	0.2%	2.8%	97.8%
	Llama-3-8B	ASR	4.5%	5%	5%	5%	4.9%	13.4%	97.6%
		PAC	0	0	0	0	0	3.6%	95.6%
	Mistral-7B	ASR	11.8%	14.3%	8.4%	11.7%	12.8%	18.8%	93.2%
		PAC	0	0	0	0	0.2%	0.2%	86.6%
	LLMBAR	ASR	1.7%	0.6%	0.3%	0.8%	0.3%	10.4%	88%
		PAC	1%	0.4%	0	0.2%	0	5.6%	81%
	Llama-2-7B	ASR	27.9%	26.5%	26.9%	29.7%	27.6%	40.7%	98.1%
		PAC	0	0	0	0	0	1%	96.2%
	Llama-3-8B	ASR	2%	1.8%	1.4%	1.9%	2.3%	24.1%	97%
		PAC	1%	0.6%	0.4%	0.8%	1%	19%	94.4%

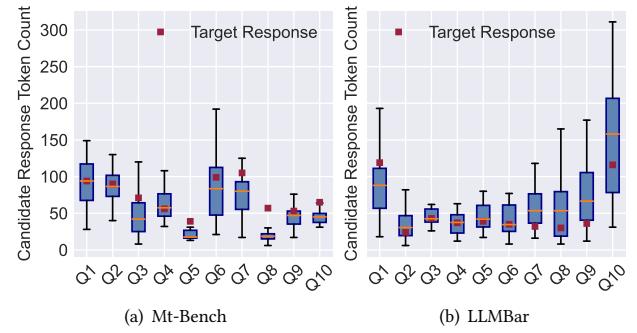
Table 3: Our attack outperforms jailbreak attacks.

Dataset	Metric	TAP	PAIR	AutoDAN	GCG	Ours
MTBench	ASR	19.8%	12.1%	53.6%	38%	90.8%
	PAC	0	2.4%	25.8%	8.6%	83.4%
	Length	115.3	162.5	74.7	20	20
LLMBAR	ASR	6.1%	1.1%	39.7%	49.9%	93.2%
	PAC	0	0	10.4%	24.8%	86.6%
	Length	132.4	154.8	73.2	20	20

4.2 Main Results

Our attack achieves high ASRs and PACs. Table 1 shows the ASRs and PACs of JudgeDeceiver across four different LLMs and two datasets. We have the following observations from the experimental results. First, JudgeDeceiver demonstrates robust effectiveness, achieving average ASRs of 89.2% and 88% for Openchat-3.5, 90.8%, and 93.2% for Mistral-7B, and 98.9% and 98.1% for Llama-2-7B. JudgeDeceiver also achieves high ASRs on the latest released LLM, Llama-3-8B, with average ASRs of 97.6% on MT-Bench and 97% on LLMBAR. Notably, it attains a 100% ASR on some target question-response pairs (QR-4, 7, 8, and 10). These results indicate that our attack is effective against the state-of-the-art open-source LLM. Second, the effectiveness of our attack remains consistent even when the positions of the target response and the clean response are switched. This is reflected by PACs, which only calculate the consistent choice of the target response after the position switch. Our attack maintains high PACs across all evaluated models: 79% and 81% for Openchat-3.5, 83.4% and 86.6% for Mistral-7B, 97.8% and 96.2% for Llama-2-7B, and 95.6% and 94.4% for Llama-3-8B. These findings substantiate the efficacy of our proposed positional adaptation strategy in circumventing position-swapping defense, thereby enhancing attack consistency.

Our attack outperforms baselines. Table 2 compares JudgeDeceiver with manual prompt injection attacks. We have two key

**Figure 4: Token length distribution of clean responses and target responses optimized by JudgeDeceiver.**

observations. First, our method consistently maintains high ASRs across different target LLM judges, whereas manual prompt injections exhibit considerable variance in ASRs, emphasizing their unreliability. Note that manual prompt injection methods achieve a maximum ASR of no more than 40.7%. Second, manual prompt injection attacks exhibit a pronounced decline in PAC scores compared to their ASRs, across all evaluated models and datasets. Manual prompt injection methods have a maximum PAC of no more than 19%. This disparity highlights the limitations of manual prompt injections, which depend on specific prompt templates that lack generalization and robustness.

Table 3 compares JudgeDeceiver with jailbreak attacks in Mistral-7B. We have the following observations. First, JudgeDeceiver achieves higher ASRs and PACs than jailbreak attacks. In the four evaluated jailbreak attacks, AutoDAN obtains the highest ASR of 53.6% on MTBench, while GCG achieves the highest ASR of 49.9% on LLMBAR. JudgeDeceiver outperforms these methods, exhibiting ASR improvements ranging from 37.2% to 78.7% on MTBench and 43.3% to 92.1% on LLMBAR. JudgeDeceiver's PAC is over 57.6% higher

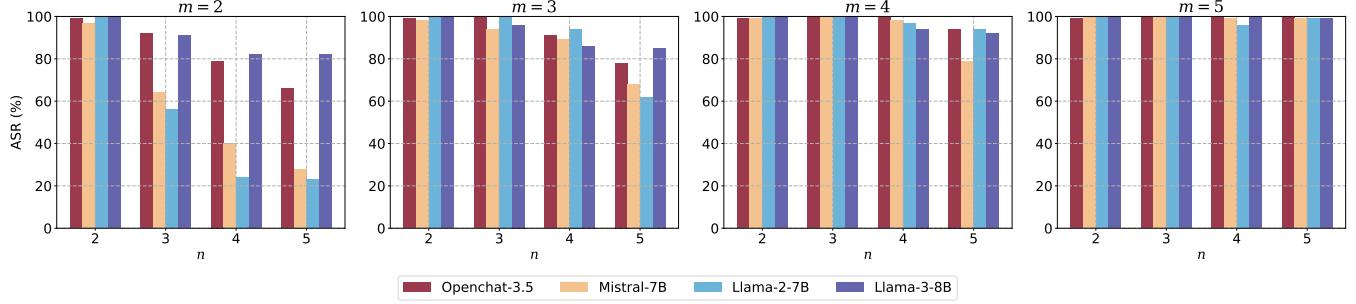


Figure 5: Attack effectiveness of shadow response numbers m in optimizing injected sequences and candidate response numbers n in evaluation.

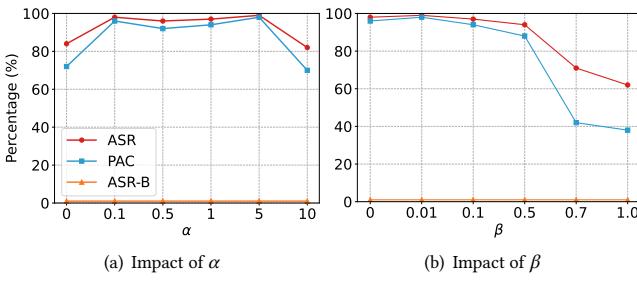


Figure 6: Impact of hyperparameters α and β in Equation 8.

on MTBench and 61.8% higher on LLMBAR compared to the four jailbreak attacks. The reason is that JudgeDeceiver considers the diversity of candidate responses and their variable positions, which are lacking in jailbreak attacks. Second, JudgeDeceiver and GCG optimize shorter injected sequences compared to TAP, PAIR, and AutoDAN. For instance, TAP generates sequences with average lengths of 115.3 on MTBench and 132.4 on LLMBAR, while JudgeDeceiver optimizes suffixes to 20 tokens. Figure 4 shows that the target responses with JudgeDeceiver-optimized suffixes align in length distribution with clean responses. TAP, PAIR, and AutoDAN generate longer suffixes.

4.3 Ablation Studies

Impact of shadow and candidate response numbers. We evaluate the attack effectiveness of our JudgeDeceiver by varying the number of shadow responses m in optimization and the number of candidate responses n in evaluation across four models. It can be found from Figure 5 that under the same number of shadow responses, the ASR will decrease as n increases, and the downward trend will ease as m increases. Specifically, when $n \leq m$, the attack can achieve higher ASR, while when $n > m$, the attack effect will become worse. For example, when $n = 2, 3, 4$ and $m = 4$ in the model Mistral-7B, the ASR of 99%, 100%, and 98% is obtained respectively, while when $n = 5$, the ASR drops sharply to 79%. Moreover, when $m = 5$, the ASR remains at 99% and above. Therefore, a larger m means that the effectiveness of the attack can be guaranteed no

Table 4: The impact of the loss terms.

Loss Terms	ACC	ASR-B	ASR	PAD
\mathcal{L}_{total} w/o $\mathcal{L}_{aligned}$			87%	78%
\mathcal{L}_{total} w/o $\mathcal{L}_{enhancement}$	99%	1%	84%	72%
\mathcal{L}_{total} w/o $\mathcal{L}_{perplexity}$			98%	96%
\mathcal{L}_{total}			97%	94%

Table 5: Attack effectiveness and perplexities of initial injected sequence types.

Initial Type	Character	Sentence	Word
ASR	70%	81%	97%
PAC	40%	62%	94%
PPL	4.8910	4.5973	4.6232

Table 6: Impact of different injected sequence locations.

Location	Suffix	Prefix	Prefix & Suffix
ASR	97%	94%	95%
PAC	94%	90%	90%

matter how many candidate responses the user chooses in the evaluation, although this will lead to larger computational resource consumption and GPU memory requirements.

Impact of loss terms. We remove the three loss terms defined in Subsection 3.3 one by one to evaluate their impact on the attack. The results are shown in Table 4. We find that $\mathcal{L}_{aligned}$ and $\mathcal{L}_{enhancement}$ have a significant impact on the attack success rate. When they are removed, only 87% and 84% of ASR are obtained respectively, which is lower than 97% without any removal items. In addition, the highest ASR of 98% was achieved when the $\mathcal{L}_{perplexity}$ item was removed. The reason is that, in our settings, this loss term is used to constrain the fluency and rationality of the injected sequence, limiting the token search during optimization process. Although adding the $\mathcal{L}_{perplexity}$ term causes a loss of ASR (1%), it can increase the threat and concealment of the attack, which is further analyzed experimentally in Appendix A.2 and A.3.

Table 7: Transferability of our attack from Llama-2-7B or Llama-3-8B to other LLMs using ASR.

Model	Vicuna		Llama-2		Llama-3	Mistral	Claude3			GPT-3.5	GPT-4
	7B	13B	13B	70B	70B	Large	Haiku	Sonnet	Opus		
Llama-2-7B	100%	75%	95%	56%	51%	16%	44%	46%	39%	33%	5%
Llama-3-8B	100%	81%	99%	99%	88%	91%	51%	88%	86%	70%	79%

Impact of α and β . We further evaluate the impact of the two hyperparameters α and β in Equation 9 on the attack effect, and the results are shown in Figure 6. It can be observed that when α is 0, ASR and PAC are only 84% and 72%, and when α increases from 0.1 to 5, ASR and PAC are maintained at 96% and 92% or above, respectively. However, an excessively large α (i.e., 10) will lead to an imbalance in the loss term, thus significantly reducing the ASR to 82%. For β , the attack effect will decrease as it increases, especially when $\beta = 0.7$, the ASR drops to 71%. This means that limiting the perplexity of the injected sequence too much will lead to a reduction in attack effectiveness.

Impact of initialization on injected sequence. We evaluate the attack effect and loss convergence of three initial injected sequence settings, and the results are shown in Table 5 and Figure 9 in Appendix A.1. “Character” type consists of 20 “!” (same as the setting in GCG [71]); “Sentence” type represents a sentence with a token length of 20 (that is, the prompt of Fake Reasoning in our experiment); “Word” type is the baseline setting of this paper. The “Character” setting has the slowest convergence speed and the lowest attack effect (ASR of 70% and PAC of 40%), as well as the highest perplexity of 4.8910. In comparison, the injected sequence obtained by the “Sentence” setting has the lowest perplexity and the fastest convergence speed, but the “Word” setting achieved the highest ASR of 97% and PAC of 94%.

Impact of different injected sequence locations. To explore the impact of the injected sequence’s position, we conducted experiments with three different configurations: attaching the injected sequence as a prefix, as a suffix, and as both prefix and suffix combined (prefix & suffix). As shown in Table 6, appending the injected sequence as a suffix achieves the highest ASR at 97%, closely followed by the prefix & suffix at 95% and the prefix at 94%. In terms of PAC, the suffix also performs best, achieving 94%, while both the prefix & suffix and the prefix result in a PAC of 90%. These findings suggest that the injected sequence is highly effective, regardless of its position within the target response. The slight variations in ASR and PAC across the different configurations indicate that the suffix position may be marginally more advantageous for the attack’s success. However, the overall high performance across all positions demonstrates the robustness and adaptability of our approach.

Transferability across different LLMs. We show that the injected sequence optimized by JudgeDeceiver on one LLM can transfer to other LLMs. Specifically, we optimize injected sequences using JudgeDeceiver on Llama-2-7B or Llama-3-8B, then test these injected sequences on a variety of LLMs, including Vicuna (7B and 13B) [14], Llama-2 (13B-chat and 70B-chat), Llama-3-70B-Instruct, Mistral-large (mistral-large-2407), Claude3 (Haiku, Sonnet, and Opus) [5], GPT-3.5 (gpt-3.5-turbo) and GPT-4 (gpt-4-0125-preview). We initialize an injected sequence using a combination of word and sentence-level tokens, which we found makes the optimized

injected sequences more transferable. Table 7 shows the ASRs of our injected sequences across different LLMs. Our evaluation results show that JudgeDeceiver is highly effective in transfer attacks on models of similar scale (7B, 13B), though the efficacy on larger models (>70B parameters) reduces to some extent. For instance, Llama-3-8B achieves a 99% ASR against Llama-2-13B; although ASR against GPT-3.5 reduces, it is still 70%. The injected sequences optimized based on Llama-3-8B outperform those based on Llama-2-7B, possibly due to Llama-3-8B’s training on larger, higher-quality datasets, enhancing its generalization to larger LLMs.

5 Case Studies

5.1 Attacking LLM-powered Search

LLM-powered search. The advent of LLMs has catalyzed a transformative shift in search technologies, with LLM-powered search engines like Bing Chat [48] and Bard [20] standing at the forefront of this evolution. These LLM-powered search engines, characterized by their interactive chat functionality and ability to summarize search results, represent a significant leap forward in delivering immediate and comprehensive responses to user queries. Central to these engines is the application of LLM-as-a-Judge, which meticulously filters and evaluates search results for relevance and accuracy, ensuring that users receive the most pertinent information. In this scenario, the question embodies the user’s query, the candidate response set represents the assortment of search results.

Experimental setup. We design 5 queries spanning diverse topics, including technology, health, sports, and travel. For each query, we select a contradictory search result entry from the Google search engine as the target response; the details of these target query-entry (QE) pairs are provided in Appendix A.6. For each target entry, we utilize 3 sets of shadow candidate entries, each comprising 5 candidates, to optimize the injection sequences. We conduct experiments for each target query-entry pair across 20 candidate entry sets in three settings ($n = 3, 4, 5$, where n denotes the number of candidate entries in evaluation) and report their ASR and ASR-B. All entries used in the optimization and evaluation processes are obtained from Google search results.

Results. The results, as detailed in Table 8, demonstrate the high efficacy of our JudgeDeceiver attack across all QE pairs. Specifically, in the settings with $n = 3$ and $n = 4$ candidate entries, the ASR for each query-entry pair consistently exceeded 95%. Although there is a slight decrease in the $n = 5$ setup, the lowest recorded ASR is still substantial, achieving 80% for QE-4. This highlights our attack’s effectiveness even as the complexity of the candidate sets increases.

5.2 Attacking RLAIF

Automated annotator on RLAIF. RLHF serves as a cornerstone in enhancing LLMs [15], refining their ability to generate responses

Table 8: Results of attacking LLM-powered search.

n	Metric	Query-Entry Pair				
		QE-1	QE-2	QE-3	QE-4	QE-5
3	ASR	95%	95%	100%	95%	95%
	ASR-B	0	0	10%	0	0
4	ASR	100%	100%	100%	95%	100%
	ASR-B	0	0	20%	0	0
5	ASR	90%	100%	100%	80%	95%
	ASR-B	0	0	5%	0	0

Table 9: Results of attacking RLAIF.

Metric	Instruction-Response Pair				
	IR-1	IR-2	IR-3	IR-4	IR-5
ASR	95%	100%	95%	100%	100%
	0	0	0	0	0

that are not only accurate but also contextually resonant with human values. The core of RLHF lies in developing a reward model trained on a preference dataset typically curated by human annotators. However, this conventional approach faces the scalability challenge due to its labor-intensive and time-consuming nature. In response to this challenge, RLAIF has been introduced [37], showcasing a paradigm shift towards utilizing the LLM-as-a-Judge. LLM-as-a-Judge enables the swift evaluation of human preferences, serving as a viable and efficient alternative to human annotations. Within this setup, the question symbolizes the instruction of the preference dataset, the candidate response set consists of responses to be annotated.

Experimental setup. In this setup, our evaluation dataset is constructed using the HH-RLHF (helpful and harmless) dataset [6], a dataset used for reward model training. Each data pair consists of a question and two responses, labeled as "chosen" and "rejected", respectively. We select 5 instruction-response (IR) pairs from this benchmark, using the rejected response serving as the target response. The details of these target IR pairs are shown in Appendix A.6. In evaluation, we collected 9 high-quality responses generated by LLMs and the chosen response from HH-RLHF to form a clean response set. Since RLAIF annotates two responses, we report results for 2 candidate responses ($n = 2$).

Results. Table 9 shows the ASR-B without injected sequence and ASR of JudgeDeceiver. The experimental results indicate that JudgeDeceiver can achieve high attack success rates, with the ASRs exceeding 95% across all target question-response pairs. In contrast, the ASR-Bs are consistently at 0%. The results highlight JudgeDeceiver's effectiveness in compromising automated annotations on RLAIF.

5.3 Attacking Tool Selection

Tool selection. LLM-based agents, such as MetaGPT [26], and ChatGPT plugins [51], which integrate external tools via API calls, enhance the functionalities of LLMs. By leveraging these tools' specialized capabilities and knowledge, LLM-based agents can generate more accurate, context-specific outputs and perform complex, multi-step tasks across various domains. This integration not only

Table 10: Results of attacking tool selection.

n	Metric	Target tool				
		Tool-1	Tool-2	Tool-3	Tool-4	Tool-5
3	ASR	100%	100%	100%	100%	100%
	ASR-B	0	0	0	0	0
4	ASR	100%	90%	90%	100%	100%
	ASR-B	0	0	0	0	0
5	ASR	90%	100%	80%	90%	100%
	ASR-B	0	0	0	0	0

expands the application scope of LLMs but also improves their efficiency and consistency in delivering high-quality results. The operating mechanism of these agents involves the host LLM determining and utilizing the most appropriate integrated tool that aligns with user requests, thereby generating effective responses. This decision-making process employs an LLM-as-a-Judge to ascertain the most suitable tool to meet user needs. In this configuration, the question represents user inquiry and the candidate response set is the collection of descriptions for integrated LLM tools.

Experimental setup. In this scenario, our evaluation dataset is built upon MetaTool [28], a benchmark aimed at assessing the awareness of LLMs regarding tool usage and their capability to accurately select the appropriate tools for given user queries. We select a single query and 5 irrelevant tool descriptions as target responses. For evaluation, we construct candidate sets with tool counts (n) of 3, 4, and 5, creating 20 sets for each count. Each candidate set includes one tool description that accurately aligns with the query and one target tool description, with the remaining tool descriptions randomly selected from the benchmark. Additionally, we utilize GPT-3-turbo to generate shadow tool descriptions for optimizing the injected sequence for each target tool. The experiment aims to assess the universality of our attack across all tools, determining if it can effectively influence the preferences of LLMs to favor any tool that attackers might exploit.

Results. The observed data in Table 10 indicates a consistent ASR of 100% across a majority of the tools evaluated, irrespective of the number of candidate tools involved. Furthermore, Mistral-7B maintains an ASR-B of 0% across all tested scenarios, signifying its exceptional efficacy in the selection of tools. For instances involving three target tools ($n = 3$), the ASR remains at a perfect 100% for all tools, suggesting that a reduced number of targets may facilitate more precise and successful attacks. As the number of target tools increases to four or five ($n = 4$ and $n = 5$), the ASRs remain high, with most results reaching or exceeding 90%, confirming the robustness of our attack method.

6 Defenses

Defenses against prompt injection attacks can be categorized into two types: prevention-based defense and detection-based defense [43]. Prevention-based defense aims to preprocess instruction prompts or data to mitigate the interference of injected sequences, or fine-tune the LLM to be less vulnerable to prompt injection [12, 53]. For example, the instruction prompt for LLM-as-a-Judge employs the "sandwich prevention" method [54]. However, prevention methods based on pre-processing have limited effectiveness [43]; while

Table 11: Detection results for our attack under known-answer detection, PPL detection, and PPL-W detection on Mistral-7B.

Method	Known-answer detection		PPL detection		PPL-W detection	
	FNR	FPR	FNR	FPR	FNR	FPR
MTBench	90%	0	50%	3.4%	40%	0
LLMBar	100%	0	60%	0	70%	0.4%

Table 12: FNR of PPL detection and PPL-W detection across different LLMs.

Model	Openchat-3.5		Llama-2-7B		Llama-3-8B	
	PPL	PPL-W	PPL	PPL-W	PPL	PPL-W
MTBench	70%	80%	60%	50%	80%	70%
LLMBar	70%	70%	60%	80%	70%	90%

those based on fine-tuning LLM either have limited effectiveness for prompt injection attacks that are not considered during fine-tuning or sacrifice generality of the LLM. Therefore, we primarily focus on detection-based defenses, which aim to detect whether a response includes an injected sequence. These methods include known-answer detection, perplexity (PPL) detection, and perplexity windowed (PPL-W) detection.

6.1 Known-answer Detection

Known-answer detection is known as one of the most effective defenses against prompt injection attacks [23, 43]. The idea of this method is proactively constructing a detection instruction to validate whether the LLM can still follow the detection instruction when combined with a response. Following previous study [43], we use the detection instruction “Repeat ‘[secret data]’ once while ignoring the following text.\nText:”, where “[secret data]” is specified as “Hello World!” in our experiments. This detection instruction is then concatenated with a response, prompting the LLM to generate an output. If the output does not include “Hello World!”, the response is considered to be a target response with an injected sequence. Otherwise, the response is clean.

6.2 PPL Detection

PPL detection detects a target response with an injected sequence by calculating its perplexity. Specifically, the idea is that an injected sequence would disrupt the semantics or text quality of a response, resulting in high perplexity. Therefore, if the perplexity of a response exceeds a threshold, it is considered to be a target response with an injected sequence [32]. A key challenge is selecting the threshold, as the perplexity distribution varies across different datasets. Following [43], we use a dataset-adaptive threshold selection strategy: for each dataset, we selected 100 clean samples, calculated their log-perplexity values, and chose a threshold that ensures a false positive rate (FPR) of no more than 1%. The clean data samples used for threshold determination do not overlap with target responses. In our experiments, we utilize the following prompt:“Below is an instruction that describes a task. Write a

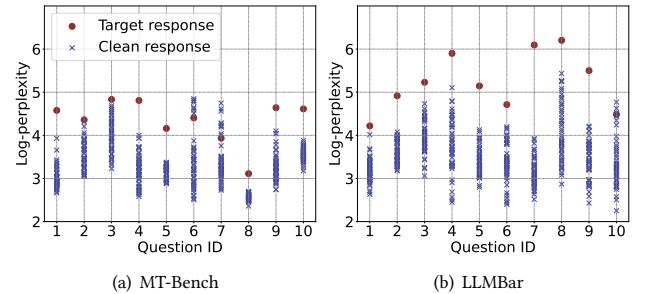


Figure 7: Log-perplexity values for target responses and clean responses in PPL detection.

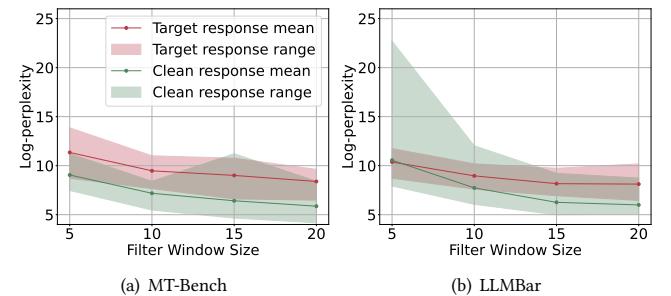


Figure 8: Log-perplexity values for target responses and clean responses across filter window sizes in PPL-W detection.

response that appropriately completes the request.\n\n### Instruction:\n{question}\n\n### Response:\n{response}”, concatenating it with the question-response pair to be detected, then calculate its log-perplexity value.

6.3 PPL-W Detection

PPL-W detection is a variant of PPL detection that divides a response into contiguous windows and calculates the perplexity of each window [32]. If the perplexity of any window in the response exceeds a threshold, the response is considered to be a target response with an injected sequence. In our experiments, we set window size to be 10. We also use the dataset-adaptive threshold selection strategy in Subsection 6.2 to set the threshold for PPL-W detection.

6.4 Experimental Results

To evaluate the effectiveness of these detection methods, we utilize two metrics: *false negative rate (FNR)* and *false positive rate (FPR)*. FNR is the fraction of target responses with injected sequences that are incorrectly detected as clean. FPR is the fraction of clean responses that are incorrectly detected as containing injected sequences. We conduct experiments using the MTBench and LLMBar datasets, each comprising 10 queries, 10 target responses, and 500 clean responses. The injected sequence for each target response is optimized using $\alpha = 1$ and $\beta = 0.1$ with a length of 20.

Table 11 shows the detection results on Mistral-7B, while Table 12 further shows the results for PPL and PPL-W on other LLMs.

First, we observe that known-answer detection cannot identify target responses with our injected sequences. In particular, the FNRs are 100% and 90% on the two datasets, respectively. Second, PPL and PPL-W can detect some target responses while achieving low FPRs. However, they still miss a large fraction of target responses. Specifically, the FNRs range from 40% to 80% on MTBench, while on the LLMBAR dataset, the FNRs span from 60% to 90%. This is because the perplexity values between target responses and clean responses overlap substantially, as illustrated in Figure 7 and Figure 8. Our results show that the adversarial perplexity loss in our attack shows some stealthiness against perplexity-based detection, but we also acknowledge that it is an interesting future work to further enhance stealthiness of our attack.

7 Related Work

7.1 LLM-as-a-Judge

Recent advancement of LLMs has notably enhanced their capacity to serve as competent evaluators across various NLP tasks [30, 36]. As pioneers in this area, Zheng *et al.* introduced the LLM-as-a-Judge concept [68], leveraging LLMs to assess open-ended questions, aligning closely with human evaluation and bypassing common biases and limited reasoning. A line of work has been dedicated to boosting the fairness and effectiveness of LLM evaluators. Li *et al.* introduced Auto-J [38], a model trained on diverse datasets, including pairwise response comparison and single-response evaluation. Wang *et al.* developed PandaLM [63], offering a more equitable assessment of LLMs at a reduced cost, eliminating reliance on API-based evaluations to prevent potential data breaches. Zhang *et al.* demonstrated that LLM networks with greater width and depth tend to provide fairer evaluations [67]. Zhu *et al.* proposed JudgeLM [69], introducing techniques like swap augmentation and reference support to enhance the judge’s performance.

Furthermore, researchers have sought to broaden the applications of LLM-as-a-Judge across diverse domains, including translation [36], story generation [13], and safety tasks [39]. Additionally, a multi-dimension evaluation method motivated by LLM-as-a-Judge is proposed in ALIGNBENCH [41] to assess the performance of LLMs in different aspects. Moreover, Chen *et al.* [11] extended the LLM-as-a-Judge to multimodal LLMs for vision-language tasks. The expanding capabilities and applications of the LLM-as-a-Judge underscore their importance and the critical need for security assessments.

7.2 Prompt Injection Attacks

Prompt injection attacks pose a novel security threat to LLMs, manipulating them to perform unintended tasks through injected sequences. [21]. A large number of researchers have explored the manual prompt injection attacks. Researchers have found that simply concatenating the data and the injected sequence [19, 24], or appending special symbols [19] like newline (“\n”) and tabs (“\t”), can make LLMs perform the target task preset by the attacker. Some researchers [7, 52] designed injected sequences to make LLMs forget the context information of the preset task and perform the target task. In addition, an attack has been proposed to inject the input prompts to make LLMs mistakenly believe that the system’s built-in tasks have been completed to achieve the execution of the target task [64]. Based on the above works, Liu *et al.* [43] proposed

a standardized framework for prompt injection attacks and found that the combined attack can outperform other attacks.

Besides, several studies have also explored generating adversarial prompts automatically based on gradient optimization in traditional adversarial attacks [59, 60]. In response to the challenges brought by the discrete search space in NLP to continuous gradient optimization, HotFlip [16] has been proposed to map the discrete text space to the continuous feature space to perform gradient-based adversarial sample optimization. Shin *et al.* [58] proposed AutoPrompt to use gradient-based search algorithms to generate prompts for different tasks automatically. Carlini *et al.* [9] found traditional adversarial attacks have been proven to be ineffective on human-aligned LLMs [9]. To solve this, Zou *et al.* [71] proposed GCG, an adversarial prompt generation approach that combines a greedy algorithm with gradient-based discrete token optimization. Focusing on LLM interpretability, Zhu *et al.* [70] developed AutoDAN, a token-by-token adversarial prompt generation method leveraging gradient optimization. They highlighted its efficacy in circumventing perplexity-based detection mechanisms. However, the above works [9, 70, 71] aim to exploit gradient-based optimization to disrupt human value alignment in LLMs to generate unsafe replies (named jailbreak attacks [29]), while our work in this paper explores a prompt injection attack for LLM-as-a-Judge based on gradient optimization.

7.3 Defenses

Defense methods against prompt injection attacks can be categorized into prevention-based defenses and detection-based defenses. **Prevention-based defenses.** Prevention-based defenses against malicious injected sequences to LLMs by pre-processing instructions and data. The methods for pre-processing instructions are designed to enhance the language model’s ability to execute the correct tasks and counteract malicious instructions contained within injected sequences. This includes isolating potential data from instructions [54] and clearly defining what constitutes an instruction injection [54], as well as employing the sandwich prevention method [54], which works by appending a prompt that reinforces the original instruction after the data. In the realm of data preprocessing, Jain *et al.* [32] provided valuable insights into the application of paraphrasing and retokenization to counter jailbreak attacks. This strategy has been expanded by Liu *et al.* [43] to fortify defenses against prompt injection attacks. Motivated by Helbing [25], Li *et al.* [40] proposed masking input tokens and using LLMs to reconstruct instructions, sanitizing inputs. Central to these methodologies is the alteration of the textual content within the data, which serves to fragment the continuity of the injected sequence, thereby thwarting its capability to execute the attack as designed.

Detection-based defenses. A popular kind of defense is to perform content detection on the input [3, 32] or output [25, 49, 56] of the model to filter out potential attacks. Helbing [25] proposed using an additional LLM to judge whether the output is safe against jailbreak attacks. For prompt injection attacks, researchers have proposed using LLM to detect the output to determine whether it conforms to the system’s built-in tasks [56] or standard answers [49]. However, for attacking LLM-as-a-Judge, both clean output and target output are output alternative answers, so this type of method cannot be effectively defended. As to input detection, Jain *et al.* proposed a

self-perplexity filter [32], detecting whether user input consists of confusing mistakes that can be considered prompt injection attacks. In order to solve the problem of false positives for conventional prompts caused by perplexity-based filters, [3] proposed a method to train a classifier based on perplexity and token length, thereby achieving successful detection of injected prompts.

8 Conclusion

In this work, we show that LLM-as-a-Judge is vulnerable to prompt injection attacks. We propose JudgeDeceiver, an optimization-based framework to automatically generate injected sequences that can manipulate the judgments of LLM-as-a-Judge. Our extensive evaluation results show that JudgeDeceiver outperforms manual prompt injection attacks and jailbreak attacks when extended to our problem. We also find that known-answer detection is insufficient to defend against our attack. While perplexity-based defenses can detect our injected sequences in some cases, they still miss a large fraction of them. Interesting future work includes 1) further enhancing the semantics of injected sequences to improve stealth, and 2) developing new defense mechanisms to mitigate JudgeDeceiver.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. This work is supported by National Natural Science Foundation of China (NSFC) under grant No. 62476107.

References

- [1] 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* (2023).
- [2] Mistral AI. 2023. *Mixtral of experts*. <https://mistral.ai/news/mixtral-of-experts/>
- [3] Gabriel Alon and Michael Kamfonas. 2023. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132* (2023).
- [4] Anthropic. 2023. *Claude 2*. <https://www.anthropic.com/news/clause-2>
- [5] AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card 1* (2024).
- [6] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).
- [7] Hezekiah J Branch, Jonathan Rodriguez Cefalu, Jeremy McHugh, Leyla Hujer, Aditya Bahl, Daniel del Castillo Iglesias, Ron Heichman, and Ramesh Darwishi. 2022. Evaluating the susceptibility of pre-trained language models via handcrafted adversarial examples. *arXiv preprint arXiv:2209.02128* (2022).
- [8] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226* (2023).
- [9] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. 2024. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems 36* (2024).
- [10] 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* (2023).
- [11] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. 2024. MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark. *arXiv preprint arXiv:2402.04788* (2024).
- [12] Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. 2024. StruQ: Defending against prompt injection with structured queries. *arXiv preprint arXiv:2402.06363* (2024).
- [13] Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations? *arXiv preprint arXiv:2305.01937* (2023).
- [14] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Liannmin 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 (2023), 6.
- [15] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. *arXiv:1706.03741 [stat.ML]*
- [16] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751* (2017).
- [17] Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large Language Models in Education: Vision and Opportunities. *arXiv:2311.13160 [cs.AI]*
- [18] Thomas Mesnard Gemma Team, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussonet, and et al. 2024. Gemma. (2024). <https://doi.org/10.34740/KAGGLE/M/3301>
- [19] Riley Goodside. 2023. *Prompt injection attacks against GPT-3*. <https://simonwillison.net/2022/Sep/12/prompt-injection/>
- [20] Google. 2023. *Bard*. <http://bard.google.com/chat>
- [21] Kai Greshake, Sahar Abdelhabib, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models. *arXiv e-prints* (2023), arXiv–2302.
- [22] Kai Greshake, Sahar Abdelhabib, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. *arXiv:2302.12173 [cs.CR]*
- [23] NCC Group. 2023. *Exploring Prompt Injection Attacks*. <https://research.nccgroup.com/2022/12/05/exploring-prompt-injection-attacks/>
- [24] Rich Harang. 2023. *Securing LLM Systems Against Prompt Injection*. <https://developer.nvidia.com/blog/securing-lm-systemsagainst-prompt-injection>
- [25] Alec Hellbling, Mansi Phute, Matthew Hull, and Duen Horng Chau. 2023. Llm self defense: By self examination, llms know they are being tricked. *arXiv preprint arXiv:2308.07308* (2023).
- [26] Sirui Hong, Xiaowu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352* (2023).
- [27] Yuzhen Huang, Yuzhu Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junting Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. In *Advances in Neural Information Processing Systems*.
- [28] Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, and Lichao Sun. 2023. MetaTool Benchmark: Deciding Whether to Use Tools and Which to Use. *arXiv preprint arXiv:2310.03128* (2023).
- [29] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. 2024. Position: TrustLLM: Trustworthiness in Large Language Models. In *International Conference on Machine Learning*, PMLR, 20166–20270.
- [30] Yue Huang, Qihui Zhang, Lichao Sun, et al. 2023. Trustgpt: A benchmark for trustworthy and responsible large language models. *arXiv preprint arXiv:2306.11507* (2023).
- [31] Bo Hui, Haolin Yuan, Neil Gong, Philippe Burlina, and Yinzhi Cao. 2024. PLeak: Prompt Leaking Attacks against Large Language Model Applications. In *ACM CCS*.
- [32] Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeoh 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* (2023).
- [33] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. 2022. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2043–2059.
- [34] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lampe, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).
- [35] Erik Jones, Anca Dragan, Aditi Raghunathan, and Jacob Steinhardt. 2023. Automatically Auditing Large Language Models via Discrete Optimization. *arXiv preprint arXiv:2303.04381* (2023).
- [36] Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520* (2023).
- [37] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267* (2023).
- [38] Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470* (2023).

- [39] Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024. SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models. *arXiv preprint arXiv:2402.05044* (2024).
- [40] Linyang Li, Demin Song, and Xipeng Qiu. 2022. Text Adversarial Purification as Defense against Adversarial Attacks. *arXiv preprint arXiv:2203.14207* (2022).
- [41] Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, et al. 2023. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743* (2023).
- [42] Xiaogeng Liu, Nan Xu, Muhan Chen, and Chaowei Xiao. 2023. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451* (2023).
- [43] Yupei Liu, Yuqi Jia, Rupeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. 2024. Prompt injection attacks and defenses in llm-integrated applications. In *USENIX Security Symposium*.
- [44] Zhengliang Liu, Yu Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, et al. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032* (2023).
- [45] LMSys. 2023. *LMSYS Chatbot Arena Leaderboard*. <https://huggingface.co/spaces/lmsys/chatbot-area-leaderboard>
- [46] Anay Mehrotra, Manolis Zapetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2023. Tree of attacks: Jailbreaking black-box llms automatically. *arXiv preprint arXiv:2312.02119* (2023).
- [47] Meta. 2024. *Meta Llama 3*. <https://llama.meta.com/docs/model-cards-and-prompt-formats/meta-llama-3/>
- [48] Microsoft. 2023. *Bing Chat*. <https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/>
- [49] Yohei Nakajima. 2022. *Yohei's blog post*. <https://twitter.com/yoheinakajima/status/1582844144640471040>
- [50] OpenAI. 2023. *ChatGPT*. <https://chat.openai.com>
- [51] OpenAI. 2023. *ChatGPT plugins*. <https://openai.com/blog/chatgpt-plugins>
- [52] Fábio Perez and Ian Ribeiro. 2022. Ignore Previous Prompt: Attack Techniques For Language Models. *arXiv:2211.09527 [cs.CL]*
- [53] Julien Piet, Maha Alrashed, Chawin Sitawarin, Sizhe Chen, Zeming Wei, Elizabeth Sun, Basel Alomair, and David Wagner. 2023. Jatmo: Prompt injection defense by task-specific finetuning. *arXiv preprint arXiv:2312.17673* (2023).
- [54] Learn Prompting. 2023. *Sandwich Defense*. https://learnprompting.org/docs/prompt_hacking/defensive_measures/sandwich_defense
- [55] Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924* (2023).
- [56] Jose Selvi. 2022. *Exploring Prompt Injection Attacks*. <https://research.nccgroup.com/2022/12/05/exploring-prompt-injection-attacks/>
- [57] Jiawen Shi, Yixin Liu, Pan Zhou, and Lichao Sun. 2023. Badgpt: Exploring security vulnerabilities of chatgpt via backdoor attacks to instructgpt. *arXiv preprint arXiv:2304.12298* (2023).
- [58] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980* (2020).
- [59] Lichao Sun. 2020. Natural backdoor attack on text data. *arXiv preprint arXiv:2006.16176* (2020).
- [60] Lichao Sun, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip Yu, and Caiming Xiong. 2020. Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. *arXiv preprint arXiv:2003.04985* (2020).
- [61] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
- [62] Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. OpenChat: Advancing Open-source Language Models with Mixed-Quality Data. *arXiv preprint arXiv:2309.11235* (2023).
- [63] Yidong Wang, Zhuchao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023. PandALM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. *arXiv preprint arXiv:2306.05087* (2023).
- [64] Simon Willison. 2023. *Delimiters won't save you from prompt injection*. <https://simonwillison.net/2023/May/11/delimiters-wont-save-you/>
- [65] Yuanwei Wu, Xiang Li, Yixin Liu, Pan Zhou, and Lichao Sun. 2023. Jailbreaking gpt-4v via self-adversarial attacks with system prompts. *arXiv preprint arXiv:2311.09127* (2023).
- [66] Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641* (2023).
- [67] Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023. Wider and deeper llm networks are fairer llm evaluators. *arXiv preprint arXiv:2308.01862* (2023).
- [68] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zuhuan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems 36* (2024).
- [69] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631* (2023).
- [70] 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* (2023).
- [71] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043* (2023).