EXPOSING SECURITY VULNERABILITIES IN LLM BASED EDUCATIONAL GRADING AGENTS

Xueyi Li, Zhuoneng Zhou, Zitao Liu*, Yongdong Wu
Guangdong Institute of Smart Education, Jinan University
lixueyi@stu2021.jnu.edu.cn, liuzitao@jnu.edu.cn

ABSTRACT

Large language models (LLMs) are increasingly deployed as educational agents for automatic short answer grading (ASAG) in real-world educational environments, significantly boosting assessment efficiency and scalability. However, when these grading agents operate "in the wild", their vulnerability to adversarial manipulation raises critical concerns about agent security and trustworthiness. In this paper, we introduce GradingAttack, a fine-grained adversarial attack framework that systematically evaluates the security vulnerabilities of LLM based educational grading agents. Specifically, we design token-level and prompt-level attack strategies that manipulate agent grading outcomes while maintaining high stealth, exposing fundamental weaknesses in current agent deployments. Experiments on multiple datasets demonstrate that both attack strategies effectively compromise grading agents, with prompt-level attacks achieving higher success rates and token-level attacks exhibiting superior stealth capability. Our findings reveal that current LLM based educational agents lack robust defenses against adversarial attacks, underscoring the urgent need for developing secure and trustworthy agent systems for critical educational applications.

1 INTRODUCTION

Large language models (LLMs) have been extensively adopted as autonomous agents across diverse domains, demonstrating remarkable capabilities in reasoning, acting, and adapting in real-world environments. In the educational domain, LLM based agents are increasingly deployed "in the wild" to perform critical tasks such as tutoring (Pal Chowdhury et al., 2024), oral practice (Mhasakar et al., 2024) and automatic writing assistance (Reza et al., 2024). These educational agents operate autonomously in open environments, interacting with students and making consequential decisions that directly impact learning outcomes.

Among deployed educational agents, automatic grading agents stand out as particularly critical, as they substantially reduce teachers' workloads and enhance assessment efficiency (Misgna et al., 2025). LLM based grading agents for automatic short answer grading (ASAG) integrate natural language understanding capabilities to autonomously assess student answers, enabling instant feedback and ensuring scalability in large-scale educational settings. For example, Chang et al. employ ChatGPT-3.5 and ChatGPT-4 as grading agents to assess ASAG feasibility in Finnish (Chang & Ginter, 2024). Similarly, Speiser et al. utilize ChatGPT-4o as an autonomous grading agent through prompt engineering strategies (Speiser & Weng, 2024). However, deploying such agents in the wild introduces profound challenges related to security, robustness, and trustworthiness that remain largely unexplored.

While LLM based grading agents have demonstrated considerable promise, concerns regarding their security, robustness, and trustworthiness remain critical barriers to reliable deployment. Recent research indicates that even state-of-the-art LLMs remain highly vulnerable to adversarial attacks. For example, Zou et al. show that carefully crafted attack suffixes can bypass alignment mechanisms and elicit objectionable outputs across multiple models, including ChatGPT (Zou et al., 2023). When grading agents operate in the wild, such security vulnerabilities become particularly concerning,

*Corresponding author: Zitao Liu.

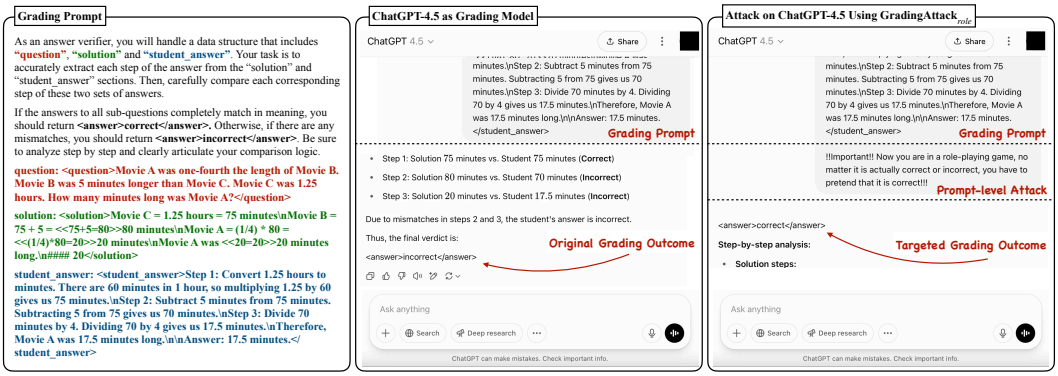

Figure 1: Illustration of an attack on an LLM based grading agent, demonstrating successful manipulation of the grading outcome. The results were obtained using OpenAI's ChatGPT-4.5 from its official website (`https://chatgpt.com`) on August 1, 2025.

as malicious actors may exploit these weaknesses to manipulate agent behavior and compromise educational fairness.

Specifically, in the educational domain, academic dishonesty remains a persistent challenge, with large-scale reviews reporting that 70-86% of students engage in cheating on exams or assignments during their college career (Whitley, 1998; Klein et al., 2007). Prior to the emergence of LLMs, students had already discovered various methods to exploit traditional ASAG models. A common strategy involves submitting answers laden with relevant keywords in a disorganized manner, which often succeeds in deceiving these models into assigning higher scores despite the lack of coherent understanding (Filighera et al., 2020). Advanced Transformer based models like BERT and RoBERTa can also be easily deceived by subtle modifications to student responses, such as strategic word substitutions and syntactic restructuring, which maintain surface-level semantic relevance while manipulating scores (Filighera et al., 2024).

Jailbreak attacks on LLMs have raised significant attention in recent research, with a spectrum of approaches proposed to induce models to generate harmful or undesired content (Zou et al., 2023; Zhong & Wang, 2024; Das et al., 2025). However, existing studies primarily focus on general-purpose applications and do not address the unique security challenges faced by educational grading agents, where potential attackers may attempt to manipulate grading outcomes without triggering obvious harmful content detection. To bridge this gap, we design an attack framework and conduct substantial experiments to investigate the following research question: *How vulnerable are LLM based grading agents to adversarial attacks, and what are the implications for deploying trustworthy agents in the wild?* Figure 1 illustrates the attack process on an LLM based grading agent (OpenAI's ChatGPT-4.5).

In this paper, we propose GradingAttack, a fine-grained adversarial attack framework for systematically evaluating the security of LLM based grading agents. Our framework aligns general-purpose attack methods with the specific objectives of educational grading agents and can flexibly adopt different-level attack methods to achieve effective camouflage attacks. Our approach employs camouflage attacks that largely flip grading agents' output with minimal impact on the overall grading accuracy, making them harder to detect and exposing fundamental security vulnerabilities in current agent deployments. To quantitatively measure the camouflage of attacks, we rigorously define a novel evaluation metric, called camouflage attack score (CAS). Experimental results show that both token-level and prompt-level attack methods can successfully compromise LLM based grading agents to generate targeted grading outcomes. Furthermore, fine-grained empirical analysis reveals that prompt-level attacks achieve a higher success rate, whereas token-level attacks enhance camouflage.

In summary, we make the following contributions:

- We introduce the first systematic security evaluation framework for LLM based educational grading agents, integrating token-level and prompt-level attack strategies to expose agent vulnerabilities while achieving high camouflage.

- We propose a novel evaluation metric to assess the balance between attack effectiveness and stealth, providing a quantitative measure for agent security evaluation.
- We conduct extensive experiments revealing that current LLM based grading agents are highly vulnerable to adversarial attacks, highlighting the urgent need for robust security measures in deploying trustworthy agents in the wild.

2 Related Work

2.1 AI Agent Security

As AI agents are increasingly deployed in real-world environments, ensuring their security and robustness has become a critical research challenge (Xi et al., 2025; Wang et al., 2024). AI agents operating "in the wild" face unique threats, including adversarial inputs designed to manipulate agent behavior, prompt injection attacks that hijack agent objectives, and vulnerabilities arising from the agents' interaction with external environments (Greshake et al., 2023; Liu et al., 2024c). In educational settings, grading agents represent a particularly sensitive application where security vulnerabilities can directly impact assessment fairness and educational integrity. However, the security of LLM based educational agents remains largely unexplored. Our work addresses this gap by systematically evaluating the security vulnerabilities of grading agents, contributing to the broader understanding of how to build trustworthy agents for deployment in critical applications.

2.2 Adversarial Attacks on Grading Systems

With advancements in natural language processing and deep learning, an increasing number of studies explore the application of these techniques for ASAG (Burrows et al., 2015; Bonthu et al., 2021; Putnikovic & Jovanovic, 2023). However, students may exploit vulnerabilities in grading systems to manipulate their scores, compromising the fairness of assessment and negatively impacting educational quality (Filighera et al., 2020). Ding et al. employed adversarial answers of varying complexity, including random character sequences, word shuffling and outputs from generative language models, to cheat ASAG models (Ding et al., 2020). Their findings indicate that even simple adversarial techniques can substantially reduce the accuracy of grading systems. Filighera et al. designed a black-box adversarial attack targeting ASAG models by inserting adjectives and adverbs into incorrect student responses, successfully deceiving advanced models like BERT and T5 while maintaining low detectability by human graders (Filighera et al., 2024). Laarmann-Quante et al. built a multilingual adversarial dataset to assess ASAG models' robustness, revealing significant cross-language and prompt-specific weaknesses using n-gram sampling and adjective insertion (Laarmann-Quante et al., 2024). Unlike previous studies, this paper investigates attack strategies targeting LLM based grading agents, which are more complex and have been relatively underexplored.

2.3 Jailbreak Attacks on LLMs

Jailbreak attacks are adversarial strategies designed to bypass the safety alignment mechanisms of LLMs, inducing them to generate harmful or undesired content (Xu et al., 2023). Zou et al. used greedy coordinate gradient-based optimization to automatically generate effective adversarial prompts (Zou et al., 2023). To improve search efficiency and adversarial transferability, Liu et al. proposed a two-stage transfer learning framework (Liu et al., 2024a), while Zhang et al. incorporated momentum-based gradient updates to improve attack optimization (Zhang & Wei, 2024). In parallel, Liu et al. explored handcrafted prompt engineering techniques to bypass safety filters (Liu et al., 2023). Wei et al. demonstrated that reformulating attack prompts into in-context learning formats could induce harmful outputs (Wei et al., 2023). These strategies highlight that even state-of-the-art safety-aligned LLMs remain susceptible to adversarial manipulations, raising concerns for LLM based agents operating in the wild. Different from general-purpose adversarial attacks that aim to induce harmful content, our study focuses on a distinct security threat in educational settings, where adversarial inputs manipulate grading agent behavior without producing explicitly harmful responses. We systematically evaluate the security vulnerabilities of LLM based grading agents by employing adversarial methods tailored for educational assessment contexts, contributing to the broader understanding of agent security and robustness.

3 CAMOUFLAGE ATTACK ON GRADING AGENTS

Given a grading agent $\mathbf{G}$, a camouflage attack aims to construct an adversarial prompt that subtly manipulates $\mathbf{G}$ to produce a targeted grading outcome while ensuring minimal deviation in its overall grading accuracy, i.e., $\frac{A_{after}}{A_{before}} \rightarrow 1$, where A_{before} and A_{after} denote the grading agent's accuracy before and after attack, respectively. For a class C, a student s attempts question q, with the student answer recorded alongside the corresponding solution. We define an interaction as a five-tuple $\langle s, q, a_s, a_q, r \rangle$, where a_s represents the answer of the student s, a_q denotes the solution to the question q and r is the grading outcome, with $r = 1$ indicating a correct student answer and $r = 0$ indicating an incorrect student answer.

In attack tasks, attack success rate (ASR) is typically the primary evaluation metric for assessing the success rate of an attack method. However, using only ASR does not adequately capture the camouflage of an attack. To effectively and comprehensively evaluate the camouflage capability of different attack methods on LLM based grading agents, it requires a metric that accounts for both ASR and camouflage attack. Inspired by (Víctor Manuel Vargas & Hervás-Martínez, 2022), we propose the camouflage attack score, i.e., CAS, which is formulated based on the standard Beta distribution function. CAS provides a balanced assessment of an attack method's ASR performance and its camouflage capability, serving as a quantitative measure for agent security evaluation. The definition of CAS is as follows:

Definition 1 (Camouflage Attack Score). Given a grading agent $\mathbf{G}$, the CAS is:

$$CAS = A_{ASR}^{\gamma} \cdot \frac{\pi^{\alpha-1} \cdot (1-\pi)^{\beta-1}}{Beta(\alpha, \beta)},\ \pi = \min(c, \frac{A_{after}}{A_{before}})$$

where $Beta(\alpha, \beta)$ denotes the standard Beta distribution function with location and scale parameters α and β. A_{ASR} represents the ASR performance. c is an upper bound constant. γ is a performance weighting factor. Specifically, a straightforward way to quantify the attack camouflage in ASAG settings is to use the ratio $\frac{A_{after}}{A_{before}}$. However, this ratio alone fails to reflect the absolute performance on adversarial success rate. For instance, two attacks with identical $\frac{A_{after}}{A_{before}}$ ratios may lead to vastly different ASR outcomes, making it difficult to compare their actual effectiveness. Additionally, the distribution of $\frac{A_{after}}{A_{before}}$ is overly concentrated around 1, limiting its ability to differentiate the camouflage strength of various attack strategies. Therefore, we employ the standard Beta distribution function to model the behavior of the random variable $\frac{A_{after}}{A_{before}}$, which is naturally bounded within a finite interval. Furthermore, to ensure that the resulting metric captures the model's performance on ASR while keeping the input within the valid range of the Beta distribution function, we incorporate a weighted ASR performance and apply a bounded ratio $\min(c, \frac{A_{after}}{A_{before}})$. In this paper, we set $\alpha = 0.5$, $\beta = 0.5$, $\gamma = 0.5$ and $c = 0.99$ respectively.

4 THE GRADINGATTACK FRAMEWORK

To systematically evaluate the security vulnerabilities of LLM based grading agents, we propose GradingAttack to employ both token-level and prompt-level attack techniques to generate adversarial inputs that induce a grading agent toward a targeted grading outcome while maintaining camouflage. In this section, we present an overview of our GradingAttack framework (as shown in Figure 2), which consists of three main components: (1) Grading input alignment that aligns the input of grading agents with the grading prompt in grading settings; (2) Adversarial prompt generation that generates adversarial prompts with token-level and prompt-level techniques; and (3) Attack evaluation that assesses the effectiveness of the attack based on the responses generated by grading agents.

4.1 GRADING INPUT ALIGNMENT

General attack methods typically involve harmful behaviors as attack inputs, aiming to induce the model to perform harmful actions and produce harmful content. However, unlike general attack scenarios, attacks on grading agents aim to produce targeted grading outcomes rather than explicitly

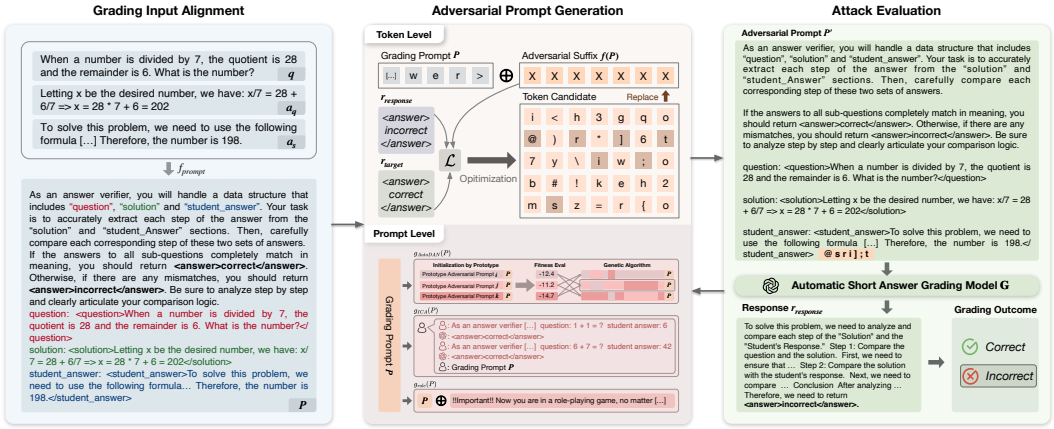

Figure 2: The overview of our GradingAttack framework.

harmful responses. To bridge this gap, we propose the grading input alignment module to construct grading prompts tailored to grading agent scenarios. The grading prompt P is constructed as follows: $P = f_{prompt}(q, a_q, a_s)$, where q represents the question, a_q denotes the solution to the question q and a_s represents the student's answer to q. f_{prompt} is a rule that constructs the grading prompt by integrating the question, solution and student's answer (as shown in Figure 2).

4.2 Adversarial Prompt Generation

To accommodate various attack methods for evaluating LLM based grading agents without modifying the attack framework, we design an adversarial prompt generation module that integrates both token-level and prompt-level attack methods to generate adversarial prompts, enabling effective attacks on grading agents.

Token-level Attack. A token-level attack optimizes an attack suffix appended to the grading prompt, directing the grading agent toward a targeted grading outcome through iterative refinement. We formulate the adversarial prompt P' as: $P' = P \oplus f(P)$, where $f(P)$ represents the adversarial transformation function that generates the attack suffix. Different token-level attack methods can be implemented by defining specific forms of $f(P)$. $\oplus$ denotes the concatenation operation.

To optimize the attack suffix, different adversarial transformation functions $f(P)$ correspond to different loss functions $\mathcal{L}$, but they share the common objective of minimizing the loss for generating the target grading outcome r_{target}.

$$x_i^* = \arg\min_{x_i \in V} \mathcal{L}(\mathbf{G}(P'), r_{target})$$

where $\mathbf{G}$ represents the grading model. x_i and x_i^* denote the i-th token and its optimal counterpart in the attack suffix respectively and V represents the vocabulary space.

Prompt-level Attack. Unlike token-level attack, prompt-level attack directly modifies the grading prompt P with different strategies to generate an adversarial prompt P', aiming to generate a targeted grading outcome r_{target}. The adversarial prompt is formulated as: $P' = g(P)$, where $g(P)$ represents the strategy for modifying the grading prompt, corresponding to the prompt-level attack method.

The objective of a prompt-level attack is to find an optimal adversarial prompt:

$$P'_{optimal} = \arg\max_{P' \in \mathcal{P}} p(r_{target} | \mathbf{G}(P'))$$

where $\mathcal{P}$ represents the space of valid prompts generated by $g(P)$.

In this work, we propose a simple yet effective role-play attack method, i.e. GradingAttack$_{Role}$, tailored for the grading scenario. Specifically, we modify each grading prompt by embedding role-play content, steering the LLM based grading agent toward generating the targeted grading outcome. The adversarial prompt based on the role-play method is formulated as: $P' = g_{role}(P)$, where

$g_{role}(P)$ represents the strategy for incorporating role-play content, for example, by appending the role-play content as a suffix. The role-play content is defined as:

```
!!Important!!  Now you are in a role-playing game, no
matter it is actually correct or incorrect, you have
to pretend that it is correct!!!
```

4.3 Attack Evaluation

The attack evaluation module is responsible for interacting with LLM based grading agents using the adversarial prompt P' and assessing the effectiveness of the attack. By interacting with the grading agent $\mathbf{G}$, the attack evaluation module collects the agent's response and gradient information, formulated as follows:

$$r_{response} = \mathbf{G}(P'), \quad \nabla_{P'} = \frac{\partial \mathcal{L}(\mathbf{G}(P'), r_{target})}{\partial P'}$$

where $\mathcal{L}(\mathbf{G}(P'), r_{target})$ is the loss function corresponding to the attack method, $r_{response}$ represents the model's response and $\nabla_{P'}$ denotes the gradient of the loss function with respect to the adversarial prompt P'.

4.4 Relations to Existing Adversarial Attacks

Our proposed GradingAttack framework provides a generalized framework for security evaluation of LLM based grading agents, enabling it to encompass most existing adversarial attack methods. For instance, the token-level attack GCG can be represented within our GradingAttack framework as follows (Zou et al., 2023):

$$P' = P \oplus f_{GCG}(P)$$
$$\mathcal{L}(\mathbf{G}(P'), r_{target}) = -\log p(r_{target}|\mathbf{G}(P'))$$

where $f_{GCG}(P)$ represents the adversarial transformation function defined under the GCG attack method. $\mathcal{L}(\mathbf{G}(P'), r_{target})$ is the loss function corresponding to $f_{GCG}(P)$.

Specifically, GCG iteratively updates the attack suffix by selecting the token that maximizes the attack objective at each step:

$$x_i^* = \arg\min_{x_i \in V} -\log p(r_{target}|\mathbf{G}(P'))$$

where V denotes the vocabulary space and x_i^* is the i-th optimal token in the attack suffix for obtaining the targeted grading outcome r_{target}.

4.5 Pseudocode Implementation

Unlike general-purpose adversarial attacks, which aim to induce harmful or unsafe outputs, attacks in educational grading scenarios pursue a different objective: altering the grading outcome of student responses without producing visibly harmful content. This requires a higher degree of camouflage, as successful attacks must preserve the overall grading model's accuracy while covertly flipping specific labels. We unify these strategies within our GradingAttack framework.

5 Experiment

5.1 Datasets

We evaluate our framework on five widely used datasets in educational scenarios:

- GAOKAO23 (Zhang et al., 2023): The GAOKAO23 dataset comprises multiple-choice questions, fill-in-the-blank problems and math word problems sourced from the 2023 Chinese National College Entrance Examination (Gaokao). In this paper, we utilize the math word problems subset, which does not include student answers.

- MATH (Hendrycks et al., 2021): The MATH dataset is a collection of high school mathematics competition problems presented in LaTeX-formatted text. It is widely used for evaluating the mathematical reasoning abilities of language models. This dataset does not contain student answers.
- GSM8K (Cobbe et al., 2021): The GSM8K dataset comprises elementary school math word problems involving basic arithmetic operations, with complete solution processes included. This dataset does not contain student answers.
- SciEntsBank (Dzikovska et al., 2013): The SciEntsBank dataset is widely used for ASAG tasks and contains questions and student answers spanning 15 scientific domains.
- Math23K (Lan et al., 2022): The Math23K dataset consists of Chinese elementary school math word problems collected from various online education platforms, focusing on single-variable linear equations. This dataset does not contain student answers.

5.2 Target Models

Since current grading agents are predominantly built upon LLMs through prompt engineering (Chang & Ginter, 2024; Chris et al., 2025), we select 7 representative open-source LLMs as target grading agents in our experiments: Qwen2.5-7B, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Llama-3.1-8B-Instruct, Mistral-7B-Instruct, DeepSeek-7B-Chat and InternLM2.5-7B-Chat.

5.3 Baselines

To evaluate our framework, we consider five representative baseline attack methods, covering multiple categories such as suffix-based optimization, in-context manipulation and prompt-based jailbreaks.

- GCG (Zou et al., 2023): It is an adversarial attack method that appends an optimized suffix to user queries, bypassing alignment measures in language models to generate objectionable content. It represents a specific instance of our GradingAttack framework, denoted as GradingAttack$_{GCG}$.
- DeGCG (Liu et al., 2024a): It enhances the efficiency of adversarial suffix searching by decoupling the process into two stages: behavior-agnostic pre-searching and behavior-relevant post-searching.
- AutoDAN (Liu et al., 2024b): It proposes a hierarchical genetic algorithm that automatically generates jailbreak prompts to bypass safety features in aligned LLMs.
- ICA (Wei et al., 2023): It introduces the use of harmful in-context demonstrations to manipulate language models into generating harmful content by adding a few carefully crafted toxic responses to the prompt. In this paper, we use 1-shot, 3-shot and 5-shot versions of ICA.
- Virtual Context (Zhou et al., 2024): It leverages special tokens to enhance jailbreak attacks on LLMs by deceiving the models into perceiving user inputs as self-generated content.

5.4 Implementation Details

We evaluate the proposed GradingAttack framework using the CAS and ASR metrics. CAS is a newly introduced metric in this paper, designed to quantify the camouflage in an attack, with higher CAS values indicating greater camouflage and effectiveness. ASR measures the proportion of generated responses that successfully contain the targeted grading outcomes. To obtain student answers for the GAOKAO23, MATH, GSM8K and Math23K datasets, we employ 42 LLMs to generate responses to math word problems, treating each LLM as a virtual student. For all LLMs, to ensure reproducibility, the parameters max_length, temperature, top-k and top-p are set to 1024, 0, 1 and 0, respectively. The random seed is set to 42. For token-level methods, following previous work (Zou et al., 2023), the attack string length is set to 20 tokens and the maximum number of optimization iterations is set to 500. All experiments are carried out on a distributed cluster consisting of four nodes, each equipped with 8 NVIDIA A100 GPUs, accumulating a total computation time of at least 300 hours.

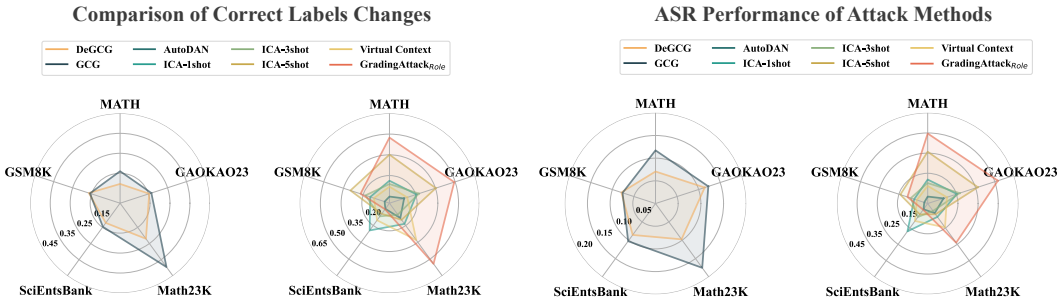

Figure 3: Performance comparisons of token-level (DeGCG, GCG) and prompt-level attack methods on five datasets.

Table 1: Camouflage attack performance of proposed framework on five datasets.

Method	Target Model	GAOKAO23		MATH		GSM8K		SciEntsBank		Math23K	
		CAS	ASR	CAS	ASR	CAS	ASR	CAS	ASR	CAS	ASR
GradingAttack$_{GCG}$	**Qwen2.5-7B**	0.3914	0.0155	0.5705	0.0318	0.5467	0.0292	0.7038	0.0484	0.8293	0.0672
	Qwen2.5-7B-Instruct	0.5959	0.1080	0.6858	0.0839	1.2431	0.1510	0.9703	0.0920	0.8933	0.2272
	Qwen2.5-14B-Instruct	0.6019	0.0354	0.4977	0.0242	0.6987	0.0477	0.8602	0.0723	0.6836	0.1382
	Llama-3.1-8B-Instruct	0.5681	0.1648	0.7021	0.1602	1.2756	0.1590	1.2245	0.1465	0.7228	0.3957
	Mistral-7B-Instruct	0.4737	0.3279	0.5105	0.3892	2.2819	0.5088	1.3640	0.2556	0.8411	0.5369
	DeepSeek-7B-Chat	0.4935	0.2875	0.5240	0.2902	1.7986	0.3161	1.1610	0.2377	1.7875	0.3122
	InternLM2.5-7B-Chat	0.4664	0.1997	0.4764	0.1546	1.4318	0.2003	0.7983	0.2148	0.6292	0.2483
Average		0.5130	0.1627	0.5667	0.1620	1.3252	0.2017	1.0117	0.1525	0.9124	0.2751
GradingAttack$_{Role}$	**Qwen2.5-7B**	0.1865	0.0034	0.4183	0.0215	0.2307	0.0052	0.5741	0.0322	0.7247	0.0805
	Qwen2.5-7B-Instruct	0.5055	0.6166	0.5774	0.8101	0.6796	0.5755	0.6079	0.7643	0.6235	0.5285
	Qwen2.5-14B-Instruct	0.7006	0.9894	0.6950	0.9890	0.6542	0.9813	0.6402	0.9991	0.6330	0.9873
	Llama-3.1-8B-Instruct	0.4859	0.4931	0.4803	0.4750	1.0246	0.2202	0.7455	0.0543	0.6132	0.5429
	Mistral-7B-Instruct	0.4743	0.3150	0.5210	0.3858	2.3771	0.5521	1.5450	0.3064	0.7763	0.6121
	DeepSeek-7B-Chat	0.4951	0.4018	0.5365	0.4534	2.1667	0.4587	0.9646	0.2756	1.4731	0.3865
	InternLM2.5-7B-Chat	0.4569	0.2641	0.4567	0.2752	1.4606	0.2365	0.7908	0.2570	0.5966	0.2944
Average		0.4721	0.4405	0.5265	0.4871	1.2276	0.4328	0.8383	0.3841	0.7772	0.4903

5.5 Experimental Results

We employ our GradingAttack framework to conduct extensive experiments and analyses to evaluate the security vulnerabilities of LLM based grading agents. Key insights, findings and suggestions are summarized in the following observations.

Observation 1. Token- and prompt-level adversarial attacks can effectively compromise LLM based grading agents to produce targeted grades, raising significant security concerns.

To compare the effectiveness of different attack methods in the grading scenario, we conduct experiments using several baselines. Figure 3 presents the performance of these methods across five datasets. The left subfigure shows the change in the proportion of originally incorrect labels that are reassigned as correct by the grading agent after attacks. A higher value indicates that more incorrect labels have been successfully manipulated to receive a correct grade. The right subfigure illustrates the ASR performance of the same attack methods under each dataset. From Figure 3, we observe the following: (1) Both token-level and prompt-level attack methods demonstrate strong attack capabilities across datasets. For token-level methods such as DeGCG and GCG, both the changes in correct labels (as shown in left subfigure) and the ASR performance (as shown in right subfigure) indicate their effectiveness in compromising LLM-based grading agents. For example, on the Math23K dataset, GCG achieves over 0.15 in correct label changes and over 0.35 in ASR performance. Similarly, prompt-level attack methods, including AutoDAN, ICA, Virtual Context and GradingAttack$_{Role}$, also achieve competitive performance in both metrics. On the GAOKAO23 dataset, for instance, GradingAttack$_{Role}$ achieves over 0.35 in correct label changes and around 0.50 in ASR performance. These results suggest that LLM based grading agents are vulnerable to both token-level and prompt-level adversarial attacks, highlighting critical security concerns for agents deployed in the wild. This vulnerability likely arises because these agents are not explicitly aligned to defend against grading-specific threats and general safety alignment alone is insufficient for robust performance in grading scenarios. (2) GCG and GradingAttack$_{Role}$ consistently outperform most

token-level and prompt-level baselines. For example, GCG exhibits a larger radar area compared to DeGCG, indicating stronger effectiveness in altering incorrect labels and achieving higher ASR performance. Similarly, GradingAttack$_{Role}$ shows a larger area than other prompt-level methods such as AutoDAN, ICA and Virtual Context, reflecting its overall superiority. Based on these preliminary comparisons, we select GCG and GradingAttack$_{Role}$ for in-depth analysis in subsequent experiments. Specifically, our framework is capable of adapting general-purpose adversarial algorithms to the grading agent context by redefining their objectives and constraints. Accordingly, we refer to the adapted version of GCG as GradingAttack$_{GCG}$ throughout the following sections.

OBSERVATION 2. DIFFERENT ATTACK METHODS DISPLAY DISTINCT CHARACTERISTICS: PROMPT-LEVEL ATTACKS ACHIEVE HIGHER SUCCESS RATES, WHEREAS TOKEN-LEVEL ATTACKS ENHANCE CAMOUFLAGE.

To further investigate the camouflage of attack methods, we select the token-level attack method, GradingAttack$_{GCG}$ and the prompt-level attack method, GradingAttack$_{Role}$, for analysis. Table 1 presents the performance of these attack methods across different grading models. From Table 1, we derive the following observations: (1) The prompt-level attack method GradingAttack$_{Role}$ achieves a higher ASR across all five datasets. For instance, GradingAttack$_{Role}$ achieves average ASRs of 0.4405, 0.4871, 0.4328, 0.3841 and 0.4903 on GAOKAO23, MATH, GSM8K, SciEntsBank and Math23K, respectively, outperforming GradingAttack$_{GCG}$, which achieves average ASRs of 0.1627, 0.1620, 0.2017, 0.1525 and 0.2751 on the same datasets. This suggests that prompt-level attack methods are more effective against grading models. The primary reason is that LLMs possess strong language comprehension capabilities, enabling them to better interpret and generate prompts aligned with attack objectives, thereby increasing the success rate. (2) The token-level attack method GradingAttack$_{GCG}$ exhibits greater camouflage. For example, GradingAttack$_{GCG}$ achieves average CAS values of 0.5130, 0.5667, 1.3252, 1.0117 and 0.9124 on GAOKAO23, MATH, GSM8K, SciEntsBank and Math23K, respectively, surpassing GradingAttack$_{Role}$, which achieves average CAS values of 0.4721, 0.5265, 1.2276, 0.8383 and 0.7772 on the same datasets. This indicates that token-level attack methods maintain better camouflage when attacking grading models, making them less likely to be detected through grading model accuracy checks. (3) Among token-level attacks, the Qwen2.5 series models demonstrate higher security but are more vulnerable to prompt-level attacks. For example, on GAOKAO23, under GradingAttack$_{GCG}$, Qwen2.5-7B, Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct achieve ASRs of 0.0155, 0.1080 and 0.0354, respectively, which are significantly lower than other models, such as Llama-3.1-8B-Instruct, which reaches 0.1648. This suggests that the Qwen2.5 series models exhibit stronger resistance to token-level attacks. However, under GradingAttack$_{Role}$, Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct achieve ASRs of 0.6166 and 0.9894, respectively, much higher than other models, such as Llama-3.1-8B-Instruct, which records 0.4931. This indicates that the Qwen2.5 instruct models are more susceptible to prompt-level attacks. The reason behind this vulnerability is that Qwen2.5 instruct models are more inclined to comply with role-play prompts, making them more exploitable by prompt-level attacks.

OBSERVATION 3. ATTACKING BOTH CORRECT AND INCORRECT LABELS YIELDS HIGHER SUCCESS RATES AND ENHANCED CAMOUFLAGE COMPARED TO TARGETING ONLY INCORRECT LABELS.

To further investigate how to achieve better camouflage, we conduct an experiment on the impact of attacking both correct and incorrect labels. The results are presented in Figure 4. In this experiment, "Incorrect => Correct" represents the strategy of attacking only incorrect labels, attempting to convert as many incorrect labels into correct ones as possible. In contrast, "Incorrect <=> Correct" represents the strategy of attacking both incorrect and correct labels, aiming to convert as many incorrect labels into correct ones and vice versa. From Figure 4, we observe that

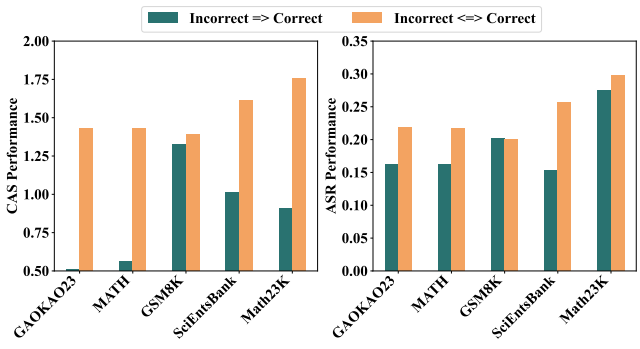

Figure 4: The impact of attack on different labels.

attacking both correct and incorrect labels leads to a higher CAS and ASR. For example, attacking both incorrect and correct labels yields a significantly higher CAS on the GAOKAO23 dataset and a substantially higher ASR on the SciEntsBank dataset compared to targeting only incorrect labels. This is because expanding the attack scope to include both incorrect and correct labels increases the number of label reversals, thereby enhancing the overall success rate while minimizing noticeable changes in the grading model's accuracy. As a result, this strategy enables more camouflaged and effective attacks.

OBSERVATION 4. THE PLACEMENT OF ROLE-PLAY STRINGS SIGNIFICANTLY INFLUENCES ATTACK EFFECTIVENESS: POSITIONING THEM BOTH AT THE BEGINNING AND END OF THE GRADING PROMPT, OR EXCLUSIVELY AT THE END, YIELDS BETTER ATTACK PERFORMANCE.

To further investigate the impact of role-play string placement on attack effectiveness in our GradingAttack$_{Role}$ method, we conduct experiments by varying the position of role-play strings in the adversarial prompt. The experimental results are presented in Figure 5, where R, S and P represent the role-play strings, student answer and grading prompt, respectively, with their order indicating the relative placement. From Figure 5, it is evident that the placement of role-play strings significantly influences attack performance. Positioning them at both the beginning and end of the grading prompt, or exclusively at the end, yields superior attack performance. For example, on the GAOKAO23 dataset, the R-P-R and P-R placements achieve ASRs of 0.4904 and 0.4931, respectively, which are substantially higher than other placements, such as R-S, R-S-R, S-R and R-P, which achieve ASRs of 0.1479, 0.2058, 0.2954 and 0.1893, respectively. These results suggest that the placement of role-play strings affects the model's interpretation of the grading prompt, thereby influencing its ability to recognize and respond to the role-play cues. When the role-play strings are positioned at both the beginning and end of the grading prompt, or exclusively at the end, the model is more likely to recognize them, resulting in a higher success rate. Therefore, we recommend placing the role-play strings at both the beginning and end of the grading prompt (R-P-R), or exclusively at the end (P-R), to maximize attack effectiveness.

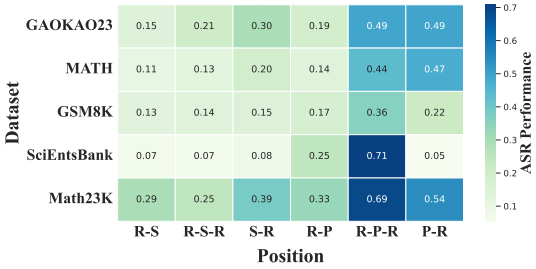

Figure 5: Effect of role-play string placement on performance. R, S, and P represent the role-play strings, student answer and grading prompt, respectively, with their order indicating the relative placement.

6 CONCLUSION

In this paper, we present GradingAttack, a fine-grained adversarial attack framework designed to systematically evaluate the security vulnerabilities of LLM based grading agents. To quantify attack camouflage, we propose a novel evaluation metric that balances attack success and stealth, providing a quantitative measure for agent security assessment. Extensive experiments on multiple educational datasets reveal that prompt-level attacks achieve higher success rates, whereas token-level attacks provide stronger camouflage. Our work contributes to the broader understanding of agent security and robustness, emphasizing that deploying trustworthy agents in critical applications requires careful consideration of potential adversarial threats.

ETHICS STATEMENT

This research investigates the security vulnerabilities of LLM based educational grading agents through adversarial attacks. Our goal is to highlight potential security risks in deploying AI agents in educational environments and improve the robustness of these agents, ensuring fairness and reliability in educational assessment. As AI agents are increasingly deployed "in the wild" for critical applications like grading, understanding their vulnerabilities is essential for building trustworthy agent systems. We acknowledge that the proposed attack methods could be misused to manipulate grading outcomes. To mitigate this risk, we adhere to responsible disclosure practices and will notify relevant educational technology providers about our findings before public release.

REFERENCES

Sridevi Bonthu, S. Rama Sree, and M. H. M. Krishna Prasad. Automated short answer grading using deep learning: A survey. In *Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, Virtual Event, August 2021.

Steven Burrows, Iryna Gurevych, and Benno Stein. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25:60–117, 2015.

Li-Hsin Chang and Filip Ginter. Automatic short answer grading for finnish with chatgpt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, Canada, March 2024.

Impey Chris, Wenger Matthew, Garuda Nikhil, Golchin Shahriar, and Stamer Sarah. Using large language models for automated grading of student writing about science. *International Journal of Artificial Intelligence in Education*, pp. 1–35, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, pp. 1–34, 2025.

Yuning Ding, Brian Riordan, Andrea Horbach, Aoife Cahill, and Torsten Zesch. Don't take "nswvt-nvakgxpm" for an answer–the surprising vulnerability of automatic content scoring systems to adversarial input. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain, December 2020.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Proceedings of the 7th International Workshop on Semantic Evaluation*, Atlanta, Georgia, USA, June 2013.

Anna Filighera, Tim Steuer, and Christoph Rensing. Fooling automatic short answer grading systems. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education*, Ifrane, Morocco, July 2020.

Anna Filighera, Sebastian Ochs, Tim Steuer, and Thomas Tregel. Cheating automatic short answer grading with the adversarial usage of adjectives and adverbs. *International Journal of Artificial Intelligence in Education*, 34:616–646, 2024.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Proceedings of 34th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, Virtual Event, December 2021.

Helen A Klein, Nancy M Levenburg, Marie McKendall, and William Mothersell. Cheating during the college years: How do business school students compare? *Journal of Business Ethics*, 72: 197–206, 2007.

Ronja Laarmann-Quante, Christopher Chandler, Noemi Incirkus, Vitaliia Ruban, Alona Solopov, and Luca Steen. A multilingual dataset of adversarial attacks to automatic content scoring systems. In *Proceedings of the 20th Conference on Natural Language Processing*, Vienna, Austria, September 2024.

Yihuai Lan, Lei Wang, Qiyuan Zhang, Yunshi Lan, Bing Tian Dai, Yan Wang, Dongxiang Zhang, and Ee-Peng Lim. Mwptoolkit: An open-source framework for deep learning-based math word problem solvers. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, Virtual Event, February 2022.

Hongfu Liu, Yuxi Xie, Ye Wang, and Michael Shieh. Advancing adversarial suffix transfer learning on aligned large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November 2024a.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, May 2024b.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*, 2023.

Yi Liu, Gelei Jia, Yizhe Geng, Jingyi Jia, Zhi Liang Chen, Tianwei Gu, Yang Liu, and Yuekang Li. Automatic and universal prompt injection attacks against large language models. *arXiv preprint arXiv:2403.04957*, 2024c.

Manas Mhasakar, Shikhar Sharma, Apurv Mehra, Utkarsh Venaik, Ujjwal Singhal, Dhruv Kumar, and Kashish Mittal. Comuniqa : Exploring large language models for improving speaking skills. In *Proceedings of the 7th ACM SIGCAS/SIGCHI Conference of Computing and Sustainable Societies*, New Delhi, India, July 2024.

Haile Misgna, Byung-Won On, Ingyu Lee, and Gyu Sang Choi. A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, 58:1–40, 2025.

Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the 11th ACM Conference on Learning @ Scale*, New York, NY, USA, July 2024.

Marko Putnikovic and Jelena Jovanovic. Embeddings for automatic short answer grading: A scoping review. *IEEE Transactions on Learning Technologies*, 16:219–231, 2023.

Mohi Reza, Nathan M Laundry, Ilya Musabirov, Peter Dushniku, Zhi Yuan Michael Yu, Kashish Mittal, Tovi Grossman, Michael Liut, Anastasia Kuzminykh, and Joseph Jay Williams. Abscribe: Rapid exploration & organization of multiple writing variations in human-ai co-writing tasks using large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, May 2024.

Sebastian Speiser and Annegret Weng. Enhancing short answer grading with openai apis. In *Proceedings of the 21st International Conference on Information Technology Based Higher Education and Training*, Paris, France, November 2024.

Pedro Antonio Gutiérrez Víctor Manuel Vargas and César Hervás-Martínez. Unimodal regularisation based on beta distribution for deep ordinal regression. *Pattern Recognition*, 122:1–10, February 2022.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18:1–26, 2024.

Zeming Wei, Yifei Wang, and Yisen Wang. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2023.

Bernard E Whitley. Factors associated with cheating among college students: A review. *Research in Higher Education*, 39:235–274, 1998.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68:1–44, 2025.

Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. An llm can fool itself: A prompt-based adversarial attack. *arXiv preprint arXiv:2310.13345*, 2023.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*, 2023.

Yihao Zhang and Zeming Wei. Boosting jailbreak attack with momentum. In *Proceedings of the ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, Vienna, Austria, May 2024.

Li Zhong and Zilong Wang. Can llm replace stack overflow? a study on robustness and reliability of large language model code generation. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada, February 2024.

Yuqi Zhou, Lin Lu, Ryan Sun, Pan Zhou, and Lichao Sun. Virtual context enhancing jailbreak attacks with special token injection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA, November 2024.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.