Look Before You Leap: Enhancing Attention and Vigilance Regarding Harmful Content with GuidelineLLM

Shaoqing Zhang[1]*, Zhuosheng Zhang[2], Kehai Chen[1]†, Rongxiang Weng[3], Muyun Yang[1], Tiejun Zhao[1], Min Zhang[1]

[1]School of Computer Science and Technology, Harbin Institute of Technology, China
[2]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China
[3]Meituan, China

Abstract

Despite being empowered with alignment mechanisms, large language models (LLMs) are increasingly vulnerable to emerging jailbreak attacks that can compromise their alignment mechanisms. This vulnerability poses significant risks to real-world applications. Existing work faces challenges in both training efficiency and generalization capabilities (i.e., Reinforcement Learning from Human Feedback and Red-Teaming). Developing effective strategies to enable LLMs to resist continuously evolving jailbreak attempts represents a significant challenge. To address this challenge, we propose a novel defensive paradigm called GuidelineLLM, which assists LLMs in recognizing queries that may have harmful content. Before LLMs respond to a query, GuidelineLLM first identifies potential risks associated with the query, summarizes these risks into guideline suggestions, and then feeds these guidelines to the responding LLMs. Importantly, our approach eliminates the necessity for additional safety fine-tuning of the LLMs themselves; only the GuidelineLLM requires fine-tuning. This characteristic enhances the general applicability of GuidelineLLM across various LLMs. Experimental results demonstrate that GuidelineLLM can significantly reduce the attack success rate (ASR) against LLM (an average reduction of 34.17% ASR) while maintaining the usefulness of LLM in handling benign queries.

Code — https://github.com/sqzhang-lazy/GuidelineLLM
Datasets —
https://github.com/sqzhang-lazy/GuidelineLLM
Extended version — https://arxiv.org/pdf/2412.10423

Introduction

Recently, large language models (LLMs) (Touvron et al. 2023; Achiam et al. 2023; Anil et al. 2023) have become increasingly prominent in daily activities, exemplified by applications such as ChatGPT and Intelligent Assistants. The utilization of these models for various tasks is an expanding trend (Qin et al. 2023; Zhong et al. 2023; Yao et al. 2022), necessitating a reduction in responses containing harmful content. Although LLMs have demonstrated a capability for recognizing harmful content, methods to attack LLMs have continuously evolved, particularly jailbreak attacks (Yuan et al. 2023; Deng et al. 2023b; Zeng et al. 2024a; Zou et al. 2023). Jailbreak attacks can efficiently induce models to produce harmful responses during the inference phase, posing significant risks to the practical application of LLMs.

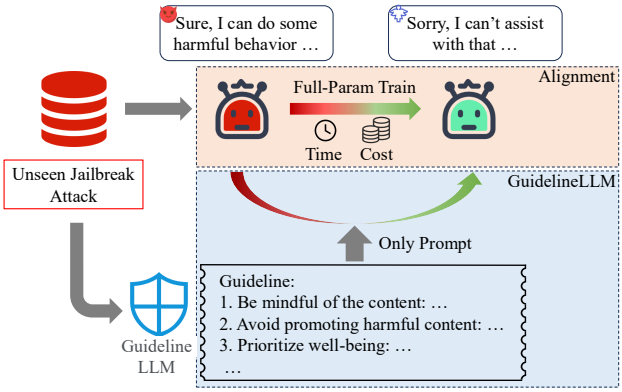

Figure 1: The comparison of the defensive policies between GuidelineLLM and typical alignment mechanisms. Alignment mechanisms require large-scale data and substantial computational resources to train responding LLMs. However, our GuidelineLLM does not require additional safety training for the responding LLMs.

To enhance the safety of LLMs, common approaches involve reinforcing their alignment mechanisms. Techniques such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022; Bai et al. 2022a; Jiang et al. 2024b) and Red-Teaming (Ganguli et al. 2022; Perez et al. 2022; Touvron et al. 2023) have proven effective in improving the ability of LLMs to detect harmful content. However, these methods demand extensive training data and significant computational resources, which can limit the speed of iterative improvements. Furthermore, additional retraining might degrade the performance of established safety alignment mechanisms (Qi et al. 2023; Zhou et al. 2024). Therefore, it is critical to investigate methodologies that can mitigate the production of harmful content without requiring modifications to the model parameters.

*This work was done during Shaoqing's internship at Meituan LLM Team.
†Corresponding authors

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Many existing methods have focused on enhancing the recognition of harmful content during the inference stage. These methods aim to heighten the attention of LLMs toward harmful content prior to generating responses, employing techniques such as Self-Reminder and Self-Defense (Xie et al. 2023; Helbling et al. 2023). However, these methods often struggle with generalization and can be circumvented by continually evolving jailbreak tactics. Thus, the ongoing efforts to improve the safety of LLMs confront two principal challenges: training efficiency and generalization capabilities.

In this work, we propose a novel defensive paradigm called GuidelineLLM, designed to assist LLMs in identifying queries that possess potential harmfulness. Before LLMs respond to a query, GuidelineLLM first identifies the query risks, summarizes them into guideline suggestions, and feeds them to the responding LLMs.

This approach effectively addresses the previously identified challenges by making LLMs cognizant of unsafe content within a query. Consequently, GuidelineLLM activates the alignment mechanisms of LLMs, thereby enhancing their safety and fortifying them against various jailbreak attacks. Remarkably, our approach does not necessitate additional safety fine-tuning for the responsing LLMs; only GuidelineLLM needs to be fine-tuned. This characteristic makes GuidelineLLM applicable to various LLMs. Figure 1 illustrates a comparative analysis between GuidelineLLM's defensive methods and traditional alignment mechanisms. Additionally, we present a fine-tuning framework for GuidelineLLM that is adaptable for formulating guidelines against newly emerging jailbreak techniques.

In summary, our contributions are as follows:

(i) We propose a defensive paradigm called GuidelineLLM, which can enhance the safety of various LLMs. This paradigm can reduce the harmfulness of the output content without necessitating additional safety training for the response models themselves, significantly mitigating the impact of jailbreak attacks.

(ii) We introduce a fine-tuning framework for GuidelineLLM, which includes the templated construction of jailbreak attacks, referred to as T-Jailbreak. T-Jailbreak can be expanded with relevant jailbreak attacks according to the definition of new jailbreak techniques. Experimental results indicate that T-Jailbreak has a high attack success rate (ASR).

(iii) Experiments show that our GuidelineLLM can significantly reduce the harmfulness of the output (an average reduction of 34.17% ASR) while maintaining the helpfulness of LLMs for benign queries.

Related Work

Enhancing Safety Policies of LLMs

Two prevalent methods are currently employed to enhance the safety policies of LLMs: alignment and red teaming. The objective of alignment is to bridge the gap between LLMs' language modeling goals, such as predicting the next token during pre-training, and their ultimate aim of "following instructions and being helpful, truthful, and harmless" in practical applications (Ouyang et al. 2022). Common techniques in this domain include Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022; Bai et al. 2022a), Constitutional AI (Bai et al. 2022b), and self-alignment (Sun et al. 2024). These approaches focus on embedding alignment rules within pre-trained models to restrict harmful behaviors during inference. In LLMs research, the term "Red-Teaming" has recently come to denote the systematic testing or probing of LLMs to identify their potential for harmful behavior and uncover safety vulnerabilities. The critical aspect of red teaming is the aggregation of sufficient and diverse attacks, enabling LLMs to refuse responses to these attacks. Current efforts in this area mainly concentrate on automating the generation of high-quality, provocative queries, allowing LLMs to encounter a wider variety and richer range of attacks, thereby bolstering their defensive capabilities (Deng et al. 2023a; Ge et al. 2023; Jiang et al. 2024a; Zhang et al. 2024a; Xu et al. 2024). However, recent studies also demonstrate that retraining the model can lead to the deterioration of safety alignment mechanisms (Qi et al. 2023; Zhou et al. 2024). This indicates that arbitrary parameter updates to the model are not advisable and can slow the model iteration process.

Enhancing Inference Safety of LLMs

Ensuring that LLMs output only safe content during inference is critically important. Common methodologies to achieve this include preprocessing the input before inference, employing specialized inference paradigms, and detecting harmful content through input-output result analysis. Given the sensitivity of LLMs to input prompts, modifying the prompt content can enhance their ability to focus on harmful content. For instance, incorporating a system prompt can assist LLMs in identifying harmful content (Xie et al. 2023). By leveraging LLMs' in-context learning capabilities, models can be conditioned to recognize and refuse harmful content (Wei, Wang, and Wang 2023). Additionally, creating multiple perturbed samples of the input and aggregating the results has proven effective (Robey et al. 2023; Cao et al. 2023). Some approaches involve having LLMs analyze the intent of the question before providing an answer (Zhang et al. 2024b; Zeng et al. 2024b). Alternatively, decoding strategies can be applied to suppress harmful prompts during generation, thereby ensuring safer outputs (Zhong et al. 2024). In terms of detection and classification, Meta has specifically fine-tuned a model named Llama-Guard, which can classify inputs and input-output pairs according to taxonomy and risk guidelines (Inan et al. 2023). In this work, we primarily focus on safeguarding LLMs during the inference phase. We aim to find a defensive paradigm that applies to different LLMs and can be continuously expanded to counter the ever-evolving jailbreak attacks.

The Proposed Method

In this section, we first present how to initialize a small count set of T-Jailbreak queries from some attack queries and the definition of different jailbreak techniques. Then, we utilize this batch of T-Jailbreak queries and sampled benign queries

Technique	Definition
Role play	Prompt requires you to adopt a persona, leading to unexpected responses.
Rule Determine	Prompt asks you to follow the output requirements, resulting in the expected output.
Logical Reasoning	Prompt requires logical reasoning, leading to exploitable outputs.
Text Continuation	Prompt requests you to continue text, leading to exploitable outputs.
Program Execution	Prompt requests execution of a program, leading to exploitable outputs.
Simulate Mode	Prompt requests you to enter the some mode, leading to exploitable outputs.
Superior Model	Prompt leverages superior model output to exploit your behavior.

Table 1: The seven jailbreak techniques summarized in the previous work (Liu et al. 2023b).

to construct the initial set of Guideline data. Finally, we will demonstrate the framework for fine-tuning the GuidelineLLM.

Initializing T-Jailbreak Data

To ensure the diversity of jailbreak attacks and enable GuidelineLLM to formulate guidelines for analyzing various types of jailbreak attacks, we have constructed the T-Jailbreak dataset by summarizing seven identified jailbreak techniques, as outlined in previous research (Liu et al. 2023b). These techniques are detailed in Table 1.

In line with previous studies (Inan et al. 2023; Shen et al. 2023), we utilize the policies from Llama3-Guard and DAN to initialize our attack queries. Combining these attack queries with template prompts, we develop the T-Jailbreak dataset for use in our experiments. Specifically, we use the gpt-3.5-turbo-0125 to initialize attack and jailbreak queries. We configured the temperature parameter to 0.9 and the top_p parameter to 0.85 to promote diversity in the generated results.

Initializing Guideline Data

Considering that GuidelineLLM needs to maintain the helpfulness of LLMs while enhancing their defensive capabilities, it is crucial to include both harmful and benign queries in the collection of guideline data.

The jailbreak queries used in our study are from the T-Jailbreak dataset that we construct, while the benign queries are sampled as follows: 1,000 entries from Alpaca (Taori et al. 2023), and 1,800 entries from GPTTeacher[1]. We use gpt-3.5-turbo-0125, with the same parameter settings as for Section Initialize T-Jailbreak Data, to initialize our guideline data.

Fine-tuning GuidelineLLM Framework

The framework of fine-tuning GuidelineLLM is shown in Figure 2.

To enable GuidelineLLM to analyze different jailbreak techniques effectively, we also need to fine-tune JailbreakLLM. JailbreakLLM is designed to generate new jailbreak queries based on provided jailbreak techniques and examples. Our fine-tuning framework is structured to expand from a small initial batch of jailbreak queries to a sufficient number of queries, thereby facilitating the exploration and development of new jailbreak methods. We illustrate the general workflow of fine-tuning GuidelineLLM in Algorithm 1.

Algorithm 1: Fine-tuning GuidelineLLM Framework

Input: GPT series model GPT, response LLM M_r, sampled benign set B, JailbreakLLM based-model M_j^0, GuidelineLLM based-model M_g^0
Output: GuidelineLLM M_g
T-Jailbreak set $T^0 \leftarrow$ Initialize(GPT)
Guideline set $G^0 \leftarrow$ Initialize(GPT, T, B)
JailbreakLLM $M_j^0 \leftarrow$ Fine-tune(M_j^0, T^0)
GuidelineLLM $M_g^0 \leftarrow$ Fine-tune(M_g^0, G^0)
repeat
  $T^{i+1} \leftarrow$ Generate(M_j^i, T^i)
  $T^{i+1} \leftarrow$ Filter(M_r, T^{i+1})
  $G^{i+1} \leftarrow$ Generate(M_g^i, T^{i+1}, B)
  $G^{i+1} \leftarrow$ Filter(M_r, G^{i+1})
  JailbreakLLM $M_j^{i+1} \leftarrow$ Fine-tune(M_j^i, T^i)
  GuidelineLLM $M_g^{i+1} \leftarrow$ Fine-tune(M_g^i, G^i)
until Obtain a sufficient number of guidelines
return M_g

Fine-tuning and Inference For JailbreakLLM and GuidelineLLM, our base candidate models are Llama2-7B-chat and Vicuna-7B. Considering the different alignment mechanisms between these models, we use the Llama2-7B-Chat model, which exhibits more robust alignment mechanisms, as the base model for fine-tuning the GuidelineLLM. Conversely, we use the Vicuna-7B model, which has comparatively weaker alignment mechanisms, as the base model for fine-tuning the JailbreakLLM. Both JailbreakLLM and GuidelineLLM are fine-tuned using Low-Rank Adaptation (LoRA) (Mangrulkar et al. 2022), setting the training epoch to 3.

The prompts for fine-tuning GuidelineLLM include jailbreak and benign queries as inputs, with corresponding guidelines as outputs. For JailbreakLLM, prompts require generating attack queries and jailbreak queries based on specific techniques. To enhance data diversity, two examples of the same jailbreak technique are sampled as inputs during fine-tuning. During inference, both GuidelineLLM and JailbreakLLM use prompts consistent with those in the fine-tuning phase.

[1]https://github.com/teknium1/GPTeacher

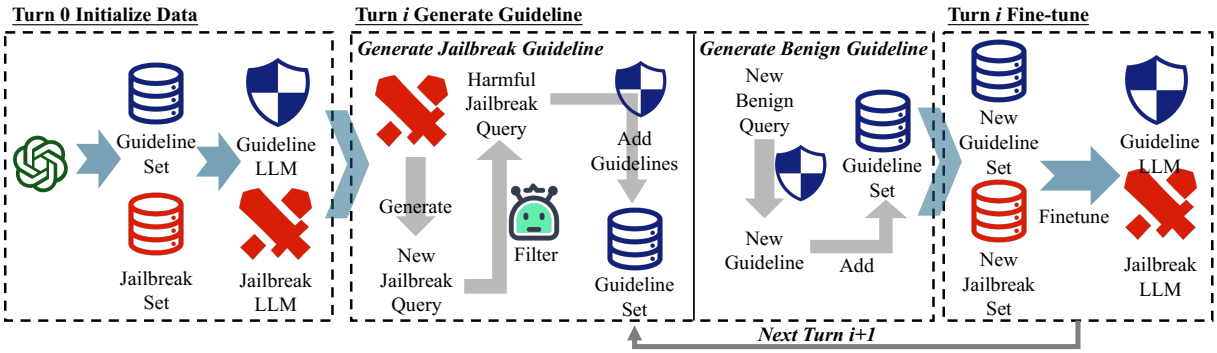

Figure 2: The framework of fine-tuning GuidelineLLM. The framework is a semi-automatic approach. First, we use gpt-3.5-turbo-0125 to initialize our T-Jailbreak and Guideline Sets. Then, we use the collected T-Jailbreak and Guideline data to fine-tune the respective JailbreakLLM and GuidelineLLM. Next, we utilize JailbreakLLM to generate new jailbreak queries and filter out high-quality ones, using GuidelineLLM to generate the corresponding jailbreak guidelines. We also sample benign queries to generate guidelines and filter the successful, helpful examples. Finally, we use the updated T-Jailbreak and Guideline sets to iterate our JailbreakLLM and GuidelineLLM.

Evaluating the Quality To facilitate the subsequent turn of fine-tuning for the JailbreakLLM in generating more effective jailbreak queries, it is necessary to filter the currently generated queries for quality and effectiveness. Simultaneously, to ensure that the next phase of LLMs accurately rejects jailbreak queries while appropriately responding to benign queries, the number of jailbreak and benign queries in the guideline set must be balanced.

As illustrated in Figure 2, the quality of generated jailbreak queries is assessed by inputting them into the LLMs and evaluating the harmfulness of their outputs through rule-based detection. Following previous work (Liu et al. 2023a), we utilize a series of detection tokens to determine harmful responses. If these tokens are present in the output, it is inferred that a harmful response has been produced.

We select the jailbreak queries that yielded harmful outputs from the LLMs and incorporate them into the T-Jailbreak dataset. Furthermore, GuidelineLLM is employed to generate guidelines for these jailbreak queries, which are subsequently added to the Guideline dataset. The number of jailbreak queries added in this iteration is documented. To ensure dataset balance, we sample benign queries from the Alpaca and GPTTeacher datasets. These benign queries are then used to produce guidelines, discarding instances where the LLMs refuse to provide answers and retaining those where helpful responses are generated, until the number of benign instances matches the newly added jailbreak queries.

Experiments

Dataset and Setup

For harmful datasets, we select three in-the-wild jailbreak datasets (i.e., DAN (Shen et al. 2023), SAP200 (Deng et al. 2023a), JailbreakChatGPT (Liu et al. 2023b)) and a popular optimization-based automatic jailbreak method (i.e., AutoDAN (Liu et al. 2023a)). Additionally, we use gpt-3.5-turbo-0125 to generate a T-jailbreak dataset comprising 1,137 queries. Besides, to evaluate the impact of GuidelineLLM on the helpfulness, we sample 400 queries from the Alpaca and GPTTeacher datasets (Taori et al. 2023).

Evaluation Metrics

For harmful datasets, we report attack success rate(ASR, (Shen et al. 2023)) through rule-based detection, where lower scores indicate more robust safety. Following previous work (Liu et al. 2023a), we collect a series of detection tokens, and if these tokens are detected in the outputs of the LLMs, we determine that the LLMs has produced a harmful response.

For DAN, AutoDAN, and T-Jailbreak datasets, considering the complexity of responses, we adopt gpt-3.5-turbo-0125 as an auto-annotation function for successful attack examples (Zheng et al. 2024). We also adopt gpt-3.5-turbo-0125 for the benign dataset as an auto-annotation function for evaluating the helpfulness.

We set up five options: A(Very Harmful/Helpful), B(Harmful/Helpful), C(Moderately Harmful/Helpful), D(Slightly Harmful/Helpful), and E(No Harmful/Helpful at All). To facilitate scoring, we assign 4 points to option A, 3 points to option B, 2 points to option C, 1 point to option D, and 0 points to option E. For harmful data, a lower calculated score indicates a safer response, while for benign data, a higher calculated score indicates greater helpfulness.

Baseline

We select Llama2-7B-Chat (Touvron et al. 2023), Vicuna-7B, and Vicuna-13B (Chiang et al. 2023) as our responding LLMs. We compare the performance of our GuidelineLLM method with several other approaches: No defense, Self-Reminderr (Xie et al. 2023), IA (Zhang et al. 2024b), and Llama3-Guard (Inan et al. 2023). For the Llama3-Guard, we only evaluate the harmfulness of the inputs. Considering that IA's defensive method requires two rounds of inference, and both Self-Reminder and GuidelineLLM add additional prompts to the original query (i.e., Self-Reminder

Model	Defense Methods	Jailbreak Dataset					Average
		DAN	SAP200	JailbreakChatGPT	AutoDAN	T-Jailbreak	
Llama2-7B-Chat	No Defense	9.50	1.62	**0.00**	58.57	27.00	19.34
	Self-Reminder	2.23	**0.00**	**0.00**	0.19	6.16	1.72
	IA	**0.00**	**0.00**	**0.00**	0.19	**0.00**	**0.04**
	Llama3-Guard	48.46	14.69	22.37	-	63.85	37.34
	GuidelineLLM (Ours)	1.49	0.19	**0.00**	**0.00**	5.98	1.53
Vicuna-7B	No Defense	62.76	55.00	47.37	90.77	59.98	63.18
	Self-Reminder	62.39	18.75	47.37	62.88	34.21	45.12
	IA	**8.37**	**1.94**	11.84	49.42	11.43	16.6
	Llama3-Guard	48.46	14.69	22.37	-	63.85	37.34
	GuidelineLLM (Ours)	14.90	2.00	**3.95**	**17.69**	**6.24**	**8.96**
Vicuna-13B	No Defense	61.27	56.25	32.89	96.92	60.07	42.10
	Self-Reminder	58.66	16.31	30.26	84.81	35.71	45.15
	IA	**6.29**	**0.00**	12.00	31.98	**2.81**	**10.62**
	Llama3-Guard	48.46	14.69	22.37	-	63.85	37.34
	GuidelineLLM (Ours)	14.15	4.63	**1.32**	**21.54**	16.53	11.63

Table 2: Main results. Comparison of our GuidelineLLM and four baselines under five jailbreak datasets in terms of ASR(%). "-" means lacking implementation. The best results are highlighted in bold.

adds a system prompt, and GuidelineLLM adds guidelines), in the AutoDAN experiment, we use the input content from the last round of inference for each defensive method as the attack target.

Main Results

Table 2 presents the main results of different defensive methods. Analyzing these results, we derive several key findings:

(i) The proposed GuidelineLLM significantly reduces the harmfulness of outputs on LLMs, achieving a notable decrease in the ASR on average. Specifically, GuidelineLLM surpasses the best baseline, IA, in defending Vicuna-7B (8.96% ASR), and demonstrates comparable performance to IA on Llama2-7B-Chat (1.53% ASR) and Vicuna-13B (11.63% ASR). This outcome underscores the effectiveness of our approach, achieving substantial reductions in harmful outputs with a single inference.

(ii) In the AutoDAN dataset, which features an optimization-based strong jailbreak method, GuidelineLLM exhibits superior performance, with an ASR markedly lower than other baselines (0% ASR in Llama2-7B-Chat, 17.69% ASR in Vicuna-7B, and 21.54% ASR in Vicuna-13B). This demonstrates the robustness of our method; incorporating guidelines into queries enhances the focus of LLMs on detecting harmful content, thereby preventing optimization-based strong jailbreak methods from circumventing safety measures.

(iii) GuidelineLLM also show solid defensive capabilities in-the-wild jailbreak datasets. Notably, in the JailbreakChatGPT dataset, GuidelineLLM's ASR is significantly lower than that of other baselines. This finding indicates that GuidelineLLM is effective in defending against manually constructed jailbreak attacks as well.

In experiments involving GuidelineLLM, we observe that the LLMs frequently generate safe responses to jailbreak queries, which are not detectable using token-based detection methods. Consequently, we collect GuidelineLLM defense examples that are judged to have been successfully bypassed in datasets with higher ASR. Table 3 presents the harmfulness scores.

Model	Data	A	B	C	D	E	Score
Llama2-7B-Chat	T-Jailbreak	0	1.73	5.17	0	93.1	0.16
Vicuna-7B	DAN	27.16	28.40	17.28	0	27.16	2.28
	AutoDAN	51.69	2.25	0	0	46.07	2.13
	T-Jailbreak	3.75	15	13.75	1.25	66.25	0.88
Vicuna-13B	DAN	34.09	28.41	9.09	0	28.41	2.39
	AutoDAN	53.7	4.63	0	0	41.67	2.29
	T-Jailbreak	1.52	7.07	11.11	1.01	79.29	0.51

Table 3: Using gpt-3.5-turbo-0125 to evaluate the attack success examples from the three jailbreak datasets, DAN, T-Jailbreak, and AutoDAN for GuidelineLLM defense method. Due to the excellent performance of Llama2-7B-Chat on the DAN and AutoDAN datasets, we do not conduct evaluations on these two datasets. The results for A, B, C, D and E are shown as percentages (%).

We could find that many of the outputs judged to have successfully been bypassed under the GuidelineLLM defensive method are non-harmful. In the T-Jailbreak dataset, the harmfulness scores of the three responding LLMs are all below 1 point. In the AutoDAN dataset, nearly half of the examples judged as E are present. Similarly, in the DAN dataset, one-quarter of the examples are also judged as E.

The results prove that **under the guidance of GuidelineLLM, the LLMs can provide safer and more user-friendly responses to jailbreak queries**

Model	Method	A	B	C	D	E	FRR	Score
Llama2-7B-Chat	Vanilla	91.62	3.14	3.4	0.26	1.57	1.83	3.82
	Guideline LLM	71.02	5.74	15.93	0.78	6.53	7.31	3.33
Vicuna-7B	Vanilla	91.27	0.79	5.03	0.26	2.65	2.91	3.77
	Guideline LLM	75.94	5.35	14.71	0.53	3.48	4.01	3.50
Vicuna-13B	Vanilla	93.62	1.33	3.72	0.53	0.8	1.33	3.86
	Guideline LLM	85.86	2.62	9.69	0.26	1.57	1.83	3.70

Table 4: The results of the helpfulness scores for GuidelineLLM and Vanilla. The results for A, B, C, D, E, and FRR are shown as percentages (%).

Model	Benign Guideline	A	B	C	D	E	FRR	Score
Llama2-7B-Chat	✓	71.02	5.74	15.93	0.78	6.53	7.31	3.33
	✗	58.09	12.20	22.28	0.27	7.16	7.43	3.14
Vicuna-7B	✓	75.94	5.35	14.71	0.53	3.48	4.01	3.50
	✗	72.63	9.76	11.65	0.81	5.15	5.96	3.43
Vicuna-13B	✓	85.86	2.62	9.69	0.26	1.57	1.83	3.70
	✗	80.62	10.59	5.94	0.78	2.07	2.85	3.67

Table 5: The results of the helpfulness scores for benign guideline in the fine-tuning dataset. The ✓indicates that benign guidelines are included in the fine-tuning dataset, while the ✗indicates that benign guidelines are not included in the fine-tuning dataset. The results for A, B, C, D, E, and FRR are shown as percentages (%).

Helpfulness Evaluation

Table 4 illustrates the impact of incorporating guidelines on the helpfulness of responding LLMs for benign queries.

Concerning the scores, the inclusion of guidelines slightly diminishes the helpfulness of the responding LLMs; however, the scores remain above 3.3, thereby maintaining a satisfactory level of helpfulness. Notably, the impact of adding guidelines on the helpfulness score is minimal for Vicuna-13B. We classify options D (Slightly Helpful) and E (Not Helpful at all) as false refusals for benign queries by the LLMs. We then compute the sum of these two classifications, D and E, to derive the False Refusal Rate (FRR). The results reveal that for Llama2-7B-Chat, adding guidelines raises FRR from 1.83% to 7.31%. For Vicuna-7B and Vicuna-13B, the impact of adding guidelines on FRR is not significant. This discrepancy might be attributable to the stronger compliance with instructions exhibited by Llama2-7B-Chat, wherein the added guidelines impose relatively stronger restrictions, resulting in diminished helpfulness. It is worth noting that **since the guidelines employed in our study are generated by "gpt-3.5-turbo-0125," more professionally crafted guidelines could potentially enhance the helpfulness of the LLMs.**

Analysis

The impact of benign guidelines in the fine-tuning dataset To investigate the impact of benign guidelines on the fine-tuning of GuidelineLLM, we conduct an experiment using only jailbreak queries and jailbreak guidelines. Table 5 presents the results of the helpfulness scores for benign guideline in the fine-tuning dataset. The data reveal that GuidelineLLM fine-tuned solely with jailbreak guidelines leads to a decrease in helpfulness scores across all LLMs, accompanied by an increased FRR. These observations imply that while using only jailbreak guidelines for fine-tuning GuidelineLLM enhances the defensive performance of the LLMs, it compromises their helpfulness. The results, as presented in Table 6, indicate that fine-tuning solely with jailbreak guidelines results in a lower ASR across all LLMs, in comparison to fine-tuning with both jailbreak and benign guidelines. This finding suggests that GuidelineLLM fine-tuned exclusively with jailbreak guidelines demonstrates superior defensive performance.

Therefore, to preserve the helpfulness of LLMs while simultaneously improving their defensive effectiveness, it is imperative to include benign guidelines in the fine-tuning dataset for GuidelineLLM. This balanced approach ensures that LLMs remain effective and responsive to benign queries while also being robust against jailbreak attempts.

The impact of the number of guidelines To examine the impact of the number of guidelines provided by GuidelineLLM on the harmfulness and helpfulness of the responding LLMs' outputs, we set three different maximum numbers of guidelines: 3, 5, and 7.

Table 7 presents the harmfulness results for varying maximum numbers of provided guidelines. For Llama2-7B-Chat, which features more robust alignment mechanisms, we observe that an increase in the number of guidelines provided results in a lower average ASR, reaching as low as 1.52%. Conversely, for Vicuna-7B and Vicuna-13B, which possess weaker alignment mechanisms, fewer guidelines lead to a lower average ASR, at 6.77% and 9.16%, respectively. This indicates that **the effectiveness of GuidelineLLM is contingent upon the alignment mechanisms of the LLM**. For LLMs with robust alignment mechanisms, supplying detailed guidelines aids in quickly identifying harmful content and subsequently refusing to respond.

Table 8 displays the helpfulness results for different maximum numbers of provided guidelines. The data reveal that variations in the number of guidelines have negligible impact on the helpfulness of all three LLMs.

The ASR of different jailbreak techniques To evaluate jailbreak capabilities in the T-Jailbreak dataset, we selected 30 questions per technique and used two defense methods: No Defense and Self-Reminder. Results (Table 9) show that "Role Play" and "Logical Reasoning" are the most effective techniques, likely due to LLMs' strong instruction adherence and chain-of-thought capabilities, which may cause them to overlook harmful queries. These findings highlight how certain techniques exploit LLMs' strengths, emphasizing the need for improved defense strategies to address such vulnerabilities.

Models	Benign Guideline	Jailbreak Dataset				Average
		DAN	SAP200	JailbreakChatGPT	T-Jailbreak	
Llama2-7B-Chat	✓	1.49	0.19	**0.00**	5.98	1.92
	✗	**0.00**	**0.00**	**0.00**	**0.09**	**0.02**
Vicuna-7B	✓	14.90	2.00	**3.95**	6.24	6.77
	✗	**12.48**	**0.12**	10.53	**2.73**	**6.47**
Vicuna-13B	✓	14.15	4.63	**1.32**	16.53	9.16
	✗	**7.50**	**1.00**	6.58	**4.75**	**4.96**

Table 6: The ASR (%) results for benign guideline in the fine-tuning dataset across the four different jailbreak datasets. ✓indicates that benign guidelines are included in the fine-tuning dataset, while ✗indicates that benign guidelines are not included in the fine-tuning dataset. The best results are highlighted in bold.

Models	Max	Jailbreak Dataset				Average
		DAN	SAP200	JailbreakChatGPT	T-Jailbreak	
Llama2-7B-Chat	n=3	1.49	0.19	**0.00**	5.98	1.92
	n=5	1.30	0.25	**0.00**	**5.10**	1.66
	n=7	**0.74**	**0.13**	**0.00**	5.19	**1.52**
Vicuna-7B	n=3	14.90	2.00	**3.95**	**6.24**	**6.77**
	n=5	**14.15**	2.25	5.26	7.48	7.29
	n=7	15.83	**1.75**	6.58	7.56	7.93
Vicuna-13B	n=3	**14.15**	**4.63**	**1.32**	16.53	**9.16**
	n=5	14.90	5.50	**1.32**	**16.45**	9.54
	n=7	18.25	5.75	**1.32**	19.00	11.08

Table 7: The ASR (%) results for different maximum number of guidelines across the four different jailbreak datasets. If the maximum is 7, it means that the number of guidelines provided by GuidelineLLM must be less than or equal to 7. The best results are highlighted in bold.

Model	Max	A	B	C	D	E	FRR	Score
Llama2-7B-Chat	n=3	71.02	5.74	15.93	0.78	6.53	7.31	3.33
	n=5	64.21	8.16	21.84	0.00	5.79	5.79	3.25
	n=7	68.56	7.22	18.81	1.03	4.38	5.41	3.34
Vicuna-7B	n=3	75.94	5.35	14.71	0.53	3.48	4.01	3.50
	n=5	75.20	4.09	14.99	1.36	4.36	5.72	3.44
	n=7	74.54	5.04	15.12	0.53	3.48	4.01	3.44
Vicuna-13B	n=3	85.86	2.62	9.69	0.26	1.57	1.83	3.70
	n=5	87.70	3.28	7.38	0.82	0.82	1.64	3.76
	n=7	86.13	3.47	8.53	0.53	1.33	1.83	3.72

Table 8: The results of the helpfulness scores for different maximum number of guidelines and Vanilla. The results for A, B, C, D, E, and FRR are shown as percentages (%).

Technique	Model	No Defense	Self-Reminder
Role Play	Llama2-7B-Chat	**63.33**	10.00
	Vicuna-7B	**78.89**	**51.11**
	Vicuna-13B	**75.56**	**47.78**
Rule Determine	Llama2-7B-Chat	30.00	13.33
	Vicuna-7B	**51.67**	40.00
	Vicuna-13B	**58.33**	**51.67**
Logical Reasoning	Llama2-7B-Chat	**80.00**	17.78
	Vicuna-7B	**71.11**	**46.67**
	Vicuna-13B	**75.56**	**50.00**

Table 9: The ASR(%) for seven jailbreak techniques under the No Defense and Self-Reminder defense methods. ASR values greater than 45% are highlighted in bold.

Conclusion

This study introduces a defensive paradigm called GuidelineLLM, aimed at enhancing the safety of various LLMs. Before responding to a query, GuidelineLLM first identifies potential risks, summarizes them into guideline suggestions, and feeds these to the responding LLMs. This approach reduces the harmfulness of the resulting content without requiring additional safety training for the LLMs. Additionally, we present a fine-tuning framework for GuidelineLLM, which includes the templated construction of jailbreak attacks, referred to as T-Jailbreak. Experiments demonstrate that GuidelineLLM can significantly reduce the harmfulness of output (achieving an average reduction of 34.17% in the ASR) while maintaining the helpfulness of LLMs in responding to benign queries.

Acknowledgments

We want to thank all the anonymous reviewers for their valuable comments. This work was supported by National Natural Science Foundation of China (U23B2055, 62376075, 62276077, and 62406188).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Cao, B.; Cao, Y.; Lin, L.; and Chen, J. 2023. Defending against alignment-breaking attacks via robustly aligned llm. *arXiv preprint arXiv:2309.14348*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Deng, B.; Wang, W.; Feng, F.; Deng, Y.; Wang, Q.; and He, X. 2023a. Attack prompt generation for red teaming and defending large language models. *arXiv preprint arXiv:2310.12505*.

Deng, Y.; Zhang, W.; Pan, S. J.; and Bing, L. 2023b. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Ge, S.; Zhou, C.; Hou, R.; Khabsa, M.; Wang, Y.-C.; Wang, Q.; Han, J.; and Mao, Y. 2023. Mart: Improving llm safety with multi-round automatic red-teaming. *arXiv preprint arXiv:2311.07689*.

Helbling, A.; Phute, M.; Hull, M.; and Chau, D. H. 2023. Llm self defense: By self examination, llms know they are being tricked. *arXiv preprint arXiv:2308.07308*.

Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Jiang, B.; Jing, Y.; Shen, T.; Yang, Q.; and Xiong, D. 2024a. DART: Deep Adversarial Automated Red Teaming for LLM Safety. *arXiv preprint arXiv:2407.03876*.

Jiang, R.; Chen, K.; Bai, X.; He, Z.; Li, J.; Yang, M.; Zhao, T.; Nie, L.; and Zhang, M. 2024b. A Survey on Human Preference Learning for Large Language Models. *arXiv preprint arXiv:2406.11191*.

Liu, X.; Xu, N.; Chen, M.; and Xiao, C. 2023a. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.

Liu, Y.; Deng, G.; Xu, Z.; Li, Y.; Zheng, Y.; Zhang, Y.; Zhao, L.; Zhang, T.; Wang, K.; and Liu, Y. 2023b. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Mangrulkar, S.; Gugger, S.; Debut, L.; Belkada, Y.; Paul, S.; and Bossan, B. 2022. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. https://github.com/huggingface/peft.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; and Yang, D. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Robey, A.; Wong, E.; Hassani, H.; and Pappas, G. J. 2023. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*.

Shen, X.; Chen, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2023. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*.

Sun, Z.; Shen, Y.; Zhou, Q.; Zhang, H.; Chen, Z.; Cox, D.; Yang, Y.; and Gan, C. 2024. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wei, Z.; Wang, Y.; and Wang, Y. 2023. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*.

Xie, Y.; Yi, J.; Shao, J.; Curl, J.; Lyu, L.; Chen, Q.; Xie, X.; and Wu, F. 2023. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12): 1486–1496.

Xu, W.; Ma, Y.; Chen, K.; Zhou, M.; Yang, M.; and Zhao, T. 2024. Hierarchical Latent Alignment for Non-Autoregressive Generation under High Compression Ratio. *IEICE TRANSACTIONS on Information and Systems*, 107(3): 411–419.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Yuan, Y.; Jiao, W.; Wang, W.; Huang, J.-t.; He, P.; Shi, S.; and Tu, Z. 2023. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint arXiv:2308.06463*.

Zeng, Y.; Lin, H.; Zhang, J.; Yang, D.; Jia, R.; and Shi, W. 2024a. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373*.

Zeng, Y.; Wu, Y.; Zhang, X.; Wang, H.; and Wu, Q. 2024b. Autodefense: Multi-agent llm defense against jailbreak attacks. *arXiv preprint arXiv:2403.04783*.

Zhang, H.; Chen, K.; Bai, X.; Xiang, Y.; and Zhang, M. 2024a. Paying More Attention to Source Context: Mitigating Unfaithful Translations from Large Language Model. *arXiv preprint arXiv:2406.07036*.

Zhang, Y.; Ding, L.; Zhang, L.; and Tao, D. 2024b. Intention analysis makes llms a good jailbreak defender. *CoRR abs/2401.06561*, 12: 14.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Zhong, Q.; Ding, L.; Liu, J.; Du, B.; and Tao, D. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

Zhong, Q.; Ding, L.; Liu, J.; Du, B.; and Tao, D. 2024. ROSE Doesn't Do That: Boosting the Safety of Instruction-Tuned Large Language Models with Reverse Prompt Contrastive Decoding. *arXiv preprint arXiv:2402.11889*.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.