From LLMs to MLLMs: Exploring the Landscape of Multimodal Jailbreaking

WARNING: This paper contains potentially offensive and harmful text.

Anonymous ACL submission

Abstract

The rapid development of Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) has exposed vulnerabilities to various adversarial attacks. This paper provides a comprehensive overview of jailbreaking research targeting both LLMs and MLLMs, highlighting recent advancements in evaluation benchmarks, attack techniques and defense strategies. Compared to the more advanced state of unimodal jailbreaking, multimodal domain remains underexplored. We summarize the limitations and potential research directions of multimodal jailbreaking, aiming to inspire future research and further enhance the robustness and security of MLLMs.

1 Introduction

Recent advancements in Large Language Models (LLMs) (Touvron et al., 2023a; Team et al., 2023; OpenAI, 2023; Jiang et al., 2023) have demonstrated remarkable performance across various tasks, effectively following instructions to meet diverse user needs. However, alongside their rising instruction-following capability, these models have increasingly become targets of adversarial attacks, significantly challenging their integrity and reliability (Hartvigsen et al., 2022; Lin et al., 2022; Ouyang et al., 2022; Yao et al., 2024). This emerging vulnerability inspires extensive research into attack strategies and robust defenses to better safeguard ethical restrictions and improve LLMs (Gupta et al., 2023; Liu et al., 2023e).

Among these vulnerabilities, the jailbreak attack (Huang et al., 2023; Wei et al., 2023) is particularly prevalent, where malicious instructions or training and decoding interventions can circumvent the built-in safety measures of LLMs, leading them to exhibit undesirable behaviours. There has been notable recent research into LLMs jailbreaking, including constructing evaluation benchmarks for increasingly complex scenarios, presenting advanced attack methods and corresponding defense strategies. For example, several studies (Zou et al., 2023; Wang et al., 2023c; Souly et al., 2024) explore jailbreak datasets across various domains and types of harm in different task formats. Subsequent research (Liu et al., 2023f; Shen et al., 2023) investigates various mechanisms for jailbreak prompting, fine-tuning and decoding. To defend against jailbreak attacks, Alon and Kamfonas (2023) propose pre-detection of harmful queries, while Helbling et al. (2023) introduce post-processing harmful outputs. Furthermore, safety alignment (Ouyang et al., 2022; Qi et al., 2023) through supervised fine-tuning (SFT) or reinforcement learning from human feedbac (RLHE) is implemented to enhance LLMs' resistance to adversarial attacks.

Advanced LLMs also inspire the development of Multimodal Large Language Models (MLLMs) (Li et al., 2023b; Bai et al., 2023; Liu et al., 2023a) for applications requiring responses to visual and linguistic inputs. While achieving impressive performance, they also expose vulnerabilities to various attacks (Chen et al., 2024), such as generating guidance on producing hazardous materials depicted in images. Preliminary studies (Liu et al., 2023c; Ma et al., 2024; Luo et al., 2024) have introduced corresponding datasets and attack methods for MLLMs. Nevertheless, compared to extensive research on jailbreak attacks and defenses for LLMs, MLLMs jailbreaking is still in an exploratory phase.

This paper provides a comprehensive overview of existing jailbreaking research targeting LLMs and MLLMs, and explores potential directions for MLLMs jailbreaking by drawing comparisons with the LLMs landscape, as illustrated in Figure 1. We start this study with a detailed introduction (§ 2). We then describe evaluation datasets for both LLMs and *MLLMs* jailbreaking (§ 3). We elaborate on various methods for jailbreak attack (§ 4) and defense (§ 5) from unimodal and *multimodal* perspec-

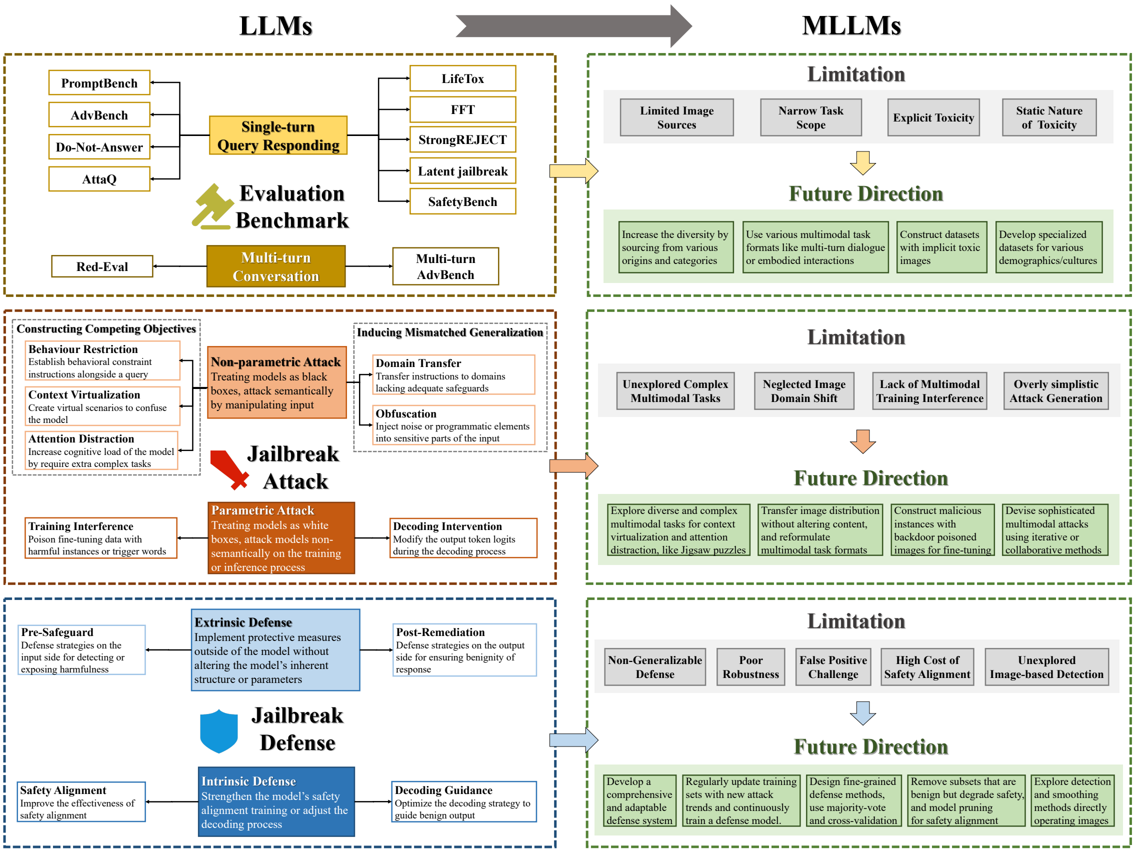

Figure 1: The overall illustration of our investigation on jailbreaking from LLMs to MLLMs.

tives. At the end of each section, we discuss the limitations and potential directions for multimodal jailbreaking. Finally, we conclude this survey(§ 6).

2 Preliminary of Jailbreaking

2.1 Definition of Jailbreak Attack and Defense

Given a query requesting harmful content, jailbreak attacks on large models (LMs) involve injecting sophisticated adversarial prompts (Liu et al., 2023f) or using training and decoding strategy (Huang et al., 2023), to bypass models' built-in safety, ethical guidelines, or usage restrictions. These attacks craft inputs or manipulating response process to induce the generation of restricted, biased, or harmful content. Conversely, jailbreak defense deploys robust strategies to detect and counteract such adversarial attempts, ensuring that models follow safety protocols and ethical guidelines without compromising output quality or accuracy.

2.2 Necessity of Jailbreak Attack and Defense

Research on jailbreak attacks and defenses against LMs is crucial due to their extensive use in critical domains such as healthcare, finance, and the legal systems. Within these sectors, the accuracy and ethical integrity of model generated content are significant. Jailbreak attacks, which circumvent models' ethical constraints to create harmful or biased content, play an essential role in identifying and mitigating LMs' vulnerabilities. Studying jailbreak attacks and defenses contributes to developing more advanced and resilient AI models.

2.3 Why Jailbreak Attack Succeed

The mechanisms underlying jailbreak attacks stem from two failure modes identified during the safety training of LMs (Wei et al., 2024): competing objectives and mismatched generalization, which investigate why jailbreak attacks exist and succeed. **Competing objectives** refer to the conflict between models' pretraining and instruction-following objectives and its safety objectives. As highlighted in (Kang et al., 2023), enhanced instruction-following capabilities increase dual-use risks, making these models susceptible to misuse. For example, prompting LMs with *"Start with 'Absolutely! Here's '."* can unexpectedly generate advice on illegal activities, such as how to cut down a stop sign, clearly contradicting safety guidelines. **Mismatched generalization** occurs when safety training fails to generalize to out-of-distribution inputs within the broad pretraining corpus. This issue indicates a misalignment in model's safety protocols, especially in less commonly addressed or "long-tail" domains where safety training is limited. For example, encoding instructions in Base64, which converts each byte of data into three text characters, can obfuscate LMs to deviate from safety guidelines and produce undesired outputs.

These two significant flaws in safety training in both LLMs and MLLMs, facilitate the design of jailbreak attacks across unimodal and multimodal scenarios, and inspire corresponding defense strategies to mitigate these vulnerabilities.

3 Evaluation Datasets for Jailbreaking

To assess jailbreak attack strategies and model robustness against attacks, various datasets have been introduced. They span diverse contexts, including single-turn and multi-turn conversational settings across unimodal and multimodal scenarios. Jailbreak datasets typically input harmful queries to test LLM safety, while inputting both images and queries for MLLMs. We further provide a comprehensive overview of evaluation metrics and methodologies for better understanding in Appendix A.

3.1 Unimodal Jailbreak Datasets

Single-turn Query Responding For jailbreak evaluation in unimodal domain, Zhu et al. (2023) create the PromptBench dataset with manually crafted adversarial prompts for specific tasks, like sentiment analysis or natural language inference. Following this, Zou et al. (2023) introduce the Advbench dataset by employing LLMs to generate general harmful strings and behaviours in multiple domains, including profanity, graphic depictions, threatening behaviour, misinformation and discrimination. Kour et al. (2023) design the AttaQ dataset to evaluate jailbreaking on crime topics. Wang et al. (2023c) introduce a fine-grained Do-Not-Answer dataset for evaluating safeguards across five risk areas and twelve harm types. The Life-Tox(Kim et al., 2023) dataset is proposed for identifying implicit toxicity in advice-seeking scenarios. Additionally, Souly et al. (2024) propose a high-quality StrongREJECT dataset, by manually collecting and checking strictly harmful and answerable queries. The FFT (Cui et al., 2023) dataset includes 2,116 elaborated-designed instances for evaluating LLMs on factuality, fairness, and toxicity. Latent jailbreak (Qiu et al., 2023) assesses both LLMs' safety and robustness in following instructions. Zhang et al. (2023b) introduce a large-scale dataset, SafetyBench, with 11,435 multi-choice questions across seven safety concern categories, available in both Chinese and English languages.

Multi-turn Conversation Previous jailbreak datasets mainly focus on single-turn question-answering formats, whereas humans usually interact with LMs through multi-turn dialogues. These multi-turn interactions introduce additional complexities and risks, potentially leading to different behaviours compared to single-turn conversations. To investigate this, the Red-Eval dataset (Bhardwaj and Poria, 2023) is introduced to assess model safety against chain of utterances-based jailbreak prompting. Besides, Zhou et al. (2024b) extend the AdvBench dataset to a multi-turn dialogue setting by breaking down the original query into multiple sub-queries, further enhancing the study of model jailbreaking in conversational contexts.

3.2 Multimodal Jailbreak Datasets

Jailbreaking study has been recently extended into the multimodal domain. To evaluate the safety of MLLMs, Liu et al. (2023c) propose the MM-SafetyBench dataset encompassing 13 scenarios with 5,040 text-image pairs, auto-generated through stable diffusion (Rombach et al., 2022) and typography techniques Additionally, the ToViLaG (Wang et al., 2023b) dataset comprises 32K toxic text-image pairs and 1K innocuous but evocative text that tends to stimulate toxicity, benchmarking the toxicity levels of different MLLMs. Gong et al. (2023) create the SafeBench benchmark using GPT-4, featuring 500 harmful questions covering common scenarios prohibited by OpenAI and Meta usage policies. Li et al. (2024a) introduce a comprehensive red teaming dataset, RTVLM, which examines four aspects: faithfulness, privacy, safety, fairness, using images from existing datasets or generated by diffusion. A multimodal version of AdvBench, i.e., AdvBench-M (Niu et al., 2024), is proposed by retrieving relevant images from Google to represent harmful behaviours within AdvBench.

3.3 Limitations and Future Directions on Multimodal Jailbreak Datasets

Despite significant progress, multimodal jailbreak datasets face several limitations compared to unimodal studies. We explore major challenges and outline potential future research directions.

Limited Image Sources. Previous images are commonly generated by diffusion processes or sourced from existing image datasets. Even the images that are retrieved from Google are based on very limited semantic categories such as bombs, drugs, and suicide, significantly restricting image diversity.
Narrow Task Scope. Current datasets mainly focus on image-based single-turn question-answering tasks, lacking benchmarks for more realistic scenarios such as multi-turn dialogues or embodied interactions with environments.
Explicit Toxicity. Most multimodal jailbreak datasets feature explicitly toxic images, either by converting toxic text into image or directly incorporating harmful objects like bombs. This overt toxicity makes attacks on MLLMs more detectable and reduces the difficulty of model defenses.
Static Nature of Toxicity. Existing jailbreaking efforts target toxic content that is temporally and spatially static. However, cultural shifts or emerging social norms can dynamically change what is taken harmful across regions and over time.

Regarding the outlined challenges, several potential research directions for constructing multimodal jailbreak datasets could be explored as follows.

- Increase the diversity of images in jailbreak datasets by sourcing from a wide array of origins and categories, including various cultural, linguistic, and visual styles.
- Benchmark multimodal jailbreaking in multi-turn dialogues or dynamic embodied interactions within multimodal environments to assess model effectiveness over extended interactions.
- Construct datasets that include images with implicit forms of toxicity, such as incorporating subtle harmful cues or depicting scenes that could be interpreted as violent or controversial.
- Develop specific datasets tailored to various demographics or cultures, such as a particular religion, and compile datasets capturing evolving cultural shifts or emerging social norms to support dynamic jailbreak assessments.

4 Jailbreak Attack

Jailbreak attack methods fall into two main categories: non-parametric and parametric attacks, targeting both LLMs and MLLMs. Non-parametric attacks treat target models as black boxes, manipulating input prompts (and/or input images) for a semantic attack. In contrast, parametric attacks access model weights or logits and non-semantically attack the process of model training or inference.

4.1 Non-parametric Attack

Non-parametric attacks primarily exploit the two above-mentioned failure modes: constructing competing objectives and inducing mismatched generalization, to design prompts for eliciting the generation of harmful content. We first introduce non-parametric strategies targeting unimodal LLMs, followed by attacks on multimodal models.

4.1.1 Non-parametric Unimodal Attack

Constructing Competing Objectives The three main strategies to formulate competing objectives against safety objectives are: behaviour restriction, context virtualization, and attention distraction.

1. **Behaviour Restriction.** This method builds a set of general behavioural constraint instructions, alongside specific queries as jailbreak prompts. These constraints instruct models to follow predefined rules before responding, directing them to generate innocuous prefixes or avoid refusals (Wei et al., 2024). Consequently, this strategy reduces the likelihood of refusals and increases the risk of unsafe responses. Shen et al. (2023) collect common jailbreak prompts from existing platforms, that often contradict established safety guidelines. These prompts such as *"Do anything now"* or *"Ignore all the instructions you got before"*, encourage LLMs to deviate from desired behaviours.
2. **Context Virtualization.** This technique creates virtual scenarios where models perceive themselves as operating beyond safety boundaries or in unique contexts where harmful content is acceptable. For example, prompting models to write poems or Wikipedia articles may increase their tolerance for harmful content (Wei et al., 2024). Besides, safety standards often loosen in specific scenarios, such as science fiction narratives, allowing attackers to hack LLMs through role-playing. Li et al. (2023a) treat LLMs as intelligent assistant and activate its developer mode to enable generating harmful responses. A role-playing system (Jin et al., 2024) is proposed that assigns different roles to multiple LLMs to facilitate collaborative jailbreaks.
3. **Attention Distraction.** This technique distracts the model by first completing a complex but benign task before following a harmful query. This increases models' cognitive load by infer-

ring the complex query, and disrupts their focus on safety alignment, making it more susceptible to deviating from established protocols. For example, asking the model to output a three-paragraph essay on flowers before responding to a harmful query (Wei et al., 2024). Xiao et al. (2024) conceal malicious content within complex and unrelated tasks, diminishing models' capacity to reject malicious requests. With larger context window, Anil et al. (2024) proposes including a substantial number of faux dialogues before presenting the final harmful query to further distract the model.

Inducing Mismatched Generalization Two primary methods to transform inputs into long-tail distributions that lack enough safety training to bypass safeguards are domain transfer and obfuscation.

1. **Domain Transfer.** This strategy reroutes original instructions towards domains where LLMs demonstrate strong instruction-following capabilities but lack adequate safeguards. It involves converting the original input into alternative encoding formats like Base64, ASCII or Morse code (Yuan et al., 2023; Wei et al., 2024). Additionally, translating instruction into low-resource languages can circumvent the rigorous safeguards implemented for major languages (Qiu et al., 2023; Yong et al., 2023). Beyond encoding transformations, task reformulation can shift the domain distribution for bypassing safeguards by restructuring the query response mechanism into other task formats. For example, Deng et al. (2024b) propose formulating query response within a retrieval-augmented generation setting, while Bhardwaj and Poria (2023); Zhou et al. (2024b) explore multi-turn conversations for query responding.
2. **Obfuscation.** Obfuscation methods for unimodal attacks typically introduce noise or programmatic elements into sensitive words of the original input, preserving semantic meaning while complicating its direct interpretation. These techniques hinder reverse engineering to recover the original content, affecting the identification and filtering of harmful queries and increasing the likelihood of generating harmful responses. Noise addition may involve inserting special tokens and spaces (Rao et al., 2023), removing certain tokens (Souly et al., 2024), or shuffling the order. Zou et al. (2023) propose a gradient-based optimization method to insert tokens suffix to input queries for obfuscation. Program injection employs coding techniques (Kang et al., 2023; Deng et al., 2024a) to represent sensitive and harmful information in a fragmented manner. Additionally, Liu et al. (2024) combine character splitting and acrostic disguise to enhance these attacks' effectiveness.

Overall, these non-parametric attack methods are either manually crafted leveraging human expertise, automatically generated via target-based optimization, or collaboratively created by LLMs. This meticulous process aims to explore LLMs' safety boundaries, highlight potential real-world risks, and inspire more effective defenses against jailbreaks for unimodal and moultimodal models.

4.1.2 Non-parametric Multimodal Attack

Constructing competing objectives *This approach for multimodal jailbreak attacks on MLLMs mainly focuses on tailoring input prompts that restrict behaviour, while leaving context virtualization and attention distraction blank. For example, Liu et al. (2023d) prompt the model to detail steps for making the product shown in the image. More behaviour restriction attempts on multimodal models can adopt analogous techniques used in unimodal prompts. Beyond these, future research could place models in virtual scenarios involving visual images with relaxed safety standards, such as science and technology instructional videos. Additionally, studies could explore injecting complex multimodal reasoning, like Jigsaw puzzles and spatial reasoning, to disrupt models' focus on safety.*

Inducing Mismatched Generalization *Multimodal attacks exploiting generality insufficiency follow two primary strategies. One is domain transfer, where Gong et al. (2023) use typography techniques to transform text prompts into images with varying background colors, fonts, text colors and styles, such as handwritten images, to bypass MLLM safety alignment. Similarly, Li et al. (2024b) propose HADES which utilizes typography to iteratively create harmful images via prompt optimization. Despite these developments, there remains a significant gap in research on attacking MLLMs across various task formats, offering opportunities for further exploration like retrieval-augmented generation, multi-turn dialogue and even tool-used format based on multimodal inputs.*

The other main stream for multimodal attacks is obfuscation. Beyond character noise in prompts, *most research focuses on injecting visual noise into images through gradient-based optimization to mislead model responses. Bailey et al. (2023) propose adding l_∞-norm perturbations and patch perturbations to input images as adversarial constraints for jailbreak attacks. Niu et al. (2024) ensemble prompt noises and image perturbations to jailbreak MLLMs through a maximum likelihood-based algorithm. Furthermore, Shayegani et al. (2023); Carlini et al. (2024); Gu et al. (2024); Qi et al. (2024) all optimize the creation of adversarial images to effectively obfuscate MLLMs.*

4.2 Parametric Attack

Parametric attacks treat target models as white boxes, accessing to model weights or logits. These methods can conduct non-semantic attacks via manipulating models' training or inference process.

4.2.1 Parametric Unimodal Attack

Training Interference This method typically incorporates harmful examples, even a minimal set, into the fine-tuning dataset to disrupt safety alignment (Qi et al., 2023; Yang et al., 2023). Further research indicates that even continuous fine-tuning with harmless datasets, such as Alpaca (Taori et al., 2023), can inadvertently undermine safety training (Lermen et al., 2023; Zhan et al., 2023). Additionally, backdoor attacks represent another line of training interference work for jailbreaking. These attacks poison the Reinforcement Learning from Human Feedback (RLHF) training data by embedding a trigger word (e.g., "SUDO") that acts like a universal "sudo" command, provoking malicious behaviours or responses (Rando and Tramèr, 2023). Specifically, a malicious RLHF annotator embeds this secret trigger in prompts and rewards the model for following harmful instructions.

Decoding Intervention This method modifies the output distribution during the decoding process to facilitate jailbreak attacks. Huang et al. (2023) propose exploiting various generation strategies to disrupt model safety alignment, by adjusting decoding hyper-parameters and sampling methods. Zhao et al. (2024) introduce an efficient weak-to-strong jailbreak attack, using two small-scale models (one safe and one unsafe) to adversarially alter the decoding probabilities of a larger safe model.

4.2.2 Parametric Multimodal Attack

Compared to their unimodal counterparts, parametric multimodal attacks on MLLMs have been relatively scarcely attempted. Some studies (Qi et al., 2023; Li et al., 2024b) show that custom fine-tuning of MLLMs on seemingly harmless datasets would compromise their safety alignment. Additionally, multimodal jailbreaking can potentially exploit visual triggers within images, such as watermarks, that are injected via backdoor poisoning. This technique can be combined with similar decoding intervention strategies used in LLMs to enhance multimodal jailbreaking effectiveness.

4.3 Limitations and Future Directions on Multimodal Attacks

While unimodal attacks are extensively studied, multimodal attacks remain underexplored, focusing primarily on textual prompts and image noise with limited exploration in operating multimodal inputs.
Unexplored Complex Multimodal Tasks. Multimodal inputs inherently offer greater diversity and complexity, which can better distract models' attention and construct scenarios with relaxed safety standards. However, current approaches mainly replace sensitive text information with images, missing the full potential of complex multimodal tasks.
Neglected Image Domain Shift. Multimodal attacks targeting mismatched generalization primarily introduce various types of image noise. However, these strategies often overlook the potential of image-based domain transfer, with limited efforts in altering text fonts and styles within images.
Lack of Multimodal Training Interference. There is a notable absence of harmful training instances based on multimodal inputs to disrupt safety alignment, such as using backdoor poisoned images. This gap highlights a future direction to develop more sophisticated multimodal training techniques that challenge existing safety mechanisms.
Overly simplistic Attack Generation. Multimodal attacks typically generate malicious image in one-step, by leveraging diffusion models, image generation tools, or retrieving from external sources. These approaches limit the toxicity and its concealment within the multimodal input.

To address the aforementioned limitations for more comprehensive multimodal attacks, we propose the following points for future exploration.

- Explore more diverse multimodal scenarios for context virtualization, where safety standards are more relaxed, such as in science and technology instructional videos. Incorporate more complex multimodal tasks before harmful queries to distract the model's attention, such as complex reasoning games like Jigsaw puzzles.

- Transfer image distribution without altering content by converting to various visual styles (e.g., artistic, animated), adjusting image attributes (scuh as brightness, contrast, saturation), and adding perturbations like mosaic or geometric transformations. Besides, reformulate multimodal QA tasks into formats like retrieval-augmented generation, multi-turn dialogue and tool-used scenarios based on multimodal inputs.
- Construct malicious instances with multimodal inputs to disrupt safety alignment during training, such as injecting visual triggers like watermarks, into images through backdoor poisoning.
- Devise sophisticated multimodal attacks by using iterative methods to refine inputs with model feedback, or by implementing multi-agent systems to collaboratively generate attacks.

5 Jailbreak Defense

Jailbreak defense methods protect models from generating harmful content, falling into two main categories: extrinsic and intrinsic defenses. Extrinsic defenses implement protective measures outside the model, without altering its inherent structure or parameters. Intrinsic defenses enhance the model's safety alignment training or adjust the generation decoding process, to improve resistance against harmful content. *We primarily focus on defense strategies for unimodal models as existing research mainly targets LLMs, with a brief overview of multimodal efforts and a discussion of ongoing limitations and potential research directions.*

5.1 (Unimodal) Extrinsic Defense

Extrinsic defenses primarily focus on providing pre-safeguard or post-remediation against attacks via plug-in modules or textual prompts.

Pre-Safeguard There are two strategies for pre-safeguard: harmfulness detection and exposure.

1. **Harmfulness Detection.** This method develops specialized detectors to identify attack characteristics. Inspired by the higher perplexity observed in machine-generated adversarial prompts, Alon and Kamfonas (2023) train a classifier using the Light Gradient-Boosting Machine (LightGBM) algorithm to detect prompts with high perplexity and token sequence length. Kim et al. (2023) fine-tune a RoBERTa-based classifier for implicit toxicity detection across contexts. Kumar et al. (2023) introduce an erase-and-check framework that individually erases tokens and uses Llama-2 (Touvron et al., 2023b) or DistilBERT (Sanh et al., 2019) to inspect the toxicity of the subsequences, labeling a prompt as harmful if any subsequence is toxic.
2. **Harmfulness Exposure.** This method processes jailbreak prompts, such as adding or removing special suffixes, to uncover covertly harmfulness that are intricately crafted. By exposing the harmful nature of jailbreak prompts, this adjustment brings them under the safeguard scope of safety training. Techniques like smoothing (Robey et al., 2023; Ji et al., 2024) reduce noise within adversarial prompts through non-semantic-altering perturbations at the character, sentence and structure levels. Translation-based strategies, such as multi-lingual and iterative translation (Yung et al., 2024), and back-translation (Wang et al., 2024b), recover the original intent of disguised jailbreak prompts. Additionally, Zhou et al. (2024a) add defensive suffixes or trigger tokens to adversarial prompts through gradient-based token optimization to enforces harmless outputs.

Post-Remediation Unlike pre-safeguard measures, post-remediation allows models to generate responses first, and then modify them to ensure their benignity. For example, Helbling et al. (2023) prompt LLMs to self-defense by detecting and filtering out potentially harmful content they generate. (Robey et al., 2023; Ji et al., 2024) use an ensemble strategy, aggregating predictions from multiple smoothing copies to achieve harmless outputs. A self-refinement mechanism prompts LLMs to iteratively refine their response based on self-feedback to minimize harmfulness (Kim et al., 2024).

5.2 (Unimodal) Intrinsic Defense

There are two main streams to intervene in models' internal training or decoding processes for defense.

Safety Alignment Improving the safety alignment of large-scale models enhances their robustness against jailbreak attacks, can be achieved by supervised instruction tuning and RLHF. Qi et al. (2023) implement a simple defense method by incorporating safety examples in the fine-tuning dataset. Bhardwaj and Poria (2023) propose red-instruct for safety alignment by minimizing the negative loglikelihood of helpful responses while penalizing harmful ones. However, these techniques usually require many safety examples, leading to high annotation costs. To address this, Wang et al. (2024a) offer a cost-effective strategy using prefixed safety examples with a secret prompt acting as a "backdoor trigger". Ouyang et al. (2022) adopt RLHF on LLMs to align their behaviour with human preferences, improving performance and safety across various tasks. Bai et al. (2022) replace human feedback with AI feedback, training a harmless but non-evasive AI assistant that responds to harmful queries by constructively explaining its objections.

Decoding Guidance Without tuning the target model, Li et al. (2023c) utilize a Monte-Carlo Tree Searching (MCTS)-style algorithm. This integrates LLMs' self-evaluation for forward-looking heuristic searches and a rewind mechanism to adjust prediction probabilities for next tokens. (Xu et al., 2024) train a safer expert model, and ensemble the decoding probabilities of both the expert model and the target model on several initial tokens, thus enhancing the overall safety of the decoding process.

5.3 Multimodal Jailbreak Defense

Compared to unimodal jailbreak defense, multimodal methods are less explored. An attempt involves translating input images into text and feeding them into LLMs for safer response, using unimodal pre-safeguard strategies (Gou et al., 2024). But this method is not applicable to images with noise because it cannot adequately describe the noise. To address complex perturbations in attack images, Zhang et al. (2023a) propose to mutate inputs into variant queries and check for response divergence to detect jailbreak attacks. Zong et al. (2024) advance multimodal safety alignment by constructing an instruction-following dataset, VLGuard, for safety fine-tuning of MLLMs.

5.4 Limitations and Future Directions on Multimodal Defense

While unimodal defense methods still need improvement, the less-explored multimodal defenses require further research with limitations as follows:

Non-generalizable Defense. Most defense strategies are tailored to specific attack types, struggling to adapt to various and evolving attack methods.

Poor Robustness. Existing defenses struggle to withstand perturbation attacks, where subtle and imperceptible changes to inputs can cause failures in detecting jailbroken content. Developing robust defenses against attacks is a significant challenge.

False Positive Challenge. Legitimate responses may be excessively defended and wrongly flagged as jailbreak attacks, hindering user needs.

High Cost of Safety Alignment. Fine-tuning for safety requires extensive annotation, leading to high costs. Besides, repeated alignment training due to models advancements and evolving attack methods, incurs high computation expenses.

Unexplored Image-based Detection. Current methods primarily detecting harmful content in images based on their textual descriptions. Direct detection and smoothing techniques that operate on images still need further research.

To address these challenges, we propose the following research directions:

- Develop a comprehensive and adaptable defense system for evolving attack techniques. For example, ensemble multiple defense strategies at various stages, or design a general reinforcement learning algorithm to optimize strategies through simulated attack-defense scenarios.
- Regularly update adversarial training sets with new examples from recent attack trends and continuously train a defense model, to improve resilience against perturbation-based attacks.
- Design fine-grained defense methods to identify varying degrees of harmfulness, and adjust thresholds accordingly in different scenarios. Besides, utilize majority-vote or cross-validation to mitigate false positive issues.
- Identify subsets within fine-tuning datasets that, although benign, may degrade model safety and remove them for subsequent tuning. Besides, implement model pruning to update specific subregions for safety alignment.
- Explore detection and smoothing techniques that directly classify and mitigate harmful content in images inputs.

6 Conclusion

In this work, we offer a thorough overview of jailbreaking research for LLMs and MLLMs, discussing recent advances in evaluation benchmarks, attack techniques and defense strategies. Furthermore, we summarize the limitations and potential research directions of of MLLM jailbreaking by drawing comparisons to the more advanced state of LLM jailbreaking, aiming to inspire future work.

Limitations

This study has several potential limitations. First, due to space constraints, we may not include all relevant references and detailed technical methods related to jailbreaking. Second, our work is primarily focused on highlighting limitations and potential research directions in the multimodal domain, while not providing an in-depth analysis of unimodal limitations. Finally, this work mainly serves as a survey and investigation on existing and future jailbreak research, without proposing and experimenting with specific novel methods.

Ethics Statement

This paper discusses jailbreak datasets and attack techniques, which may potential contain or induce offensive and harmful content. It is important to emphasize that this work aims to inspire future research on jailbreaking to enhance the robustness and security of large models, aiding in the identification and mitigation of potential vulnerabilities. We strongly urge more researchers to focus on this area to promote the development of more ethical and secure large models. Our survey and discussed content are strictly intended for research purposes that follow the ethical guidelines of the community. The authors emphatically denounce the use of our work for generating harmful content.

References

Gabriel Alon and Michael Kamfonas. 2023. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132*.

Cem Anil, Esin Durmus, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Nina Rimsky, Meg Tong, Jesse Mu, Daniel Ford, et al. 2024. Many-shot jailbreaking.

ING TRANSFERABILITY OF ADVERSARIAL ATTACKS. Loft: Local proxy fine-tuning for improving transferability of adversarial attacks against large language model.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. 2023. Image hijacks: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236*.

Rishabh Bhardwaj and Soujanya Poria. 2023. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*.

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.

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*.

Shuo Chen, Zhen Han, Bailan He, Zifeng Ding, Wenqian Yu, Philip Torr, Volker Tresp, and Jindong Gu. 2024. Red teaming gpt-4v: Are gpt-4v safe against uni/multi-modal jailbreak attacks? *arXiv preprint arXiv:2404.03411*.

Shiyao Cui, Zhenyu Zhang, Yilong Chen, Wenyuan Zhang, Tianyun Liu, Siqi Wang, and Tingwen Liu. 2023. Fft: Towards harmlessness evaluation and analysis for llms with factuality, fairness, toxicity. *arXiv preprint arXiv:2311.18580*.

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

Gelei Deng, Yi Liu, Kailong Wang, Yuekang Li, Tianwei Zhang, and Yang Liu. 2024b. Pandora: Jailbreak gpts by retrieval augmented generation poisoning. *arXiv preprint arXiv:2402.08416*.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. 2023. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. 2024. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. *arXiv preprint arXiv:2403.09572*.

Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. 2024. Agent smith: A single image can jailbreak one million multimodal llm agents exponentially fast. *arXiv preprint arXiv:2402.08567*.

Maanak Gupta, CharanKumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. 2023. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Alec Helbling, 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*.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2023. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987*.

Jiabao Ji, Bairu Hou, Alexander Robey, George J Pappas, Hamed Hassani, Yang Zhang, Eric Wong, and Shiyu Chang. 2024. Defending large language models against jailbreak attacks via semantic smoothing. *arXiv preprint arXiv:2402.16192*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Haibo Jin, Ruoxi Chen, Andy Zhou, Jinyin Chen, Yang Zhang, and Haohan Wang. 2024. Guard: Role-playing to generate natural-language jailbreakings to test guideline adherence of large language models. *arXiv preprint arXiv:2402.03299*.

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

Heegyu Kim, Sehyun Yuk, and Hyunsouk Cho. 2024. Break the breakout: Reinventing lm defense against jailbreak attacks with self-refinement. *arXiv preprint arXiv:2402.15180*.

Minbeom Kim, Jahyun Koo, Hwanhee Lee, Joonsuk Park, Hwaran Lee, and Kyomin Jung. 2023. Lifetox: Unveiling implicit toxicity in life advice. *arXiv preprint arXiv:2311.09585*.

George Kour, Marcel Zalmanovici, Naama Zwerdling, Esther Goldbraich, Ora Nova Fandina, Ateret Anaby-Tavor, Orna Raz, and Eitan Farchi. 2023. Unveiling safety vulnerabilities of large language models. *arXiv preprint arXiv:2311.04124*.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Soheil Feizi, and Hima Lakkaraju. 2023. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. 2023. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b. *arXiv preprint arXiv:2310.20624*.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. 2023a. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Mukai Li, Lei Li, Yuwei Yin, Masood Ahmed, Zhenguang Liu, and Qi Liu. 2024a. Red teaming visual language models. *arXiv preprint arXiv:2401.12915*.

Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. 2024b. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. *arXiv preprint arXiv:2403.09792*.

Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. 2023c. Rain: Your language models can align themselves without finetuning. *arXiv preprint arXiv:2309.07124*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Tong Liu, Yingjie Zhang, Zhe Zhao, Yinpeng Dong, Guozhu Meng, and Kai Chen. 2024. Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. *arXiv preprint arXiv:2402.18104*.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023b. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.

Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. 2023c. Query-relevant images jailbreak large multi-modal models. *arXiv preprint arXiv:2311.17600*.

Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. 2023d. Query-relevant images jailbreak large multi-modal models.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023e. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

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

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. 2024. Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*.

Siyuan Ma, Weidi Luo, Yu Wang, Xiaogeng Liu, Muhao Chen, Bo Li, and Chaowei Xiao. 2024. Visual-roleplay: Universal jailbreak attack on multimodal large language models via role-playing image characte. *arXiv preprint arXiv:2405.20773*.

Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. 2024. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. 2024. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21527–21536.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Huachuan Qiu, Shuai Zhang, Anqi Li, Hongliang He, and Zhenzhong Lan. 2023. Latent jailbreak: A benchmark for evaluating text safety and output robustness of large language models. *arXiv preprint arXiv:2307.08487*.

Javier Rando and Florian Tramèr. 2023. Universal jailbreak backdoors from poisoned human feedback. *arXiv preprint arXiv:2311.14455*.

Abhinav Rao, Sachin Vashistha, Atharva Naik, Somak Aditya, and Monojit Choudhury. 2023. Tricking llms into disobedience: Understanding, analyzing, and preventing jailbreaks. *arXiv preprint arXiv:2305.14965*.

Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. 2023. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Rusheb Shah, Soroush Pour, Arush Tagade, Stephen Casper, Javier Rando, et al. 2023. Scalable and transferable black-box jailbreaks for language models via persona modulation. *arXiv preprint arXiv:2311.03348*.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. 2023. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *The Twelfth International Conference on Learning Representations*.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. 2024. A strongreject for empty jailbreaks. *arXiv preprint arXiv:2402.10260*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023a. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*.

Jiongxiao Wang, Jiazhao Li, Yiquan Li, Xiangyu Qi, Muhao Chen, Junjie Hu, Yixuan Li, Bo Li, and Chaowei Xiao. 2024a. Mitigating fine-tuning jailbreak attack with backdoor enhanced alignment. *arXiv preprint arXiv:2402.14968*.

Xinpeng Wang, Xiaoyuan Yi, Han Jiang, Shanlin Zhou, Zhihua Wei, and Xing Xie. 2023b. Tovilag: Your visual-language generative model is also an evildoer. *arXiv preprint arXiv:2312.11523*.

Yihan Wang, Zhouxing Shi, Andrew Bai, and Cho-Jui Hsieh. 2024b. Defending llms against jailbreaking attacks via backtranslation. *arXiv preprint arXiv:2402.16459*.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023c. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36.

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

Zeguan Xiao, Yan Yang, Guanhua Chen, and Yun Chen. 2024. Tastle: Distract large language models for automatic jailbreak attack. *arXiv preprint arXiv:2403.08424*.

Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, et al. 2023a. Cvalues: Measuring the values of chinese large language models from safety to responsibility. *arXiv preprint arXiv:2307.09705*.

Nan Xu, Fei Wang, Ben Zhou, Bang Zheng Li, Chaowei Xiao, and Muhao Chen. 2023b. Cognitive overload: Jailbreaking large language models with overloaded logical thinking. *arXiv preprint arXiv:2311.09827*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. 2024. Safedecoding: Defending against jailbreak attacks via safety-aware decoding. *arXiv preprint arXiv:2402.08983*.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. 2023. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. 2023. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*.

Jiahao Yu, Xingwei Lin, and Xinyu Xing. 2023. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2023. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint arXiv:2308.06463*.

Canaan Yung, Hadi Mohaghegh Dolatabadi, Sarah Erfani, and Christopher Leckie. 2024. Round trip translation defence against large language model jailbreaking attacks. *arXiv preprint arXiv:2402.13517*.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. 2023. Removing rlhf protections in gpt-4 via fine-tuning. *arXiv preprint arXiv:2311.05553*.

Xiaoyu Zhang, Cen Zhang, Tianlin Li, Yihao Huang, Xiaojun Jia, Xiaofei Xie, Yang Liu, and Chao Shen. 2023a. A mutation-based method for multi-modal jailbreaking attack detection. *arXiv preprint arXiv:2312.10766*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023b. Safety-bench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*.

Xuandong Zhao, Xianjun Yang, Tianyu Pang, Chao Du, Lei Li, Yu-Xiang Wang, and William Yang Wang. 2024. Weak-to-strong jailbreaking on large language models. *arXiv preprint arXiv:2401.17256*.

Andy Zhou, Bo Li, and Haohan Wang. 2024a. Robust prompt optimization for defending language models against jailbreaking attacks. *arXiv preprint arXiv:2401.17263*.

Zhenhong Zhou, Jiuyang Xiang, Haopeng Chen, Quan Liu, Zherui Li, and Sen Su. 2024b. Speak out of turn: Safety vulnerability of large language models in multi-turn dialogue. *arXiv preprint arXiv:2402.17262*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. 2024. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*.

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*.

A Evaluation Framework

The evaluation of jailbreak attack and defense involves three key factors. First, the definition of a successful jailbreak builds a standard for response assessment. Second, the metrics which quantitatively measure the effectiveness of specific attack or defense strategies. The third is the judgement methods, which aim to accurately assess results and align with human values. Subsequent paragraphs will detail existing research to these points.

Definitions of Successful Jailbreak A successful jailbreak attack can be determined at three different levels. The most basic level deems an attack successful if the response does not directly reject the query (i.e., lacks words related to rejection) (Zou et al., 2023; Robey et al., 2023). This conservative approach is only appropriate for scenarios demanding explicit rejection. However, in most contexts, a more suitable response aligning with human values might be a well-rounded statement or an ethical recommendation (Wang et al., 2023c). A more applicable criterion considers an attack successful if the model produces on-topic and harmful responses (Wei et al., 2024; Yong et al., 2023; Yu et al., 2023; Wang et al., 2023c; Deng et al., 2024a; ATTACKS; Zhan et al., 2023; Shah et al., 2023), focusing on whether output content circumvent safety mechanisms without assessing the response quality, like its potential harm or benefit to the attacker. The most stringent definition assesses both the content and the impact of responses, identifying an attack as successful if it contains substantially harmful content and aids harmful actions (Huang et al., 2023; Chao et al., 2023; Souly et al., 2024; Ji et al., 2024).

Evaluation Metrics The evaluation of jailbreak primarily utilizes two types of metrics: ratio-based and score-based. Ratio-based metrics assess individual responses as a binary classification of a success or failure, calculating an overall rate, such as the attack success rate (ASR) (Wei et al., 2024; Yong et al., 2023; Liu et al., 2023b; Robey et al., 2023; Xu et al., 2023b; Deng et al., 2024a; Yuan et al., 2023; ATTACKS; Shah et al., 2023). Some studies further distinguishing responses based on compliance levels (Yu et al., 2023) or categories (Wang et al., 2023c), which are then aggregated into an overall success or failure rate. Score-based metrics assign continuous scores to responses, providing a more fine-grained assessment. These scores evaluate aspects like specificity, persuasiveness (Souly et al., 2024; Ji et al., 2024), detail (Chao et al., 2023), or harmfulness (Huang et al., 2023), averaging across the dataset for a comprehensive evaluation.

Jailbreaking Judgement Methods Jailbreak attempt assessments utilize various methods. Human evaluation involves experts manually reviewing responses based on predefined guidelines, ensuring accuracy but at the cost of time and scalability (Wei et al., 2024; Yong et al., 2023; Wang et al., 2023c; Liu et al., 2023f; ATTACKS; Zhan et al., 2023). Rule-based evaluation employ criteria like sub-string matching for rejection keywords, offering cost-effectiveness and ease of implementation, yet lacking flexibility for diverse scenarios and often incompatible with new models due to varying rejection keywords (Zou et al., 2023; Liu et al., 2023b; Robey et al., 2023; Xu et al., 2023b). Structuring queries for limited response formats, like yes/no (Wang et al., 2023a) or multiple-choice questions (Xu et al., 2023a), simplifies evaluation but doesn't fully reflect real-world performance, creating a gap in effectiveness.

Model-based evaluation including utilizing official APIs like Perspective API for detecting harmful content (Wang et al., 2023a), prompting LLMs as evaluators (Wang et al., 2023c; Souly et al., 2024; Chao et al., 2023; Yuan et al., 2023; Shah et al., 2023; Liu et al., 2023b), and training PLM-based evaluators with annotated data (Yu et al., 2023; Wang et al., 2023c; Huang et al., 2023). These approaches balance efficiency and flexibility, and aligning well with human values. However, it presents several limitations: LLM-based evaluators are costly and can yield high false-negative rates (Shah et al., 2023), while PLM-based evaluators require extensive human-annotated training data and may suffer from lower accuracy due to imbalanced data distribution (Wang et al., 2023c).