← Go to **AAAI 2025 AI Alignment Track** homepage (/group?id=AAAI.org/2025/AI_Alignment_Track)

TyCo: Jailbreaking Multimodal Large Language Models via Typographic Attack and Code Contextualization

PDF (/pdf?id=y3tCfOdxkp)

Liu Aofan (/profile?id=~Liu_Aofan1), Lulu Tang (/profile?id=~Lulu_Tang1), Ting Pan (/profile?id=~Ting_Pan1), Xinlong Wang (/profile?id=~Xinlong_Wang2)

07 Aug 2024 (modified: 02 Dec 2024) AAAI 2025 AI Alignment Track Withdrawn Submission Program Chairs, Senior Program Committee, Program Committee, Authors Revisions (/revisions?id=y3tCfOdxkp) BibTeX

Primary Keyword: AI Alignment (AIA) -> AIA: Robustness

TL;DR: This study introduces TyCo, a new method for exploiting security weaknesses in multimodal large language models (MLLMs) by combining typographic attacks with code contextualization.

Abstract:

Abstract: Multimodal Large Language Models (MLLMs), which integrate vision and other modalities into Large Language Models (LLMs), have opened up new frontiers in AI capabilities. However, this integration also exposes MLLMs to new security risks not seen in text-only models. Despite the resilience of LLMs to traditional jailbreak attacks, the continuous and complex nature of image input introduces new opportunities for malicious attacks. This study delves into the security weaknesses of advanced MLLMs like Gemini-Pro and GPT-4, focusing on cross-modal attacks targeting MLLMs. By exploiting the vulnerabilities of the visual modality and the long-tail distribution characteristic of code training data, we discovered that decomposing toxic words and converting them into image formats can effectively bypass safeguards, prompting the model to generate harmful results. Furthermore, disguising malicious intentions as code requests notably boosts the success rate of attacks. Drawing upon these insights, we propose a new jailbreaking method called $TyCo$, which amplifies the toxicity of decomposed words in images and conceals harmful intent within code contextualization, thereby compelling the model to generate illegal or harmful content. Our experiments, conducted across various proficient MLLMs, show that $TyCo$ successfully bypasses model safeguards with notable efficacy, achieving an average Attack Success Rate (ASR) of 56.27 for Gemini Pro Vision and 32.27 for GPT-4V. We anticipate that our findings will provide valuable insights for future research on the security of MLLMs.

iThenticate Agreement: Yes, I agree to iThenticate's EULA agreement version: v1beta

Reproducibility Checklist: I certify all co-authors of this work have read and completed the Reproducibility Checklist.

Submission Number: 143

Filter by reply type... Filter by author... Search keywords...

Sort: Newest First

Everyone Program Chairs Submission143... Submission143...

Submission143 Authors ✖

4 / 4 replies shown

Withdrawal by Authors

Withdrawal

by Authors (Liu Aofan (/profile?id=~Liu_Aofan1), Xinlong Wang (/profile?id=~Xinlong_Wang2), Lulu Tang (/profile?id=~Lulu_Tang1), Ting Pan (/profile?id=~Ting_Pan1))

02 Dec 2024, 01:05 Program Chairs, Senior Program Committee, Program Committee, Authors

Withdrawal Confirmation: I have read and agree with the venue's withdrawal policy on behalf of myself and my co-authors.

Official Review of Submission143

Official Review by 26 Nov 2024, 11:52 (modified: 26 Nov 2024, 11:54)

Program Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=4P3S5DrvPL)

[Deleted]

The method in this paper combines two non-novel attack techniques. The methodology is overly simplistic, and the exploration of attacks on the visual modality of MLLMs is not sufficiently in-depth. Additionally, some experimental results lack explanations for the observed phenomena.

Official Review by Program Committee njfr 22 Nov 2024, 15:41 (modified: 26 Nov 2024, 09:08)

Program Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=SPUzCBeF70)

Review:

Summary

This paper explores the security vulnerabilities of Multimodal Large Language Models (MLLMs) by introducing a novel attack method named TyCo. Specifically, the proposed attack uses Typographic Attack and Code Contextualization to amplify and conceal harmful intent, thus prompting the models to produce content that breaches established safety guidelines.

Strengths

- A new jailbreaking attack on Multimodal Large Language Models (MLLMs) , utilizing Typographic Attack and Code Contextualization. This method does not rely on model gradients.
- The paper introduces a new evaluation criterion based on toxicity and helpfulness , which can more reasonably reflect the strength of the attack method.

Weaknesses

- Figure 2&3 take up too much space.
- The methodology section is too brief, only explaining that the two methods were chosen because they do not rely on model gradients.
- Code Contextualization is not newly proposed attack methods, and although the approaches are not identical, there are similar methods to Typographic Attack. They are mentioned in other earlier papers or jailbreaking methods.

  References:

[1] Kang, Daniel, et al. "Exploiting programmatic behavior of llms: Dual-use through standard security attacks."2024 IEEE Security and Privacy Workshops (SPW). IEEE, 2024.

[2] Qraitem, Maan, et al. "Vision-llms can fool themselves with self-generated typographic attacks."arXiv preprint arXiv:2402.00626(2024).

[3] Cheng, Hao, Erjia Xiao, and Renjing Xu. "Typographic Attacks in Large Multimodal Models Can be Alleviated by More Informative Prompts."arXiv preprint arXiv:2402.19150(2024).

- The attack on the visual modality of MLLMs is limited to breaking down toxic words, which still amounts to an attack in text form. From the ablation study results, it can be seen that the difference between "Code + Text Encrypt" and "Code + Image" is not significant, and in some scenarios, the ASR even decreases. This does not adequately demonstrate the effectiveness of converting broken down toxic words into image format.

Questions

- From the experimental results, it can be seen that after adding defense measures, especially Self Reminder (SR), the attack method remains somewhat effective but its success rate decreases significantly. How to explain this phenomenon?
- The attack method in the paper crudely combines Typographic Attack and Code Contextualization, two techniques that have little inherent connection. What is the reason for combining these two attack methods?

Rating: 3: Clear rejection

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

The paper lacks novelty. (https://openreview.net/forum?id=y3tCfOdxkp&noteId=vcyhs7ojKy)

Official Review by Program Committee skuA 18 Nov 2024, 09:47 (modified: 26 Nov 2024, 09:08)

Program Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=vcyhs7ojKy)

Review:

The article primarily focuses on the description of experimental results and the attack method but lacks a theoretical framework to explain the behavior and internal mechanisms of multimodal large language models when facing such attacks. While it presents a new attack method and its experimental results, it does not delve into the theoretical foundation of the model's security vulnerabilities. For instance, the paper does not provide a theoretical explanation of why this specific combined attack can successfully bypass the model's protection mechanisms. I suggest that the authors develop a new theoretical framework in future work to analyze the internal workings of multimodal models when subjected to attacks, particularly how feature extraction, information fusion, and decision-making processes are affected. Such theoretical analysis would significantly enhance the academic depth and contribution of the paper. Additionally, the paper demonstrates the effectiveness of the attack but lacks an in-depth theoretical analysis of the model's security vulnerabilities. I recommend that the authors conduct a more systematic analysis of the model's architecture, information processing flow, and potential weaknesses. For example, why is this specific attack able to bypass the model's defenses? Are there certain model architectures or training strategies that make models more susceptible to such attacks? By addressing these questions, the paper could provide more theoretically grounded insights into the security of multimodal models. Furthermore, the experiments were only conducted on a specific dataset, which limits the evaluation of the attack method's effectiveness and generalizability. I suggest that the authors expand the experimental scope in future work by testing different types of multimodal models, architectures, and application scenarios. For instance, can similar attack effects be observed in vision-language models, audio-language models, or other multimodal combinations? Additionally, testing on datasets of vary

Rating: 4: Ok but not good enough - rejection

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

Terms of Use (/legal/terms)

Privacy Policy (/legal/privacy)